Foundations and Trends® in Computer Graphics and Vision

Semantic Image Segmentation: Two Decades of Research

Suggested Citation: Gabriela Csurka, Riccardo Volpi and Boris Chidlovskii (2022), "Semantic Image Segmentation: Two Decades of Research", Foundations and Trends® in Computer Graphics and Vision: Vol. 14, No. 1-2, pp 1–162. DOI: 10.1561/0600000095.

Gabriela Csurka
Naver Labs Europe
Gabriela.Csurka@naverlabs.com

Riccardo Volpi
Naver Labs Europe
Riccardo.Volpi@naverlabs.com

Boris Chidlovskii
Naver Labs Europe
Boris.Chidlovskii@naverlabs.com

now
the essence of knowledge
Boston — Delft

Contents

Semantic Image Segmentation: Two Decades of Research

Gabriela Csurka[1], Riccardo Volpi[2] and Boris Chidlovskii[3]

[1]*Naver Labs Europe, France; Gabriela.Csurka@naverlabs.com*
[2]*Naver Labs Europe, France; Riccardo.Volpi@naverlabs.com*
[3]*Naver Labs Europe, France; Boris.Chidlovskii@naverlabs.com*

ABSTRACT

Semantic image segmentation (SiS) plays a fundamental role in a broad variety of computer vision applications, providing key information for the global understanding of an image. This survey is an effort to summarize two decades of research in the field of SiS, where we propose a literature review of solutions starting from early historical methods followed by an overview of more recent deep learning methods including the latest trend of using transformers. We complement the review by discussing particular cases of the weak supervision and side machine learning techniques that can be used to improve the semantic segmentation such as curriculum, incremental or self-supervised learning.

State-of-the-art SiS models rely on a large amount of annotated samples, which are more expensive to obtain than labels for tasks such as image classification. Since unlabeled data is instead significantly cheaper to obtain, it is not surprising that Unsupervised Domain Adaptation (UDA) reached a broad success within the semantic segmentation community. Therefore, a second core contribution of this monograph is to summarize five years of a rapidly growing

Gabriela Csurka, Riccardo Volpi and Boris Chidlovskii (2022), "Semantic Image Segmentation: Two Decades of Research", Foundations and Trends® in Computer Graphics and Vision: Vol. 14, No. 1-2, pp 1–162. DOI: 10.1561/0600000095.

field, Domain Adaptation for Semantic Image Segmentation (DASiS) which embraces the importance of semantic segmentation itself and a critical need of adapting segmentation models to new environments. In addition to providing a comprehensive survey on DASiS techniques, we unveil also newer trends such as multi-domain learning, domain generalization, domain incremental learning, test-time adaptation and source-free domain adaptation. Finally, we conclude this survey by describing datasets and benchmarks most widely used in SiS and DASiS and briefly discuss related tasks such as instance and panoptic image segmentation, as well as applications such as medical image segmentation.

We hope that this monograph will provide researchers across academia and industry with a comprehensive reference guide and will help them in fostering new research directions in the field.

Preface

Semantic image segmentation (SiS) plays a fundamental role towards a general understanding of the image content and context. In concrete terms, the goal is to label image pixels with the corresponding semantic classes and to provide boundaries of the class objects, easing the understanding of object appearances and the spatial relationships between them. Therefore, it represents an important task towards the design of artificial intelligent systems. Indeed, systems such as intelligent robots or autonomous cars should have the ability to coherently understand visual scenes, in order to perceive and reason about the environment holistically.

Hence, semantic scene understanding is a key element of advanced driving assistance systems (ADAS) and autonomous driving (AD) (Teichmann *et al.*, 2018; Hofmarcher *et al.*, 2019) as well as robot navigation (Zurbrügg *et al.*, 2022). The information derived from visual signals is generally combined with other sensors such as radar and/or LiDAR to increase the robustness of the artificial agent's perception of the world (Yurtsever *et al.*, 2020). Semantic segmentation fuels applications in the fields of robotic control and task learning (Fang *et al.*, 2018; Hong *et al.*, 2018b), medical image analysis (see Section 4.3), augmented reality (DeChicchis, 2020; Turkmen, 2019), satellite imaging (Ma *et al.*, 2019) and many others.

The growth of interest in these topics has also been caused by recent advances in deep learning, which allowed a significant performance boost in many computer vision tasks – including semantic image segmentation. Understanding a scene at the semantic level has long been a major topic in computer vision, but only recent progress in the field has allowed machine learning systems to be robust enough to be integrated into real-world applications.

The success of deep learning methods typically depends on the availability of large amounts of annotated training data, but manual annotation of images with pixel-wise semantic labels is an extremely tedious and time consuming process. As the major bottleneck in SiS is the high cost of manual annotation, many methods rely on graphics platforms and game engines to generate synthetic data and use them to train segmentation models. The main advantage of such synthetic rendering pipelines is that they can produce a virtually unlimited amount of labeled data. Due to the constantly increasing photo-realism of the rendered datasets, the models trained on them yield good performance when tested on real data. Furthermore, they allow to easily diversify data generation, simulating various environments and weather/seasonal conditions, making such data generation pipeline suitable to support the design and training of SiS models for the real world.

While modern SiS models trained on such simulated images can already perform relatively well on real images, their performance can be further improved by domain adaptation (DA) – and even with *unsupervised domain adaptation* (UDA) not requiring any target labels. This is due to the fact that DA allows to bridge the gap caused by the *domain shift* between the synthetic and real images. For the aforementioned reasons, sim-to-real adaptation represents one of the leading benchmarks to assess the effectiveness of *domain adaptation for semantic image segmentation* (DASiS).

The aim of our monograph is to overview the research field of SiS. On the one hand, we propose a literature review of semantic image segmentation solutions designed in the last two decades – including early historical methods and more recent deep learning ones, also covering the recent trend of using transformers with attention mechanism. On the other hand, we devote a large part of the monograph to survey methods

designed *ad hoc* for DASiS. While our work shares some similarities with some of the previous surveys on this topic, it covers a broader set of DASiS approaches and departs from these previous attempts pursuing different directions that are detailed below.

Amongst the existing works surveying SiS methods, we can mention Thoma (2016) who gives a brief overview of some of the early semantic segmentation and low-level segmentation methods. Li *et al.* (2018a) and Zhou *et al.* (2018) discuss some of the early deep learning-based solutions for SiS. A more complete survey on deep SiS models has been proposed by Minaee *et al.* (2020), while Zhang *et al.* (2020a) focus on reviewing semi- and weakly supervised semantic segmentation models. We cover most of these methods in Section 1, where we provide a larger spectrum of the traditional SiS methods in Section 1.1. Then, in Section 1.2, we organize the deep SiS methods according to their *most important characteristics*, such as the type of encoder/decoder, attention or pooling layers, solutions to reinforcing local and global consistency. In contrast to the previous surveys, this section also includes the latest SiS models that use attention mechanisms and transformers as encoder and/or decoder. One of the core contributions of this section is Table 2.1, which presents a broad set of deep models proposed in the literature, and summarized according to the above mentioned characteristics. Finally, in Section 1.3 we review not only semi- and weakly supervised SiS solutions, but also new trends whose goal is improving semantic segmentation, such as curriculum learning, incremental learning and self-supervised learning.

In Section 2, we present and categorize a large number of approaches devised to tackle the DASiS task. Note that previous DA surveys (Gopalan *et al.*, 2015; Csurka, 2017; Kouw and Loog, 2021; Zhang and Gao, 2019; Venkateswara and Panchanathan, 2020; Singh *et al.*, 2020; Csurka, 2020; Wang and Deng, 2018; Wilson and Cook, 2020) address generic domain adaptation approaches that mainly cover image classification and mention only a few adaptation methods for SiS. Similarly, in recent surveys on domain generalization (Wang *et al.*, 2020b; Zhou *et al.*, 2020a), online learning (Hoi *et al.*, 2018) and robot perception (Garg *et al.*, 2020), several DA solutions are mentioned, but yet DASiS received only marginal attention here. The most complete survey – and therefore most similar to the content of our Section 2 –

is by Toldo *et al.* (2020a), which also aimed at reviewing the recent trends and advances developed for DASiS. Nevertheless, we argue that our survey extends and enriches it in multiple ways. First, our survey is more recent in such a quickly evolving field as DASiS, so we address an important set of recent works appeared after their survey. Second, while we organize the DASiS methods according to how domain alignment is achieved similarly to (Toldo *et al.*, 2020a) – namely on *image, feature or output level* – we complement it with different ways of grouping DASiS approaches, namely based on their most important *characteristics*, such as the backbone used for the segmentation network, the type and levels of domain alignments, any complementary techniques used and finally the particularity of each method compared to the others. We report our schema in Table 2.1, which represents one of the core contributions of this monograph. Third, we survey a large set of complementary techniques in Section 2.3 that can help boost the adaptation performance, such as self-training, co-training, self-ensembling and model distillation.

Finally, in Section 2.4 we propose a detailed categorization of some of the *related DA tasks* – such as multi-source, multi-target domain adaptation, domain generalization, source-free adaptation, domain incremental learning, etc. – and survey solutions proposed in the literature to address them. None of the previous surveys has such a comprehensive survey on these related DA tasks, especially what concerns semantic image segmentation.

To complement the above two sections, which represent the core contributions of our monograph, we further provide in Section 3 a list of the datasets and benchmarks typically used to evaluate SiS and DASiS methods, covering the main metrics and discuss different SiS and DASiS evaluation protocols. Furthermore, in Section 4 we propose a short overview of the literature for three tasks strongly related to SiS, namely instance segmentation in Section 4.1, panoptic segmentation in Section 4.2 and medical image segmentation in Section 4.3.

We hope that our monograph, with its comprehensive survey of the main trends in the field of semantic image segmentation, will provide researchers both across academia and in the industry with a solid basis and background to help them develop new methods and foster new research directions.

1

Semantic Image Segmentation (SiS)

Semantic image segmentation (SiS) – sometimes referred to as content-based image segmentation – is a computer vision problem where the task is to determine to which semantic class each pixel of an image belongs to. Typically, this problem is approached in a supervised learning fashion, by relying on a dataset of images annotated at pixel level, and training with them a machine learning model to perform the task. This task is inherently more challenging than image classification, where the aim is to predict a single label for a given image. Furthermore, the task is more than the extension of image classification to pixel-level classification, as in contrast to image classification where each image can be considered independently from the others, in SiS the neighboring pixels are strongly related with each other and their labeling should be considered together, tackling the problem as an image partitioning into semantic regions. Hence, while the models in general indeed try to minimize the *pixel-level cross-entropy loss* between the *ground-truth* (GT) segmentation map and the *predicted* segmentation map, additional constraints or regularizing terms are necessary in general to ensure, for example, local labeling consistency or to guide segmentation boundary smoothness.[1]

[1] For more details on different losses for SiS we refer the reader to Section 1.2.10

The name of the task, *semantic image segmentation*, reflects the goal of determining the nature, *i.e.* semantics, of different parts of an image. Semantic labels in general refer to *things* such as "car", "dog", "pedestrian" or *stuff* such as "vegetation", "mountain", "road", "sky". Things and stuff are terms extensively used in the literature, where the former includes classes associated with *countable* instances and the latter indicates classes associated with the *layout of a scene*. Note that a related, still different problem is low-level image segmentation (not addressed in this survey), which consists in an unsupervised partitioning of an image into coherent regions according more to some low-level cues, such as color, texture or depth. Another related field is instance segmentation (discussed in Section 4.1), which differs from semantic segmentation as the latter treats multiple object instances with the same semantics as a single entity, while the former treats multiple objects of the same class as distinct individual objects (or instances). The extension of instance segmentation to panoptic segmentation, where *stuff* is also taken into account, is further discussed in Section 4.2.

The aim of this section is to provide a comprehensive literature review of SiS methods proposed since the beginning of the field. It is organized as follows. In Section 1.1 we first provide an overview of the historical SiS methods preceding the deep learning era. Then, in Section 1.2, we focus on deep learning-based models proposed for SiS, following Minaee *et al.* (2020), and propose to categorize them by their main principles. In particular, we collect the methods in Table 2.1 detailing their main characteristics, such as the encoder and decoder used, whether they rely on attention modules, how they tackle the semantic consistency within regions, on what kind of data the models were tested on, and what are the main characteristics of each model.

Finally, we conclude this section with Section 1.3 where we discuss some of the semantic segmentation solutions that depart from the classical setting, such as exploiting the unlabeled data (Section 1.3.1), relying on weak or none annotations (Section 1.3.2), exploiting curriculum learning strategies (Section 1.3.3), learning the semantic classes incrementally (Section 1.3.4) or fine-tuning a self-supervised pre-trained model (Section 1.3.5). Note that the models proposed in this section have generally been tested *in domain* – that is, training and testing

data come from the same data distribution. The case when training and testing data come from two different distributions, – *i.e.* the model trained on a source domain (*e.g.* synthetic environment) needs to be adapted to a new target domain (*e.g.* real world), – is discussed in detail in Section 2.

1.1 Historical SiS Methods

Methods preceding the deep learning era mainly focused on three directions to approach the segmentation problem: 1) local appearance of semantic classes, 2) local consistency of the labeling between locations, and 3) how to incorporate prior knowledge into the pipeline to improve the segmentation quality. These three aspects are addressed independently in the semantic segmentation pipeline as illustrated in Figure 1.1. They can also be approached within a unified probabilistic framework such as a Conditional Random Field (CRF), as described in Section 1.1.2. The latter methods enable at training time a joint estimation of the model parameters and therefore ensure at test time a globally consistent labeling. Yet, they carry a relatively high computational cost. Note that the three aspects are also addressed by the deep learning models, where they are jointly learned in an end-to-end manner, together with the main supervised task, as we will see in Section 1.2.

In what follows we briefly discuss how the above three components were addressed and combined by the methods proposed before the deep learning era.

1.1.1 Modeling local appearance

The local appearance can be defined at different levels, including a representation proposed at every pixel location (He *et al.*, 2004; Kumar and Hebert, 2005; Schroff *et al.*, 2006; Li *et al.*, 2009), patches on a regular grid (Verbeek and Triggs, 2007a; Larlus *et al.*, 2010), positions of interest points (Leibe *et al.*, 2004; Cao and Fei-Fei, 2007; Yang *et al.*, 2007) or regions obtained through low-level segmentation referred to as *super-pixels* (Borenstein and Ullman, 2004; Cao and Fei-Fei, 2007; Yang *et al.*, 2007). Note that a sparse description in general enables

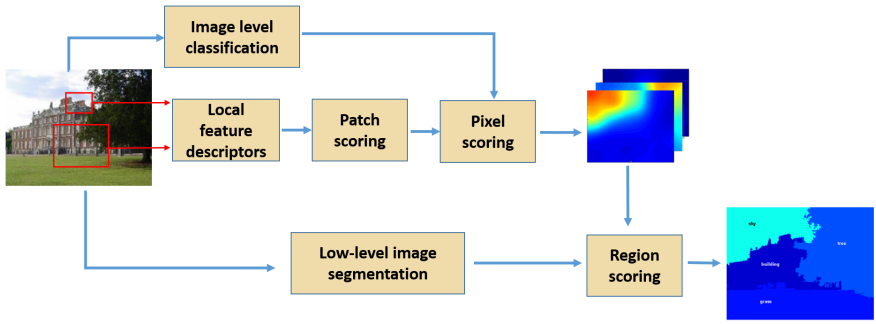

Figure 1.1: In the model proposed in Csurka and Perronnin (2011), the local appearance, global and local consistencies are addressed independently. First, for patches considered at multiple scale SIFT (Lowe, 2004) and local color statistics are extracted and transformed into high-level Fisher vector representations (Perronnin and Dance, 2007) allowing fast and efficient patch scoring. The global consistency is addressed by an image-level classifier, which is used to filter out improbable classes, while the local consistency is ensured by low-level segmentation assigning to each super-pixel the semantic label based on averaged class probabilities. Figure based on Csurka and Perronnin (2011).

faster processing and still provides excellent accuracy compared to the dense description. The same method can sometimes consider the combination of multiple representations such as using interest points and regions (Cao and Fei-Fei, 2007; Yang *et al.*, 2007) or using dense sampling and regions (Kumar and Hebert, 2005).

Amongst early local appearance features we can mention raw image representations (Schroff *et al.*, 2006), a combination of Gaussian filter outputs, colors, and locations computed for each pixel called textons (Shotton *et al.*, 2009), SIFT (Lowe, 2004), and local color statistics (Clinchant *et al.*, 2007). As mentioned above, the local features are often computed on image patches extracted either on a (multi-scale) grid (Verbeek and Triggs, 2007b; Csurka and Perronnin, 2011) or at detected interest point locations (Cao and Fei-Fei, 2007; Yang *et al.*, 2007).

These local representations are often clustered into so called *visual words* (Csurka *et al.*, 2004; Jurie and Triggs, 2005) and the local image entity (pixel, patch, super-pixels) is labeled by simply assigning the corresponding feature to the closest visual word (Schroff *et al.*, 2006) or

fed into a classifier (Plath *et al.*, 2009). Alternatively, these low-level local features can also be used to build higher level representations such as Semantic Texton Forest Shotton *et al.* (2009), Bag of Visual Words (Csurka *et al.*, 2004), Fisher Vectors (Perronnin and Dance, 2007), which are fed into a classifier that predicts class labels at patch level (Csurka and Perronnin, 2011; Ladický *et al.*, 2009), pixel level (Shotton *et al.*, 2009) or region level (Yang *et al.*, 2007; Gonfaus *et al.*, 2010; Hu *et al.*, 2012).

Topic models, such as probabilistic Latent Semantic Analysis (Hofmann, 2001) and Latent Dirichlet Allocation (Blei *et al.*, 2003) consider the bag-of-words as a mixture of several *topics* and represent a region as a distribution over visual words. Such representations have been extended to image segmentation by explicitly incorporating spatial coherency in the model to encourage similar latent topic assignment for neighboring regions with similar appearance (Cao and Fei-Fei, 2007) or by combining topic models with Random Fields (Orbanz and Buhmann, 2006; Verbeek and Triggs, 2007a; Larlus *et al.*, 2010).

1.1.2 Reinforcing local and global consistency

To reinforce the segmentation consistency, the local appearance representation and its context are generally incorporated within a Random Field (RF) framework, mainly the Markov Random Field (MRF) (Verbeek and Triggs, 2007a; Gould *et al.*, 2008; Kato and Zerubia, 2012) or the Conditional Random Field (CRF) (Shotton *et al.*, 2009; He *et al.*, 2004; Verbeek and Triggs, 2007b). While the MRF is generative in nature, the CRF directly models the conditional probability of the labels given the features.

Note that the *unary potentials* in these RF models can be pixels (Shotton *et al.*, 2009), patches (Verbeek and Triggs, 2007b; Plath *et al.*, 2009; Larlus *et al.*, 2010) or super-pixels (Lucchi *et al.*, 2011; Lempitsky *et al.*, 2011) represented by a corresponding appearance feature as described in Section 1.1.1.

In these probabilistic frameworks, label dependencies are modeled by a random field (MRF or CRF), and an optimal labeling is determined usually by energy minimization. Prior information can be imposed

through *clique potentials* between the nodes in the RF graph (as illustrated in Figure 1.2). The most often used *edge potentials* are the Potts model (Wu, 1982), which penalizes class transitions between neighboring nodes, and the contrast-sensitive Potts model (Boykov and Jolly, 2001), which includes a term reducing the cost of a transition in high contrast regions likely corresponding to object boundaries.

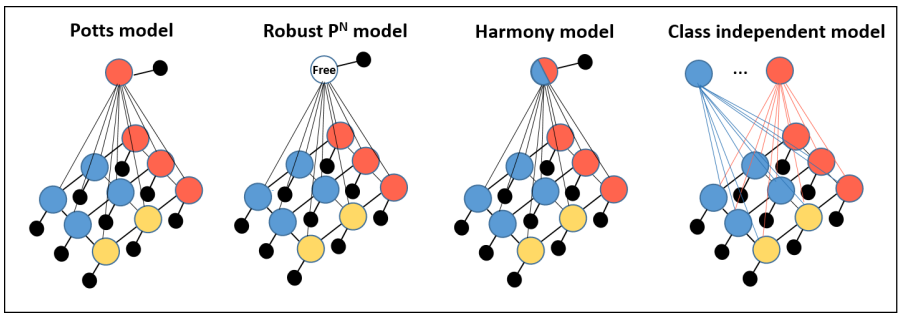

Figure 1.2: Conditional Random Field (CRF) with different increasingly more sophisticated edge potentials. From left to right: **Potts model** (Wu, 1982; Boykov and Jolly, 2001) penalizing all local nodes with a label different from the global node; **The Robust P^N model** (Kohli *et al.*, 2009) that adds an extra "free label" to the Potts model in order to not penalize local nodes; **Harmony model** (Gonfaus *et al.*, 2010) allowing different labels to coexist in a power set; **Class independent model** (Lucchi *et al.*, 2011) modeling each classes with its own global node to make the inference more tractable. Figure based on Lucchi *et al.* (2011).

To enforce region-level consistency, higher order potentials can be added to the CRF model in order to ensure that all pixels within a low-level region have the same label (see examples in Figure 1.2). As such, Kohli *et al.* (2009) propose the Robust P^N model that adds an extra *free label* to the Potts model in order not to penalize local nodes. Krähenbühl and Koltun (2011) propose a fully connected dense CRF that models the pairwise dependencies between all pairs of pixels with pairwise edge potentials defined by a linear combination of Gaussian kernels, making the inference highly efficient.

The associative hierarchical CRF model by Ladický *et al.* (2009) incorporates context from multiple quantization levels (pixel, segment, and segment union/intersection) in a joint optimization framework using graph cut-based move-making algorithms. The Harmony poten-

tials (Gonfaus *et al.*, 2010) model global preferences where any possible combination of class labels can be encoded; this enforces the consistency between local and global label assignments of the nodes. In the Pylon model (Lempitsky *et al.*, 2011), each image is represented by a hierarchical segmentation tree, and the resulting energy – combining unary and boundary terms – is optimized using the graph cut. Plath *et al.* (2009) use a CRF-based on an multi-scale pre-segmentation of the image into patches, which couples local image features with image-level multi-class SVM to provide the local patch evidences. Instead, Lucchi *et al.* (2011) propose to model each class with its own global node to make the inference more tractable (see Figure 1.2). In the LayoutCRF model (Winn and Shotton, 2006), the pairwise potentials are asymmetric and impose local spatial constraints which ensures a consistent layout whilst allowing to cope with object deformations.

In alternative to the RF framework, segmentation methods often ensure local consistency by relying on images decomposed into super-pixels. Such unsupervised partitioning of the image is obtained with low-level segmentation methods such as a Mean Shift (Comanicu and Meer, 2002) or hierarchical image segmentation (Arbelaez *et al.*, 2011). A class label is assigned for each super-pixel either in a post processing step (Csurka and Perronnin, 2011) (see *e.g.* Figure 1.1) or by relying on region descriptors and a model predicting class labels at super-pixel level (Yang *et al.*, 2007; Pantofaru *et al.*, 2008). The main limitation of these methods is that there is no possible recovery if a region includes multiple classes. To overcome this limitation, several works propose to consider multiple segmentations, exploiting overlapping sets of regions (Gould *et al.*, 2009; Pantofaru *et al.*, 2008), a hierarchy of regions (Gu *et al.*, 2009; Hu *et al.*, 2012), or a graph of regions (Chen *et al.*, 2011).

1.1.3 Using prior knowledge

Amongst different types of prior knowledge, the global image classification is most often considered – as easy to obtain, – where the global scale information is used to filter or to improve the estimation at local scale (Csurka and Perronnin, 2011; Plath *et al.*, 2009; Verbeek and Triggs, 2007b). Further priors considered for SiS are object shape priors

used to guide the segmentation process (Kumar *et al.*, 2005; Yang *et al.*, 2007) or bounding boxes obtained from object detectors (Lempitsky *et al.*, 2009). Li *et al.* (2009) employ the user tags provided by Flickr as an additional cue to infer the presence of an object in the image, while He *et al.* (2006) use an environment-specific class distribution prior to guide the segmentation. Gould *et al.* (2008) and He *et al.* (2004) explicitly model spatial relationships between different classes.

1.2 Deep Learning-based SiS

In this section we describe the main types of deep learning-based SiS pipelines, grouping the corresponding methods, similarly to (Minaee *et al.*, 2020), based on their main principles. Concretely, in Section 1.2.1 we present a few works where classical models have been revisited with deep features, and in Section 1.2.2 we discuss how deep networks were combined with graphical models. Methods based on Fully Convolutional Networks are surveyed in Section 1.2.3 and those using decoders or deconvolutional networks are presented in Section 1.2.4. Several models using Recurrent Neural Networks are briefly reviewed in Section 1.2.5 and those having pyramidal architectures in Section 1.2.6. Dilated convolutions which easily aggregate multi-scale contextual information are discussed in Section 1.2.7; attention mechanisms exploited in SiS are addressed in Section 1.2.8. Finally, we end the section with very recent transformer-based models reviewed in Section 1.2.9. In addition, we propose Table 1.1, where most deep SiS methods are briefly summarized according to their main characteristics, the type of data they are evaluated on, as well as the specificity of each method compared to the others.

1.2.1 Deep features used in classical models

Following the preliminary work of Grangier *et al.* (2009) who show that convolutional neural networks (CNNs) fed with raw pixels can be trained to perform scene parsing with decent accuracy, several methods have been proposed to replace hand crafted appearance representations (see Section 1.1.1) with representations obtained from deep convolutional

Table 1.1: Summary of the state-of-the-art methods, schematized according to their characteristics. Encoder and Decoder columns indicate which modules are used to encode and decode samples, respectively. Attention column indicates which strategy - if any - is used to apply this technique. High-resolution (HR) columns indicate which method is used to refine predictions. Data type column reports information related to the specific tasks tackled. Specificity column reports important elements that are not covered by the previously described common characteristics. Acronyms and abbreviations are defined on p. 111-116.

Citation	Model Name	Encoder	Decoder	(Self-) Attention	HR Segm	Data type	Specificity
Pinheiro and Collobert (2014)	rCNN	RNN	-	-	MSP	IP	different input patch sizes
Byeon *et al.* (2015)	2D-LSTM	LSTM	-	-	-	IP	connected LSTM blocks
Dai *et al.* (2015)	BoxSup	FCN	-	-	ObjP	Obj	weakly (BBox) supervised
Long *et al.* (2015a)	FCN	FCN	-	-	C2FR	Obj	skip connections
Mostajabi *et al.* (2015)	Zoom-Out	CNNs	-	-	SP	Obj	purely feed-forward architecture
Noh *et al.* (2015)	DeConvNet	CNN	DCN	-	ObjP	Obj	candidate object proposals
Ronneberger *et al.* (2015)	UNet	FCN	DCN	-	-	Med	skip connections
Yu and Koltun (2015)	MSCA	dCN	-	-	CRF-RNN	Obj	dilated convolutions
Zheng *et al.* (2015)	CRF/RNN	FCN	-		CRF	Obj	RNN inference
Chandra and Kokkinos (2016)	DG-CRF	FCN	-	-	dCRF	Obj	Gaussian-CRF modules
Chen *et al.* (2016)	DLab-Att	DeepLab	-	PFw	CRF	Obj	multi-scale soft spatial weights
Ghiasi and Fowlkes (2016)	LRR	FCN	-	-	FP	Obj/AD	reconstr. by deconvolution
Liu *et al.* (2016)	ParseNet	FCN	-	-	-	Obj	append global representation
Paszke *et al.* (2016)	ENet	dFCN	PUP	-	-	AD	optimized for fast inference
Visin *et al.* (2016)	ReSeg	RNNs	-	-	-	AD	Gated Recurrent Unit (GRU)
Badrinarayanan *et al.* (2017)	SegNet	FCN	FCN	-	-	AD	keep maxpool locations
Chen *et al.* (2017b)	DeepLab	RbFCN	-	-	ASPP+dCRF	Obj	CNN driven cost function
Pohlen *et al.* (2017)	FRRN	FCN	DCN	-	-	AD	high-res stream
Chaurasia and Culurciello (2017)	LinkNet	FCN	DCN	-	-	AD	skip connections
Fourure *et al.* (2017)	GridNet	FCN	DCN	-	-	AD	multiple interconnected streams
Lin *et al.* (2017a)	RefineNet	FCN	RCU	-	C2FR	IP	Chained Residual Pooling
Zhao *et al.* (2017)	PSPNet	RbFCN	SPP	-	-	IP	auxiliary loss
Chen *et al.* (2018b)	Deeplab+	DeepLab	FCN	-	-	Obj/AD	depthwise separable convolution
Li *et al.* (2018b)	PAN	RbFCN	GPA	FPA	-	Obj/AD	spatial pyramid attention
Wang *et al.* (2018a)	DUC-HDC	HDC	DUC	-	-	Obj/AD	hybrid dFCN/dense upsampling
Xiao *et al.* (2018)	UperNet	ResNet	FPN	-	SPP	Obj/Mat	multi-task framework
Yang *et al.* (2018)	DenseASPP	denseNet	-	-	ASPP	AD	densely connected ASPP
Zhao *et al.* (2018c)	PSANet	RbFCN	-	PSA	-	IP	bi-direction information propagation
Fu *et al.* (2019b)	SDN	DenseN	sDCN	-	-	Obj	hierarchical supervision
Fu *et al.* (2019a)	DANet	dFCN	-	PCAM	-	IP	dual self-attention

Table 1.1: Continued

Citation	Model Name	Encoder	Decoder	(Self-) Attention	HR Segm	Data type	Specificity
He *et al.* (2019)	APCNet	RbFCN	-	-	MSP	Obj	Adaptive Context Module
Teichmann and Cipolla (2019)	ConvCRF	FCN	-		dCRF	Obj	CRF inference as convolutions
Wang *et al.* (2020a)	Axial-DLab	DeepLab	-	AAL	-	AD/PanS	position-sensitive SelfAtt
Yuan *et al.* (2020)	OCR	HRNEt	MLP	selfAtt	ObjP	IP	object-contextual representation
Ali *et al.* (2021)	XCiT	ViT	UperNet	XCA	-	IP	cross-covariance image transformers
Chu *et al.* (2021)	Twins	HTr	SFPN	LGA+GSA	-	IP	spatially separable self-attention
Guo *et al.* (2021a)	SOTR	FPN+sTrL	PUP	HTwinT	-	Obj	twin (column/row) attention
Jain *et al.* (2021)	SeMask	HTr	SFPN	SwT+SMA	-	IP	semantic priors through attention
Liu *et al.* (2021d)	SwinTr	UperNet	SFPN	SwT	-	Obj/InstS	self-att within local windows
Strudel *et al.* (2021)	Segmenter	ViT	MaskT	MHSA	-	IP	class embeddings
Ranftl *et al.* (2021)	DPT	sTrL	RCU	MHSA	-	IP	ViT-Hybrid architecture
Wang *et al.* (2021c)	PVT	sTrL+FPN	SFPN	SRA	-	IP/Inst	progressively shrinking pyramid
Xie *et al.* (2021)	SegFormer	HTr	MLP	SelfAtt	-	IP	positional-encoding-free
Zheng *et al.* (2021)	SETR	sTrL	FPN	MHSA		IP	sequence-to-sequence prediction

networks. For example, Farabet *et al.* (2012) train a multi-scale CNN to learn good features for region classification which are used to represent nodes in a segmentation tree. Instead, Farabet *et al.* (2013) consider a deep dense feature extractor that produces class predictions for each pixel independently from its neighbors and use these predictions as unary potentials in a CRF graph. They also propose an alternative, where an image is represented first as a hierarchy of super-pixels and the average class distribution for each node is computed from the pixel-level predictions of the network.

Mostajabi *et al.* (2015) propose a purely feed-forward architecture for semantic segmentation, where they concatenate the super-pixel representations with deep features extracted from a sequence of nested regions of increasing extent. These context regions are obtained by "zooming out" from the super-pixels all the way to scene-level resolution. Hariharan *et al.* (2015) propose to concatenate features from intermediate layers into so called *hyper-column* representation used as pixel descriptors for object segmentation.

1.2.2 Graphical models

Inspired by the shallow image segmentation methods that integrate local and global context (see Section 1.1.2), several works propose to combine the strengths of CNNs with CRFs by training them jointly in an end-to-end manner. As such, the model proposed by Chandra and Kokkinos (2016) relies on Gaussian-CRF modules that collect the unary and pairwise terms from the network and propose image hypothesis (scores); these scores are then converted into probabilities using the softmax function and thresholded to obtain the segmentation. In addition, they introduce a multi-resolution architecture to couple information across different scales in a joint optimization framework showing that this yields systematic improvements.

Schwing and Urtasun (2015) propose to jointly train the parameters of a CNN used to define the unary potentials as well as the smoothness terms, taking into account the dependencies between the random variables. Chen *et al.* (2017b) treat every pixel as a CRF node and exploit long-range dependencies using CRF inference to directly optimize a deep

CNN driven cost function. Teichmann and Cipolla (2019) reformulate the CRF inference in terms of convolutions; this allows them to improve the efficiency of the CRF, which is known for being hard to optimize and slow at inference time.

1.2.3 Fully convolutional networks (FCNs)

The preliminary multi-scale convolutional network developed by Farabet *et al.* (2013) learns to produce a fairly accurate segmentation map. Still, the model needs to act on a multi-scale pyramid of image windows. Instead, the Fully Convolutional Network (FCN) proposed by Long *et al.* (2015a), transforming the fully connected layers into convolution layers, enables the net to predict directly a dense high resolution output (class presence heatmaps) from arbitrary-sized inputs. To further improve the segmentation quality, they propose to obtain such prediction maps at different levels of the network, where the lower resolution outputs are upsampled using billinearly initialized deconvolutions and fused with the coarser but higher resolution feature maps (see illustration in Figure 1.3).

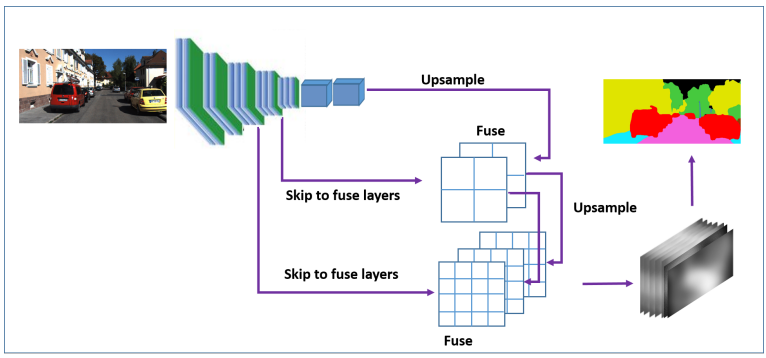

Figure 1.3: The FCN (Long *et al.*, 2015a) transforms the fully connected layers of the classification network to produce class presence heatmaps. The model takes coarse, high layer information at low resolution and upsample it using billinearly initialized deconvolutions. At each upsampling stage, the prediction maps are refined by fusing them with coarser but higher resolution feature maps. The model hence can take an arbitrary size image and produce the same size output, suitable for spatially dense prediction tasks such as SiS.

Liu *et al.* (2016) extend FCN by adding a global context vector obtained by global pooling of the feature map, showing that it reduces local confusion. This image-level information is appended to each local feature and the combined feature map is sent to the subsequent layer of the network.

The idea of using only fully convolution layers has been largely adopted as encoder for many SiS models (see Table 1.1).

1.2.4 Encoder-decoder networks

SiS models based on encoder-decoder architectures are composed of an encoder – where the input image is compressed into a latent-space representation that captures the underlying semantic information – and a decoder that generates a predicted output from this latent representation. In general, there are connections between the corresponding encoder and decoder layers allowing the spatial information to be used by the decoder and its upsampling operations (see Figure 1.4). DeConvNet (Noh *et al.*, 2015), SegNet (Badrinarayanan *et al.*, 2017), UNet (Ronneberger *et al.*, 2015), and LinkNet (Chaurasia and Culurciello, 2017) (see Figure 1.4). One such model is DeConvNet (Noh *et al.*, 2015) where the encoder computes low-dimensional feature representations via a sequence of pooling and convolution operations, while the decoder, stacked on top of the encoder, learns to upscale these low-dimensional features via subsequent unpooling and deconvolution operations; the maxpooling locations are kept during encoding and sent to the unpooling operators in the corresponding level. The trained network, applied to a set of candidate object proposals, aggregates them to produce the semantic segmentation of the whole image.

Another popular method is SegNet (Badrinarayanan *et al.*, 2017), where the encoder is similarly composed of consecutive convolutions, followed by max-pooling sub-sampling layers to increase the spatial context for pixel labelling. However, instead of deconvolution operations, trainable convolutional filters are used by the decoder and combined with the unpooling operations.

Instead of unpooling, in UNet (Ronneberger *et al.*, 2015), the corresponding features maps from the encoder are copied and concatenated

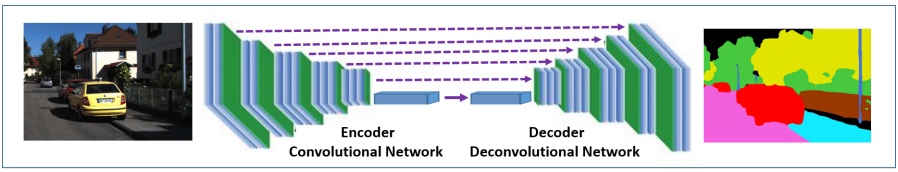

Figure 1.4: DeConvNet (Noh *et al.*, 2015) is composed of a multi-layer convolutional network as encoder and a multi-layer deconvolution network as decoder. The latter is built on the top of the output of the convolutional network, where a series of unpooling, deconvolution and rectification operations are applied yielding a dense pixel-wise class prediction map. There are connections between the corresponding encoder and decoder layers (dashed lines). In the case of DeConvNet (Noh *et al.*, 2015) the maxpooling locations in the encoder are kept at each level and sent to the unpooling operators in the corresponding level. UNet and LinkNet extend DeConvNet by skip connections; in UNet (Ronneberger *et al.*, 2015) the corresponding features maps from encoder are copied and concatenated to the layers obtained by up-convolutions, and in LinkNet (Chaurasia and Culurciello, 2017) the input of each encoder layer is bypassed to the output of the corresponding decoder.

to the layers obtained by up-convolutions, and in LinkNet (Chaurasia and Culurciello, 2017), the input of each encoder layer is bypassed to the output of the corresponding decoder layer. In addition, since the decoder is sharing knowledge learned by the encoder at every layer, the decoder can use fewer parameters yielding a more efficient network.

Pohlen *et al.* (2017) propose a Full Resolution Residual Network (FRRN) that has two processing streams: the residual one which stays at the full image resolution and a Conv-DeconvNet which undergoes a sequence of pooling and unpooling operations. The two processing streams are coupled using full-resolution residual units.

Fu *et al.* (2017) stack multiple shallow deconvolutional networks to improve accurate boundary localization which is extended by Fu *et al.* (2019b) by redesigning the deconvolutional network with intra-unit and inter-unit connections – to generate more refined recovery of the spatial resolution – and by training it with hierarchical supervision.

To maintain high-resolution representations through the encoding process, Sun *et al.* (2019b) and Yuan *et al.* (2020) consider using HRNet (Sun *et al.*, 2019a) as backbone instead of ResNet or VGG, since it enables connecting the high-to-low resolution convolution streams in parallel.

The GridNet (Fourure *et al.*, 2017) architecture follows a grid pattern which is composed of multiple paths called *streams* from the input image to the output prediction. The streams are interconnected with convolutional and deconvolutional units, where information from low and high resolutions can be shared. The two-dimensional grid structure allows information to flow horizontally in a residual resolution-preserving way or vertically through down- and up-sampling layers. The authors show that this architecture generalizes several encoder-decoder networks such as DeConvNet (Noh *et al.*, 2015), UNet (Ronneberger *et al.*, 2015) or FRRN (Pohlen *et al.*, 2017) (see Figure 1.5).

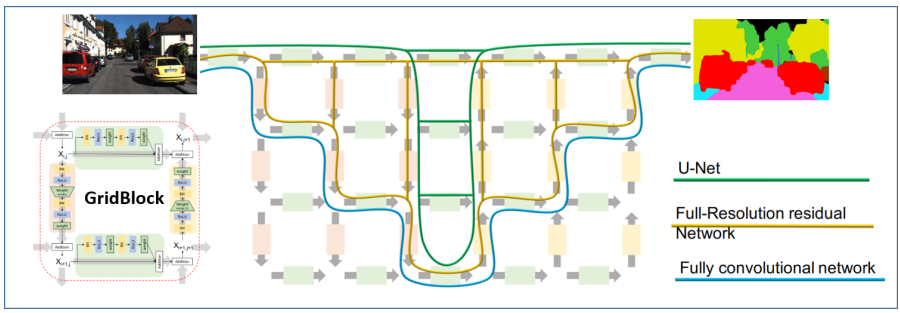

Figure 1.5: Fourure *et al.* (2017) propose the GridNet architecture showing that it generalizes several encoder-decoder networks such as DeConvNet (Noh *et al.*, 2015) (blue connections), UNet (Ronneberger *et al.*, 2015) (green connections) or FRRN (Pohlen *et al.*, 2017) (yellow connections). Figure based on Fourure *et al.* (2017).

1.2.5 Recurrent neural networks

Another group of methods consider using recurrent neural network (RNN) instead of CNNs (Pinheiro and Collobert, 2014; Gatta *et al.*, 2014); they are the first to show that modeling the long distance dependencies among pixels is beneficial to improve the segmentation quality. Pinheiro and Collobert (2014) is the first to use recurrent network for SiS exploiting the fact that RNN allows to consider a large input context with limited capacity of the model. Their model is trained with different input patch sizes (the instances) recurrently to learn increasingly large contexts for each pixel, whilst ensuring that the larger context is coherent with the smaller ones. Similarly, Gatta *et al.* (2014) propose

unrolled CNNs through different time steps to include semantic feedback connections. However, in contrast to classical RNNs, the architecture is replicated without sharing the weights and each network is fed with the posterior probabilities generated by the previous softmax classifier. Local and global features are learned in an unsupervised manner and combined.

Visin *et al.* (2016) propose the ReSeg structured prediction architecture which exploits the local generic features extracted by CNN and the capacity of RNN to retrieve distant dependencies. The model is a sequence of ReNet layers composed of four RNNs that sweep the image horizontally and vertically in both directions providing relevant global information and are followed by upsampling layers to recover the original image resolution in the final predictions. ReNet layers are stacked on top of pre-trained convolution layers, benefiting from generic local features. Zheng *et al.* (2015) propose RNNs to perform inference on the CRFs with Gaussian pairwise potentials where a mean-field iteration is modeled as a stack of CNN layers.

Byeon *et al.* (2015) introduce two-dimensional Long Short Term Memory (LSTM) networks, which consist of 4 LSTM blocks scanning all directions of an image (see Figure 1.6). This allows the model to take into account complex spatial dependencies between labels, where each local prediction is implicitly affected by the global contextual information of the image. Liang *et al.* (2016) develop the Graph LSTM model, which considers an arbitrary-shaped super-pixel as a semantically consistent node of the graph and spatial relations between the super-pixels as its edges.

1.2.6 Pyramidal architectures

While deep CNNs can capture rich scene information with multi-layer convolutions and nonlinear pooling, local convolutional features have limited receptive fields. Different categories may share similar local textures, *e.g.* "road" and "sidewalk", hence it is important to take into account the context at multiple scales to remove the ambiguity caused by local regions. Therefore several works have been proposed to solve this with pyramidal architectures, which furthermore help

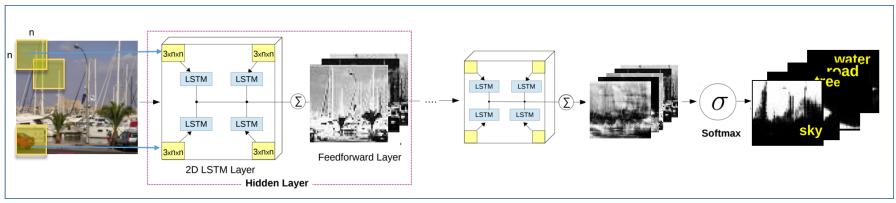

Figure 1.6: In the 2D LSTM model (Byeon *et al.*, 2015), the input image is divided into non-overlapping windows of size $n \times n$ which are are fed into four separate LSTM memory blocks. The current window of LSTM block is connected to its surrounding directions and the output of each LSTM block is then passed to the feed-forward layer where all directions are summed. At the last layer, the outputs of the final LSTM blocks are summed and sent to the softmax layer. Figure based on Byeon *et al.* (2015).

to obtain more precise segmentation boundaries. Amongst them is the work by Farabet *et al.* (2013), who transform the input image through a Laplacian pyramid where different scale inputs are fed into a pyramid of CNNs and the feature maps obtained from different scales are then combined. Ghiasi and Fowlkes (2016) develop a multi-resolution reconstruction architecture based on a Laplacian pyramid that uses skip connections from higher resolution feature maps and multiplicative confidence-weighted gating to successively refine segment boundaries reconstructed from lower-resolution maps.

The main idea behind RefineNet (Lin *et al.*, 2017a) is similarly to refine coarse resolution predictions with finer-grained ones in a recursive manner. This is achieved by short-range and long-range residual connections with identity mappings which enable effective end-to-end training of the whole system. Furthermore, a chained residual pooling allows the network to capture background context from large image regions.

The pyramid scene parsing network (PSPNet) (Zhao *et al.*, 2017) extends Spatial Pyramid Pooling (SPP), proposed by He *et al.* (2014), to semantic segmentation. The pyramid parsing module is applied to harvest different sub-region representations, followed by up-sampling and concatenation layers to form the final feature representation, which – carrying both local and global context information – is fed into a convolution layer to get the final per-pixel prediction (see Figure 1.7).

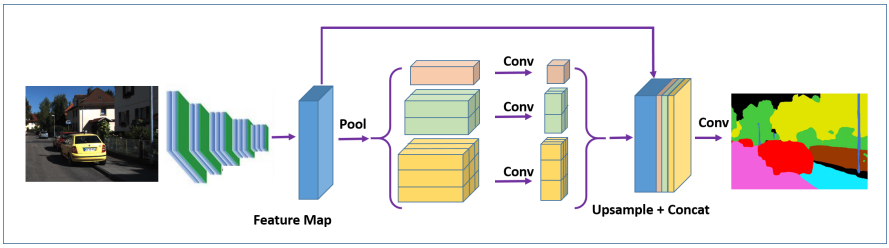

Figure 1.7: The PSPNet (Zhao *et al.*, 2017) gets the feature map of the last convolution layer of the encoder on which a pyramid parsing module is applied to harvest different sub-region representations. These representations are further upsampled and concatenated to the initial feature map to form the final feature representation. In this way local and global clues are fused together to make the final prediction more reliable. Figure based on Zhao *et al.* (2017).

Xiao *et al.* (2018) propose a Unified Perceptual Parsing framework (UperNet) which combines the Feature Pyramid Network (FPN) (Lin *et al.*, 2017b) with a Pyramid Pooling Module (PPM) (Zhao *et al.*, 2017). The model is trained in a multi-task manner with image-level (scenes, textures) and pixel-level (objects, object parts, materials) annotations.

1.2.7 Dilated convolutions

Dilated convolution-based networks (Chen *et al.*, 2017b; Yu and Koltun, 2015) aggregate multi-scale contextual information where, instead of sub-sampling, dilated convolutions are used as they support exponential expansion of the receptive field without loss of resolution nor coverage.

Many recent SiS methods adopt the dilated convolutions. For example, Paszke *et al.* (2016) extend SegNet (Badrinarayanan *et al.*, 2017) with dilated convolutions and make it asymmetric, using a large encoder and a small decoder. The encoder-decoder network by Wang *et al.* (2018a) uses hybrid dilated convolutions in the encoding phase and dense upsampling convolutions to generate pixel-level prediction. Chen *et al.* (2017b) propose to perform Spatial Pyramid Pooling with dilated convolutions where parallel atrous convolution layers with different rates capture multi-scale information. Yang *et al.* (2018) densely connect ASSP layers where the output of each dilated convolution layer is concatenated with input feature map and then fed into the next dilated

layer. He *et al.* (2019) introduce multi-scale contextual representations with multiple adaptive context modules, where each of such modules uses a global representation to guide the local affinity estimation for each sub-region. The model then concatenates context vectors from different scales with the original features for predicting the semantic labels of the input pixels.

The architecture of the popular DeepLab family (Chen *et al.*, 2017b; Chen *et al.*, 2017c) combines several ingredients including dilated convolution to address the decreasing resolution, ASPP to capture objects as well as image context at multiple scales, and CRFs to improve the segmentation boundaries (see Figure 1.8). Chen *et al.* (2018b) use the DeepLabv3 (Chen *et al.*, 2017c) framework as encoder in an encoder-decoder architecture.

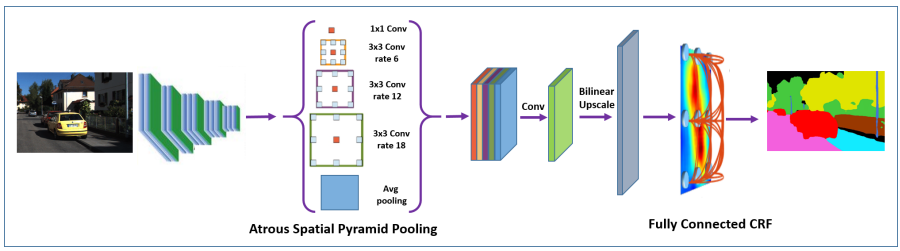

Figure 1.8: The DeepLab model (Chen *et al.*, 2017b) relies on a deep CNN with atrous convolutions to reduce the degree of signal downsampling and a bilinear interpolation stage that enlarges the feature maps to the original image resolution. A fully connected CRF is finally applied to refine the segmentation result and to better capture the object boundaries. Figure based on Chen *et al.* (2017b).

1.2.8 Attention mechanism

Attention and self-attention mechanism is widely used for many visual tasks. Amongst the methods for SiS, we can mention the work of Chen *et al.* (2016) who propose a simple attention mechanism that weighs multi-scale features at each pixel location. These spatial attention weights reflect the importance of a feature at a given position and scale.

Li *et al.* (2018b) propose a Pyramid Attention Network (PAN) where Feature Pyramid Attention modules are used to embed context features from different scales and Global Attention Upsample modules on each decoder layer to provide global context as guidance during global average pooling to select category localization details.

Fu *et al.* (2019a) introduce Dual Attention Networks (DANs) to adaptively integrate local features with their global dependencies. This is achieved by two self-attention mechanisms, the Position Attention Module capturing the spatial dependencies between any two positions of the feature maps, and the Channel Attention Module that exploits dependencies between channel maps. The outputs of the two attention modules are fused to enhance the feature representations (see illustration in Figure 1.9).

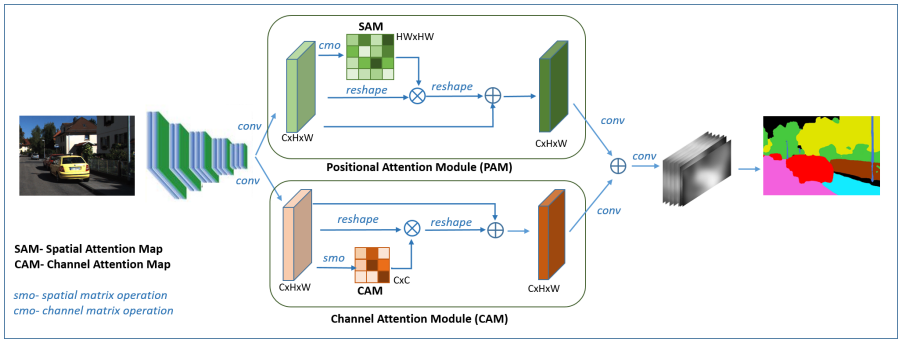

Figure 1.9: The Dual Attention Networks (Fu *et al.*, 2019a) aggregates the output of the Position Attention Module that aims at capturing the spatial dependencies between any two positions of the feature maps with the channel attention module exploiting the inter-dependencies between channel maps. Specifically, the outputs of the two attention modules are transformed by a convolution layer before fused by an element-wise sum followed by a convolution layer to generate the final prediction maps. Figure based on Fu *et al.* (2019a).

To aggregate long-range contextual information in a flexible and adaptive manner, Zhao *et al.* (2018c) propose the Point-wise Spatial Attention Network (PSANet) where each position in the feature map is connected with all the other ones learning self-adaptive attention masks which are sensitive to location and category information. The contextual information collected with a bi-directional information propagation path is fused with local features to form the final representation of complex scenes.

Wang *et al.* (2020a) propose an axial-attention block (AAL) which factorizes 2D self-attention into two 1D ones. It consists of two axial-attention layers operating along height-axis and width-axis sequentially.

1.2.9 Transformer-based models

Transformer-based models belong to the most recent networks that rely on self-attention, aimed to capture global image context and to address the segmentation ambiguity at the level of image patches. Amongst the first, Strudel *et al.* (2021) extend the recent Vision Transformer (ViT) model (Dosovitskiy *et al.*, 2021) to handle semantic segmentation problems. In contrast to convolution-based approaches, ViT allows to model global context starting from the first layer and throughout all the network. The model relies on the output embeddings corresponding to image patches and obtains class labels from these embeddings with a point-wise linear decoder or a mask transformer decoder (see Figure 1.10).

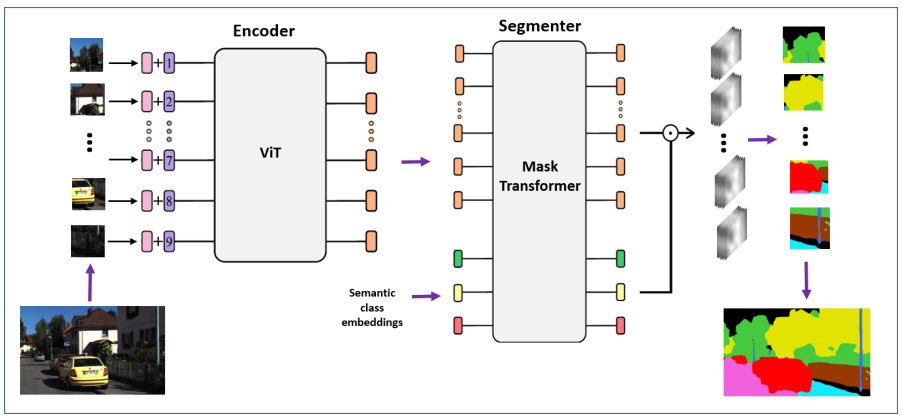

Figure 1.10: The Segmenter model (Strudel *et al.*, 2021) projects image patches to a sequence of embeddings encoded with the Vision Transformer (ViT) (Dosovitskiy *et al.*, 2021) and the Mask Transformer takes the output of the encoder and semantic class embeddings to predict segmentation class masks for each image patch; these masks are then combined to get the full semantic prediction map. Figure based on Strudel *et al.* (2021).

The pyramid Vision Transformer (PVT) (Wang *et al.*, 2021c) is another extension of ViT where incorporating the pyramid structure from CNNs allows the model to better handle dense predictions. Ranftl *et al.* (2021) introduce a transformer for dense prediction, including depth estimation and semantic segmentation. The model takes region-wise output of convolutional networks augmented with positional embed-

ding, assembles tokens from various stages of the ViT into image-like representations at various resolutions, and progressively combines them into full-resolution predictions using a convolutional decoder.

Xie *et al.* (2021) combine hierarchical transformer-based encoders to extract coarse and fine features with lightweight multi-layer perceptron decoders to aggregate information from different layers, thus combining both local and global attentions to render a more powerful representation. Guo *et al.* (2021a) also follow the hierarchical approach; they use Feature Pyramid Network (FPN) to generate multi-scale feature maps, which are then fed into a transformer – to acquire global dependencies and to predict per-instance category – and into a multi-level upsampling module to dynamically generate segmentation masks guided by the transformer output.

Liu *et al.* (2021d) introduce the Swin Transformer for constructing hierarchical feature maps and promote it as a general-purpose backbone for major downstream vision tasks. The key idea is a hierarchy of shifted window based multi-headed self-attention layers, where each layer contains a Swin Attention Block (SwT) followed by a SeMask Attention Blocks (SAB) to capture the semantic context in the encoder network. The Semantic-FPN like decoder ensures the connections between windows of consecutive layers.

Jain *et al.* (2021) incorporate semantic information into the encoder with the help of a semantic attention operation. This is achieved by adding Semantic Layers composed of SeMask Attention Blocks after the Swin Transformer Layer to capture the semantic context in a hierarchical encoder network. At each stage, the semantic maps, decoded from the Semantic Layers, are aggregated and passed through a weighted cross entropy loss to supervise the semantic context.

Chu *et al.* (2021) propose an architecture with two Twin Transformers; the first one, called Twins-PCPVT, replaces the positional encoding in PVT by positional encodings generated by a position generator. The second one, called Twins-SVT, interleaves locally-grouped attention (LGA) – capturing the fine-grained and short-distance information – with global sub-sampled attention (GSA), which deals with the long-distance and global information.

Ali *et al.* (2021) propose a novel way of computing self-attention, where attention matrices are computed over feature channels rather than on input tokens. The resulting Cross-covariance Image Transformer (XCiT) model, hence, has the intriguing property of scaling linearly with respect to the input size – in terms of computation and memory. This allows training on higher-resolution images and/or smaller patches.

1.2.10 SiS specific losses

The most common loss used in deep SiS methods is the standard *pixel-wise cross-entropy* loss, where the aim is to minimize the difference between the class predictions and the ground-truth annotations for all pixels:

$$\mathcal{L}_{ce} = -\mathbb{E}_{(\mathcal{X},\mathcal{Y})} \left[\sum_{h,w} \mathbf{y}^{(h,w)} \cdot \log(p(F(\mathbf{x}^{(h,w)}))) \right],$$

where F is the segmentation model, $p(F(\mathbf{x}^{(h,w)}))$ is a vector of class probabilities at pixel $\mathbf{x}^{(h,w)}$ and $\mathbf{y}^{(h,w)}$ is a one-hot vector with 1 at the position of the pixel's true class and 0 elsewhere.

As SiS data is compiled in real world environments, most of them are often imbalanced, with dominant portions of data assigned to a few majority classes while the rest belong to minority classes, thus forming under-represented categories. As a consequence, deep SiS methods trained with the conventional cross-entropy loss tend to be biased towards the majority classes during inference (Rahman and Wang, 2017; Buló *et al.*, 2017).

To mitigate this class-imbalance problem, SiS datasets can undergo the *resampling* step, by over-sampling minority classes and/or under-sampling majority classes. However, such approaches change the underlying data distributions and may result in sub-optimal exploitation of available data, increasing the risk of over-fitting when repeatedly visiting the same samples from minority classes.

An alternative to resampling is the *cost-sensitive* learning procedure, which introduces class-specific weights, often derived from the original data statistics. They use statically-defined cost matrices (Caesar *et al.*, 2015; Mostajabi *et al.*, 2015) or introduce additional parameter learning

steps (Khan *et al.*, 2018). Due to the spatial arrangement and strong correlations of classes between adjacent pixels, cost-sensitive learning techniques outperform resampling methods. However, the increasing complexity and a large number of under-represented (minority) classes in the recent SiS datasets make cumbersome the accurate estimation of the cost matrices.

Some approaches to the class-imbalance problem take into account the SiS specificity. One such method introduces a generalized max-pooling operator acting at the pixel-loss level (Buló *et al.*, 2017). It provides an adaptive re-weighting of contributions of each pixel, based on the loss they actually exhibit. Image pixels that incur higher losses during training are weighted more than pixels with a lower loss, thus indirectly compensating potential inter-class and intra-class imbalances within the dataset.

Another approach is to revise the standard cross-entropy loss, which optimizes the network for overall accuracy, and to address the intersection-over-union (IoU) measure instead. As described in Section 3.1.1, the IoU measure gives the similarity between the prediction and the ground-truth for every segment present in the image; it is defined as the intersection over the union of the labeled segments, averaged over all classes. Methods that optimize the IoU measure proceed either by direct optimization (Rahman and Wang, 2017) or by deploying the convex surrogates of sub-modular losses (Berman *et al.*, 2018).

Another group of SiS specific losses is driven by the observation that segmentation prediction errors are more likely to occur near the segmentation boundaries. Borse *et al.* (2021) introduce a boundary distance-based measure and include it into the standard segmentation loss. They use an inverse transformation network to model the distance between boundary maps, which can learn the degree of parametric transformations between local spatial regions.

1.3 Beyond Classical SiS

In this section, we present SiS approaches that go beyond the classical methods described above. A substantial part is focused on solutions that address the shortage of pixel-wise image annotation: in Section 1.3.1

we detail methods which exploit unlabeled samples; in Section 1.3.2 we describe approaches that rely on weak labels such as image-level annotations or bounding boxes, instead of ground-truth segmentation maps. Furthermore, Section 1.3.3 considers the case when the training is decomposed into easier-to-harder tasks learned sequentially; Section 1.3.4 reviews methods where the model's underlying knowledge is incrementally extended to new classes; finally, Section 1.3.5 focuses on the effects of self-supervised visual pre-training on SiS.

1.3.1 Semi-supervised SiS

In conventional semi-supervised learning (SSL), to overcome the burden of costly annotations, the model makes usage of a small number of labeled images and a large number of unlabeled ones. In the case of SiS, most semi-supervised methods exploit a small set of labeled images with pixel-level annotations and s set of images with image-level annotation, like image class labels or object bounding boxes. Below we present semi-supervised extension of SiS models presented in Section 1.2.

Amongst the early semi-supervised SiS works, we first mention Hong *et al.* (2015), who propose an encoder-decoder framework where image-level annotations are used to train the encoder and pixel-wise ones are used to train the decoder. Oquab *et al.* (2015) rely on FCN and introduce a max-pooling layer that hypothesizes the possible location of an object in the image. Pathak *et al.* (2015) propose a constrained CNN where a set of linear constraints are optimized to enforce the model's output to follow a distribution over latent "ground-truth" labels as closely as possible. Papandreou *et al.* (2015) develop an Expectation-Maximization (EM) method for training CNN semantic segmentation models under weakly and semi-supervised settings. The algorithm alternates between estimating the latent pixel labels, subject to the weak annotation constraints, and optimizing the CNN parameters using stochastic gradient descent (SGD). Hong *et al.* (2016) combine the encoder-decoder architecture with an attention model and exploit auxiliary segmentation maps available for different categories together with the image-level class labels.

Another principle of semi-supervised learning, consistency regularization by data augmentation, has been also successfully applied to SiS (French *et al.*, 2019; French *et al.*, 2020; Ouali *et al.*, 2020; Luo and Yang, 2020; Olsson *et al.*, 2021; Chen *et al.*, 2021b; Yuan *et al.*, 2021a) and extended by Lai *et al.* (2021) towards a directional context-aware consistency between pixels under different environments.

To further improve the consistency regularization methods, contrastive learning is used 1) by Zhou *et al.* (2021) to decrease inter-class feature discrepancy of and increase inter-class feature compactness across the dataset, 2) by Zhong *et al.* (2021) to simultaneously enforce the consistency in the label space and the contrastive property in the feature space, and 3) by Alonso *et al.* (2021) to align class-wise and per-pixel features from both labeled and unlabeled data stored in a memory bank.

Yang *et al.* (2022) show that re-training by injecting strong data augmentations on unlabeled images allows the construction of strong baselines, but such strong augmentations might yield incorrect pseudo labels. To avoid the potential performance degradation incurred by incorrect pseudo labels, they perform selective re-training via prioritizing reliable unlabeled images based on holistic prediction-level stability in the entire training course.

He *et al.* (2021c) observe that semi-supervised SiS methods in the wild severely suffer from the long-tailed class distribution and propose a *distribution alignment* and *random sampling* method to produce unbiased pseudo labels that match the true class distribution estimated from the labeled data. Similarly, to cope with long-tailed label distribution, Hu *et al.* (2021) propose an adaptive equalization learning framework that adaptively balances the training of well and badly performed categories, with a confidence bank to dynamically track category-wise performance during training.

Finally, several methods use Generative Adversarial Networks (GANs) (Goodfellow *et al.*, 2014) to train a discriminator able to distinguish between confidence maps from labeled and unlabeled data predictions (Hung *et al.*, 2018), to refine low-level errors in the predictions through a discriminator that classifies between generated and ground-truth segmentation maps (Mittal *et al.*, 2021), or to generate

fake visual data forcing the discriminator to learn better features (Souly *et al.*, 2017).

Learning from partially labeled images, where some regions are labeled and others not, is a particular case of semi-supervised segmentation (Verbeek and Triggs, 2007b; He and Zemel, 2009).

1.3.2 Weakly-supervised SiS

In contrast to semi-supervised learning, weakly-supervised SiS (Borenstein and Ullman, 2004) relies only on weak annotations such as image captions, bounding box or scribble annotations (see example in Figure 1.11).

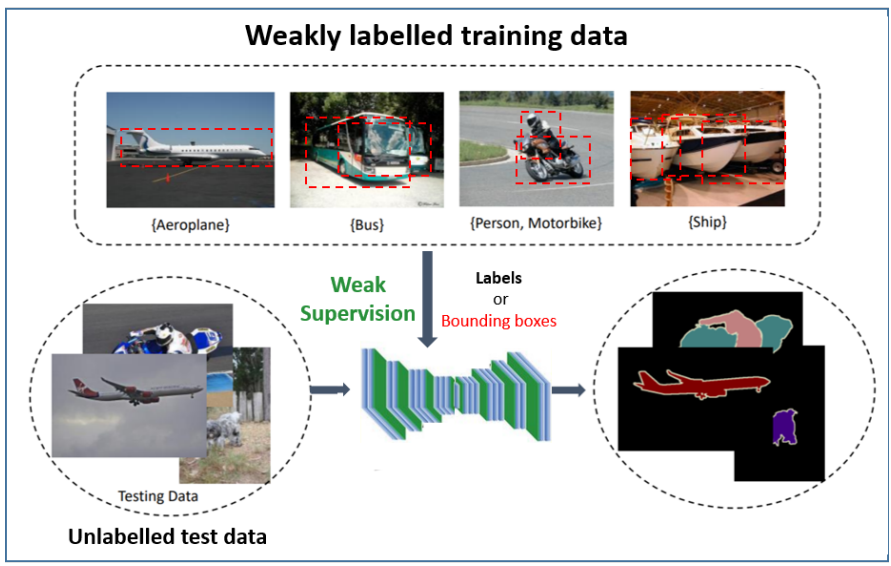

Figure 1.11: Weakly-supervised SiS aims at using either the image-level or the bounding box annotations as supervision to learn a pixel-level image segmentation (Image courtesy of Xinggang Wang).

Early methods using only image-level annotations in general rely on multiple instance learning where each image is viewed as a bag of patches or super-pixels, and the final prediction is accomplished by aggregation of the class predictions on these patches or super-pixels (Galleguillos

et al., 2008; Vezhnevets and Buhmann, 2010). In contrast, CNN based weakly-supervised SiS models exploit the observation that CNNs have remarkable localization ability despite being trained on image-level labels (Zhou *et al.*, 2016). These *Classification Activation Maps* (CAM) allow the selection of discriminative regions for each semantic class that can be used as pixel-level supervision for segmentation networks.

To improve such initial CAMs, AffinityNet (Ahn and Kwak, 2018) learns from data how to propagate local activations by performing a random walk to the entire object area, predicting semantic affinity between a pair of adjacent image coordinates. The seed-expand-constrain (SEC) model (Kolesnikov and Lampert, 2016) *seeds* weak localization cues, *expands* them with image-level class predictions and *constrains* with a CRF the segmentation to coincide with object boundaries. A similar framework is used by Huang *et al.* (2018b), except that the region expansion to cover the whole objects is done with the Seeded Region Growing algorithm (Adams and Bischof, 1994). The method was extended by Lee *et al.* (2019b) where, instead of CAM, they rely on Grad-CAM (Selvaraju *et al.*, 2017) to generate and combine a variety of localization maps obtained with random combinations of hidden units. Redondo-Cabrera *et al.* (2018) combine two Siamese CAM modules to get activation masks that cover full objects and a segmenter network which learns to segment the images according to these activation maps.

Roy and Todorovic (2017) propose to train a CRF-RNN (Zheng *et al.*, 2015) where, for each object class, bottom-up segmentation maps – obtained from the coarse heatmaps – are combined with top-down attention maps and, to improve the object boundaries, refined in the CRF-RNN over iterations. Wang *et al.* (2018b) mine common object features from the initial rough localizations and expand object regions with the mined features. To supplement non-discriminative regions, saliency maps are then considered to refine the object regions.

Kwak *et al.* (2017) propose a Super-pixel Pooling Network, which utilizes super-pixel segmentation as a pooling layout to reflect low-level image structure, and use them within deCoupledNet (Hong *et al.*, 2015) to learn semantic segmentation. The WILDCAT model (Durand *et al.*, 2017) is based on FCN, where all regions are encoded into multiple class modalities with a multi-map transfer layer, and pooled separately for

each class to obtain class-specific heatmaps. Sun *et al.* (2020a) propose two complementary neural co-attention models to capture the shared and unshared objects in paired training images.

Several methods consider adding a separate localization branch that performs the object detection and thus helps adjust the output of the segmentation branch. In this spirit, Qi *et al.* (2016) select positive and negative proposals from the predicted segmentation maps for training the object localization branch and uses an aggregated proposal to build pseudo labeled segmentation to train the segmentation branch.

Another group of weakly-supervised SiS methods considers that the object bounding boxes in an image are available and obtained either manually, as much less costly than pixel-level annotation, or automatically by pretrained object detectors such as R-CNNs (Girshick *et al.*, 2014). As such, Xia *et al.* (2013) introduce a simple voting scheme to estimate the object's shape in each bounding box using a subsequent graph-cut-based figure-ground segmentation. Then, they aggregate the segmentation results in the bounding boxes to obtain the final segmentation result.

Dai *et al.* (2015) iterate between 1) automatically generating segmentation masks, and 2) training an FCN segmentation network under the supervision of these approximate masks. The segmentation masks are obtained with *multi-scale combinatorial grouping* (MCG) (Pont-Tuset *et al.*, 2016) of unsupervised region proposals (Arbelaez *et al.*, 2011). A similar approach has been proposed by Khoreva *et al.* (2017) who combine MCG with a modified GrabCut (Rother *et al.*, 2004) to train a DeepLab model. Ji and Veksler (2021) train a per-class CNN using the bounding box annotations to learn the object appearance and to segment these bounding boxes into object-class versus background labels. These bounding box segments are then combined to get pseudo-labeled image segmentations which can be used to train a DeepLab model.

Instead, Song *et al.* (2019) train an FCN model with a box-driven class-wise masking model to generate class-aware masks, and rely on the mean filling rates of each class as prior cues. Kulharia *et al.* (2020) learn pixel embeddings to simultaneously optimize high intra-class feature affinity and increasing discrimination between features across different classes. The model uses per-class attention maps that saliently guides

the per-pixel cross entropy loss to focus on foreground pixels and to refine the segmentation boundaries.

More recent methods rely on the effectiveness of transformer networks (see also Section 1.2.9) to generate high-quality localization for different semantic classes (class aware CAMs) that can be used to generate pseudo labels for supervising the segmentation network. However, as discussed by Gao *et al.* (2021) the attention maps of visual transformers are in general semantic-agnostic (not distinguishable to object classes) and therefore are not competent to semantic-aware localization. They propose instead the Token Semantic Coupled Attention Map, that relies on a semantic coupling module which combines the semantic-aware tokens with the semantic-agnostic attention map. Similarly, Xu *et al.* (2022) exploit class-specific transformer attentions and develop an effective framework to learn class-specific localization maps from the class-to-patch attention of different class tokens. Instead, Ru *et al.* (2022) propose an *Affinity from Attention* module to learn semantic affinity from the multi-head self-attention and a *Pixel-Adaptive Refinement* of the initial CAM based pseudo labeling via a random walk process.

Amongst other types of weak annotations, we can mention *scribble* supervision (Lin *et al.*, 2016; Vernaza and Chandraker, 2017; Tang *et al.*, 2018) and the even more constrained *point supervision* (Bearman *et al.*, 2016), where a single pixel from each class is manually annotated in every image. Xu *et al.* (2015) design a unified framework to handle different types of weak supervision (image-level, bounding boxes and scribbles), formulating the problem as a max margin clustering, where supervision comes as additional constraints in the assignments of pixels to class labels.

Crawling the web is another source of weak image supervision. Jin *et al.* (2017) use images with simple background – crawled from the web – to train shallow CNNs to predict class-specific segmentation masks, which then are assembled into one deep CNN for end-to-end training. Shen *et al.* (2017) use a large scale co-segmentation framework to learn an initial dilated FCN segmentation model which is refined using pseudo-labeled masks and image-level labels of webly crawled images.

Hong *et al.* (2017) propose to crawl the web for video sequences and to exploit relevant spatio-temporal volumes within the retrieved videos. In the method proposed by Fan *et al.* (2018), images are fed into a salient instance detector to get proxy ground-truth data and to train DeepLab for segmentation, and respectively, Mask R-CNN (He *et al.*, 2017) for instance segmentation. Shen *et al.* (2018) rely on two SEC (Seed-Expand-Constrain) models (Kolesnikov and Lampert, 2016) where an initial SEC model – trained on weakly labeled target data – is used to filter images from the web, a second SEC model learns from these weakly labeled images. Note that many weakly supervised methods rely on Grab-Cut (Rother *et al.*, 2004) to improve bounding box binary segmentations, and on CRFs (Section 1.1.2) to refine the final segmentations.

1.3.3 Curriculum learning based SiS

Curriculum learning (Bengio *et al.*, 2009) refers to the practice where the training process first approaches *easier tasks* and then progressively solve the *harder tasks*. Soviany *et al.* (2021) classify the curriculum learning methods into data-level and model-level curriculum learning, where the former group ranks the training samples/tasks from easy to hard and a special module selects which examples/tasks should be considered at a given training step, while the latter group starts with a simpler model and increases progressively the model capacity. Curriculum based SiS methods belong in general to the former group.

In this setup, Kumar *et al.* (2011) propose to use self-paced learning (SPL) algorithm for object segmentation which chooses a set of easy images at each iteration to update the model. Zang *et al.* (2017) incorporate the SPL into a fine-tuning process for object segmentation in videos. The model learns over iterations from easy to complex samples in a self-paced fashion thus allowing the model to cope with data ambiguity and complex scenarios.

Feng *et al.* (2020) propose an easy-to-hard curriculum self-training approach for semi-supervised SiS where the number of confident pseudo-labels selected from each class is progressively increased where more difficult (lower confidence) samples are added at later phases. Jesson *et al.* (2017) combine curriculum learning with hard negative mining for

lung nodules segmentation where the model initially learns how to distinguish nodules from their immediate surroundings and then continuously increases the proportion of difficult-to-classify global context.

1.3.4 Class-incremental learning for SiS

Class-incremental learning is a branch of continual learning where the goal is to extend the underlying knowledge of a model to new classes. In general, the assumption is to have a model trained on an initial class-set which at different stages is fine-tuned on different new data, where images contain annotation for one or more new classes (see Figure 1.12).

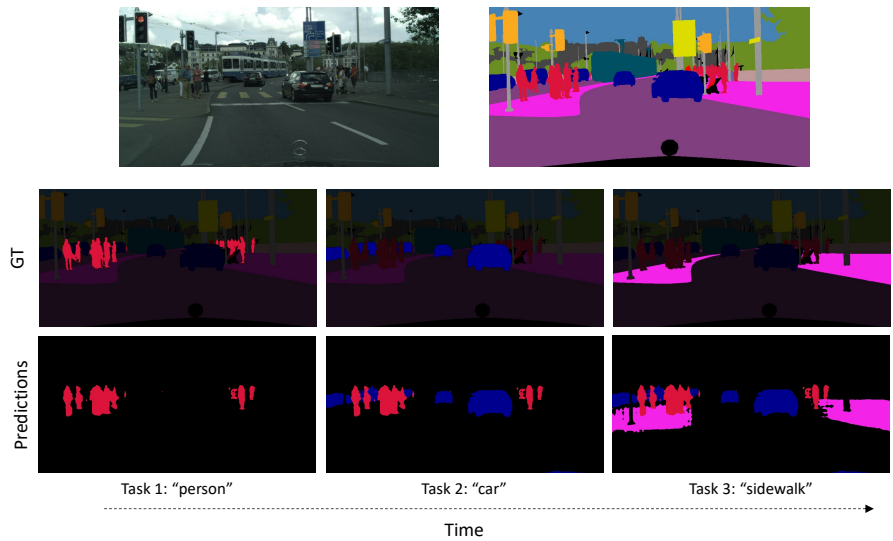

Figure 1.12: Class-incremental learning for SiS. In standard semantic segmentation, models are trained by relying on datasets where each image is annotated with a mask covering all the classes of interest *(top)*. Instead, in class-incremental learning the goal is to extend the underlying knowledge of a model in a sequence of steps, where at each step only one/a few classes are annotated – and the rest is "background". In the example in *(middle)*, the model is extended with the classes "person", "car" and "sidewalk" in three different steps. After each stage, the model can predict more classes *(bottom)*.

The class-incremental learning problem has a long history in image classification; yet, this problem has been only recently addressed in the SIS context (Cermelli *et al.*, 2020; Michieli and Zanuttigh, 2021;

Douillard *et al.*, 2021; Maracani *et al.*, 2021; Cha *et al.*, 2021). Most works focus on the scenario where the original dataset on which the model was trained is not available, and propose to fine-tune the model on samples available for the current new class (Cermelli *et al.*, 2020; Michieli and Zanuttigh, 2021; Douillard *et al.*, 2021; Maracani *et al.*, 2021) – without storing samples from the new classes over time. This is a realistic scenario under the assumption that the model has been trained by a third-party and therefore one does not have access to the original training set or the access is prohibited by the rights to use the original training samples for copyright issues. Recently, in contrast to the above methods, Cha *et al.* (2021) have proposed a memory-based approach to class-incremental SiS.

In class-incremental learning – and in continual learning in general – a major challenge that the models need to face is to avoid *catastrophic forgetting*. While learning on new samples – where only the new classes are annotated (and the rest of the image is considered as background) – the model may overfit on them and the performance on the previous ones may decrease. In the context of SiS, this is typically referred to as *background shift* (Cermelli *et al.*, 2020), given that pixel annotation from previously learned classes are in general annotated as "background" in the new samples (see Figure 1.12).

Cermelli *et al.* (2020) propose to tackle the background shift via distillation, in order to avoid forgetting previously learned categories. In addition, they propose an initialization strategy, devised *ad hoc*, to mitigate the vulnerability of SiS class-incremental learners to the background shift issue. Douillard *et al.* (2021) mitigate the catastrophic forgetting by generating pseudo-labels for old classes where the confidence of the pseudo ground-truth is weighed via its entropy. They further rely on a distillation loss that preserves short- and long-distance relationships between different elements in the scene.

Instead, Maracani *et al.* (2021) propose two strategies to recreate data that comprise the old classes; the first one relies on GANs and the second exploits images retrieved from the web. Cermelli *et al.* (2022) propose a class-incremental learning pipeline for SiS where new classes are learned by relying on global image-level labels instead of pixel-level annotations, hence related to weekly supervised SiS (see Section 1.3.2).

Differently from previous approaches, Cha *et al.* (2021) consider a memory bank in which a few hundred past samples are stored to mitigate forgetting. Furthermore, they propose to model the "unknown" classes other than the "background" one, which further helps avoiding forgetting and preparing the model to learn future classes.

Finally, Cermelli *et al.* (2021) introduce a new task called Incremental Few-Shot Segmentation (iFSS), where the goal is class-incremental learning by relying on few samples for each new class. They propose a method that can learn from only a few samples while at the same time avoiding catastrophic forgetting. This is done by relying on a prototype-based distillation and on batch renormalization (Ioffe, 2021) to handle *non-iid* data.

1.3.5 Self-supervised SiS

Under the shortage of human annotations, self-supervised learning represents another alternative to learn effective visual representations. The idea is to devise an auxiliary task, such as rotation prediction (Gidaris *et al.*, 2018), colorization (Zhang *et al.*, 2016a), or contrastive learning (Chen *et al.*, 2020b), and to train a model for this task instead of a supervised one.

Zhan *et al.* (2017) are the first to apply self-supervised learning in the SiS context; they propose *Mix-and-Match*, where in the *mix* stage sparsely sampled patches are mixed, and in the *match* stage a class-wise connected graph is used to derive a strong triplet-based discriminative loss for fine-tuning the network.

Most recently, motivated by the success of BERT (Devlin *et al.*, 2019) in NLP and by the introduction of Vision Transformers (ViT) (Dosovitskiy *et al.*, 2021), a variety of masked image models for self-supervised pre-training has been proposed. Aiming to reconstruct masked pixels (El-Nouby *et al.*, 2021; He *et al.*, 2022; Xie *et al.*, 2022b), discrete tokens (Bao *et al.*, 2022; Zhou *et al.*, 2022) or deep features (Baevski *et al.*, 2022; Wei *et al.*, 2021), these methods have demonstrated the ability to scale to large datasets and models and achieve state-of-the-art results on various downstream tasks, including SiS. In particular, the masked autoencoder (MAE) (He *et al.*, 2022) accelerates pre-training by

using an asymmetric architecture that consists of a large encoder that operates only on unmasked patches followed by a lightweight decoder that reconstructs the masked patches from the latent representation and mask tokens. MultiMAE (Bachmann *et al.*, 2022) leverages the efficiency of the MAE approach and extends it to multi-modal and multitask settings. Rather than masking input tokens randomly, the Masked Self-Supervised Transformer model (MST) (Li *et al.*, 2021c) proposes to rely on the attention maps produced by a teacher network, to dynamically mask low response regions of the input, and a student network is then trained to reconstruct it.

Instead, Fang *et al.* (2022) propose the Corrupted Image Modeling (CIM) for self-supervised visual pre-training. CIM uses an auxiliary generator to corrupt the input image where some patches are randomly selected and replaced with plausible alternatives. Given such a corrupted image, an enhancer network learns to either recover all the original image pixels, and to predict whether a visual token is replaced by a generator sample or not. After pre-training, the enhancer can be used as a high-capacity visual encoder that achieves compelling results in image classification and semantic segmentation.

The model fine-tuning in self-supervised SiS makes a transition towards the domain adaptation that we present in detail in the next section. We consider both fine-tuning and domain adaptation as instances of transfer learning. Fine-tuning uses the pre-trained models to initialize the target model parameters and update these parameters during training. In general, it requires labeled data from the target domain and does not use data from the source domain. Also, the target task is often different from the source task. Instead, domain adaptation is the process of adapting model(s) trained on source domain(s), by transferring information to improve model performance on the target domain(s). In general, labels are not available in the target set but source and target are associated with the same task. However, note that – as we will see in Section 2.4 – most recent DA models go beyond classical DA, thus making the distinction between the two approaches more subtle.

2

Domain Adaptation for SiS (DASiS)

The success of deep learning methods for SiS discussed in Section 1.2 typically depends on the availability of large amounts of annotated training data. Manual annotation of images with pixel-wise semantic labels is an extremely tedious and time consuming process. Progress in computer graphics and modern high-level generic graphics platforms, such as game engines, enable the generation of photo-realistic virtual worlds with diverse, realistic, and physically plausible events and actions. The computer vision and machine learning communities realized that such tools can be used to generate datasets for training deep learning models (Richter *et al.*, 2016). Indeed, such synthetic rendering pipelines can produce a virtually unlimited amount of labeled data, leading to good performance when deploying models on real data, due to constantly increasing photorealism of the rendered datasets. Furthermore, it becomes easy to diversify data generation; for example, when generating scenes that simulate driving conditions, one can simulate seasonal, weather, daylight or architectural style changes, making such data generation pipeline suitable to support the design and training of computer vision models for diverse tasks, such as SiS.

While some SiS models trained on simulated images can already perform relatively well on real images, their performance can be further improved by domain adaptation (DA) – and in particular *unsupervised domain adaptation* (UDA) – by bridging the gap caused by the domain shift between the synthetic and real images. For the aforementioned reasons, sim-to-real adaptation represents one of the leading benchmarks to assess the effectiveness of domain adaptation for semantic image segmentation.

The main goal of DASiS is to ensure that SiS models trained on synthetic images perform well on real target data, by leveraging annotated synthetic and non-annotated real data. A classical DASiS framework relies on either SYNTHIA (Ros *et al.*, 2016) or GTA (Richter *et al.*, 2016) dataset as a source, and the real-world Cityscapes (Cordts *et al.*, 2016) dataset as a target. Some known exceptions include domain adaptation between medical images (Bermúdez-Chacón *et al.*, 2018; Perone *et al.*, 2019), aerial images (Lee *et al.*, 2021), weather and seasonal condition changes of outdoor real images (Wulfmeier *et al.*, 2017), and adaptation between different Field of View (FoV) images (Gu *et al.*, 2021).

Early DASiS methods have been directly inspired by adaptation methods originally designed for image classification (Csurka, 2020; Wang and Deng, 2018). However, SiS is a more complex task as predictions are carried out at the pixel level, where neighbouring pixels are strongly related (as discussed in Section 1). DA methods for image classification commonly embed entire images in some latent space and then align source and target data distributions. Directly applying such a strategy to SiS models is sub-optimal, due to the higher dimensionality and complexity of the output space. To address this complexity, most DASiS methods take into account the spatial structure and the local image context, act at multiple levels of the segmentation pipeline and often combine multiple techniques.

Therefore, to overview these methods, we step away from grouping the DA methods into big clearly distinguishable families, as it is done in recent surveys on image classification (Csurka, 2020; Wang and Deng, 2018). We instead identify a number of critical characteristics of existing DASiS pipelines and categorize the most prominent methods according to them. From this point of view, Table 2.1 is one of our major

Table 2.1: Summary of the DASiS methods, according to their characteristics. Segmentation Network: The neural network used as a backbone, Image Level: alignment at image level (by using style transfer), from source to target S→T, target to source S←T or both S↔T. Network Level: alignment at feature level. Shared: the parameters of the segmentation network are shared (✓) or at least partially domain specific (-). Output Level: alignment or regularization at output level. Specificity reports important elements that are not covered by the previously described common characteristics. Acronyms and abbreviations defined on p.111-116.

Citation	Segm. Net	Image Level	Net Level	Shared	CW feat.	Output Level	Complementary	Specificity
Hoffman *et al.* (2016)	FCNs	-	mDC	✓	✓	mInstLoss	-	class-size hist transfer
Chen *et al.* (2017a)	dFCN	-	DC	✓	✓	-	SelfT	static obj prior
Perone *et al.* (2019)	UNet	Aug	EMA	-	-	SemCons	SelfEns	Dice loss/medical
Chen *et al.* (2018c)	DLab/PSPN	-	DC	-	-	-	Distill	spatial aware adaptation
Huang *et al.* (2018a)	ENet	-	DC	-	-	-	-	Jensen-Shannon divergence
Hong *et al.* (2018a)	FCNs	-	DC	✓	✓	-	-	CGAN/target-like features
Hoffman *et al.* (2018b)	FCN	S↔T	DC	-	-	SemCons	-	CycleGAN
Li *et al.* (2018d)	UNet	S↔T	-	-	✓	-	-	PatchGAN/semantic-aware gradient
Murez *et al.* (2018)	dFCN	S↔T	DC	-	-	SemCons	-	dilated DenseNet
Saito *et al.* (2018b)	FCN	-	-	✓	-	MCD	CoT	minimize/maximize discrepancy
Saito *et al.* (2018a)	FCN	-	-	✓	-	MCD	CoT	adversarial drop-out
Tsai *et al.* (2018)	DeepLab	-	-	✓	-	mDC	-	multi-level predictions
Wu *et al.* (2018)	FCN	S→T	DM	✓	-	-	-	channel-wise Gramm-matrix align
Zhu *et al.* (2018)	FCN	S↔T	-	✓	-	-	-	conservative loss
Zou *et al.* (2018)	FCN	-	-	✓	-	-	SelfT	self-paced curriculum/spacial priors
Chang *et al.* (2019a)	DeepLab	S↔T	-	-	-	DC	-	domain-specific encoders/percL
Chen *et al.* (2019c)	FCN/DRN	S↔T	DC	-	-	SemCons	-	KL cross-domain consistency
Chen *et al.* (2019a)	DeepLab	-	-	✓	-	-	SelfT	max squares loss/img-wise weights
Choi *et al.* (2019)	ASPP	S→T	EMA	-	-	SemCons	SelfEns	target-guided+cycle-free augm.
Du *et al.* (2019)	FCN	-	DC	✓	✓	-	PL	class-wise adversarial reweighting
Lee *et al.* (2019a)	PSPNet	-	-	✓	-	MCD	CoT	sliced Wasserstein discrepancy

Citation	Segm. Net	Image Level	Net Level	Sha-red	CW feat.	Output Level	Comple-mentary	Specificity
Luo *et al.* (2019b)	FCN/DLab	-	-	✓	-	MCD	CoT	local consistency/self-adaptive weight
Li *et al.* (2019c)	DeepLab	S↔T	-	✓	-	DC	SelfT	perceptual loss (percL)
Lian *et al.* (2019)	FCN/PSPN	-	-	✓	-	-	CurrL	self-motivated pyramid curriculum
Luo *et al.* (2019a)	FCN/DLab	-	DC	✓	-	-	-	signif.-aware information bottleneck
Shen *et al.* (2019)	ASPP	-	mDC	✓	✓	ConfMap	SelfT	cls+adv-conf./class-balance weighs
Xu *et al.* (2019b)	DeepLab	Aug	EMA	-	-	SemCons	SelfEns	self-ensembling attention maps
Vu *et al.* (2019a)	DLab/ASPP	-	-	✓	-	DC	TEM	entropy map align./class-ratio priors
Huang *et al.* (2020a)	DeepLab	-	-	✓	✓	DC	TEM	local contextual-relation
Lv *et al.* (2020)	FCN/DLab	-	-	✓	-	SemCons	SelfT	course-to-fine segm. interaction
Musto and Zinelli (2020)	FCN/DLab	S↔T	-	✓	-	SemCons	SelfT	spatially-adaptive normalization
Pan *et al.* (2020)	DeepLab	-	-	-	-	DC	CurrL	align entropy maps/easy-hard split
Toldo *et al.* (2020b)	DeepLab	S↔T	DC	✓	-	SemCons	-	MobileNet
Wang *et al.* (2020f)	ASPP	-	DC	✓	✓	-	SelfT	disentangled things and stuff
Yang *et al.* (2020d)	Unet	S↔T	-	✓	-	-	SelfT	phase cons./cond. prior network
Yang and Soatto (2020)	FCN/DL	S→T	-	✓	-	-	SelfT	Fourier transform/low-freq. swap
Yang *et al.* (2020a)	DeepLab	-	advF	✓	-	DC	TEM	adv. attack/feature perturbation
Yang *et al.* (2020b)	FCN/DLab	T→S	-	✓	-	DC	SelfT	reconstruction from predictions
Zhang *et al.* (2020c)	FCN	-	DC	✓	-	LocCons	-	patch+cluster+spatial consistency
Zheng and Yang (2020)	PSPNet	-	-	✓	-	MCD	CoT	memory regularization (KL)
Araslanov and Roth (2021)	DeepLab	Aug	EMA	-		SemCons	SelfT	self-sup/imp. sampling/focal loss
Cheng *et al.* (2021)	DLab/FCN	S↔T	-	-	-	SemCons	SelfT	dual perceptual loss/dual path DASS
Guo *et al.* (2021b)	DeepLab	-	-	✓	-	-	SelfT	meta-learning/meta-loss correction
Toldo *et al.* (2021)	DeepLab	-	Clust.	✓	✓		TEM	discriminative clustering
Truong *et al.* (2021)	DeepLab	-	-	✓	-	-	TEM	bij. max. likelihood/local consistency
Wang *et al.* (2021d)	DeepLab	S→T	-	✓	-	DC	SelfT	target-guided uncertainty rectifying
Wang *et al.* (2021b)	FCN	S↔T	EMA	-	-	-	SelfEns	AdaIn/class-balanced reweighting
Yang *et al.* (2021)	DeepLab	S→T	-	✓	-	contrL	SelfT	adv. attack/adv. self-supervised loss
Chen *et al.* (2022)	SwinTr	-	DC	✓	✓	-	Distill	Momentum Transformer

contributions. It is detailed in Section 2.2, where we describe the different domain alignment techniques that are applied at input image, feature and output prediction levels. In Section 2.3, we describe complementary machine learning strategies that can empower domain alignment and improve the performance of a segmentation model on target images. Before presenting these different methods, in the next section we first formalize the UDA problem and list the most popular domain alignment losses optimized by a large majority of DA approaches.

2.1 Brief Introduction into UDA

Let $\mathcal{D}_\mathcal{S} = \mathcal{X}_\mathcal{S} \times \mathcal{Y}_\mathcal{S}$ be a set of paired sample images with their ground-truth annotated segmentation maps ($\mathcal{X}_\mathcal{S} = \{\mathbf{x}_i\}_{i=1}^M$ and $\mathcal{Y}_\mathcal{S} = \{\mathbf{y}_i\}_{i=1}^M$, respectively), drawn from a source distribution $P_\mathcal{S}(\mathcal{X}, \mathcal{Y})$. In the SiS context, $\mathbf{x}$ and $\mathbf{y}$ represent images and their pixel-wise annotations, respectively, $\mathbf{x} \in \mathcal{R}^{H \times W \times 3}$ and $\mathbf{y} \in \mathcal{R}^{H \times W \times C}$, where (H, W) is the image size and C is the number of semantic categories. Let $\mathcal{D}_\mathcal{T} = \mathcal{X}_\mathcal{T} = \{\mathbf{x}_i\}_{i=1}^N$ be a set of unlabeled samples drawn from a target distribution $P_\mathcal{T}$, such that $P_\mathcal{S} \neq P_\mathcal{T}$ due to the domain shift. In the UDA setup, both sets are available at training time ($\mathcal{D} = \mathcal{D}_\mathcal{S} \cup \mathcal{D}_\mathcal{T}$) and the goal is to learn a model performing well on samples from the target distribution.

Often, the segmentation network used in DASiS has an encoder-decoder structure (see Figure 2.1) and the domain alignment can happen at different levels of the segmentation network, including the output of the encoder, at various level of the decoder, or even considering the label predictions as features (as discussed later in this section). Hence, the features used to align the domains can be extracted at image level, region level or pixel-level. Therefore, when we use the notation of F_S and F_T to refer to any of the above source respectively target feature generator. Note that it is frequent that the feature encoders F_S and F_T share their parameters θ_{F_S} and θ_{F_T} – in this case, we simply refer to them as F and θ_F.

In the following, we will cover some basic components and commonly used losses that constitute the foundation of most UDA and DASiS approaches.

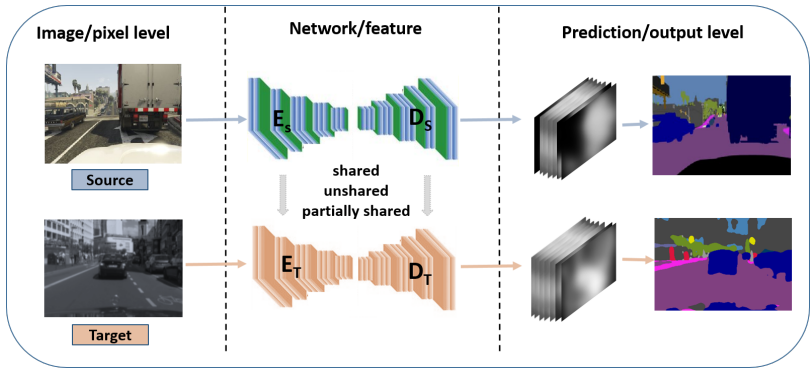

Figure 2.1: Several DASiS models adopt a Siamese architecture with one branch per domain, where the alignment is carried out at different levels of the pipeline: at *pixel level*, by transferring the image style from one domain to another; at *network level*, by aligning the image features, derived from activation layers and sharing (full or partial) the network parameters; and finally at the *output level*, by aligning the class prediction maps.

Distribution Discrepancy Minimization. In image classification, one popular approach to address UDA is to minimize the distribution discrepancy between source and target domains in some latent feature space – that is, the space of the learned visual representation. This space often corresponds to some activation layers; most often the last layer the classifier is trained on, but other layers can be considered as well. One popular measure is the empirical Maximum Mean Discrepancy (MMD) (Borgwardt *et al.*, 2006), that is written as

$$\mathcal{L}_{mmd} = \left\| \frac{1}{M} \sum_{\mathbf{x}_s \in \mathcal{X}_S} \phi(F_S(\mathbf{x}_s)) - \frac{1}{N} \sum_{\mathbf{x}_t \in \mathcal{X}_T} \phi(F_T(\mathbf{x}_t)) \right\| ,$$

where ϕ is the mapping function corresponding to a Reproducing Kernel Hilbert Space (RKHS) kernel defined as a mixture of Gaussian kernels.

Adversarial Training. An alternative to minimizing the distribution discrepancy between source and target domains is given by adversarial training (Goodfellow *et al.*, 2014). Multiple studies have shown that domain alignment can be achieved by learning a domain classifier C_{Disc} (the *discriminator*) with the parameters θ_D to distinguish between the feature vectors from source and target distributions and by using an

adversarial loss to increase *domain confusion* (Ganin *et al.*, 2016; Tzeng *et al.*, 2015; Tzeng *et al.*, 2017). The main, task-specific deep network – in our case SiS – aims to learn a representation that fools the domain classifier, encouraging encoders to produce domain-invariant features. Such features can then be used by the final classifier, trained on the source data, to make predictions on the target data. Amongst the typical adversarial losses, we mention the *min-max game* proposed by Ganin *et al.* (2016)

$$\mathcal{L}_{adv} = \min_{\theta_F, \theta_C} \max_{\theta_D} \{\mathbb{E}_{\mathbf{x}_s \in \mathcal{X}_\mathcal{S}}[\mathcal{L}_{Task}(F(\mathbf{x}_s), \mathbf{y}_s)] - \lambda \cdot \mathbb{E}_{\mathbf{x} \in \mathcal{X}_\mathcal{S} \cup \mathcal{X}_\mathcal{T}}[\mathcal{L}_{Disc}(F(\mathbf{x}), \mathbf{y}_d)]\},$$

where $\mathcal{L}_{Task}$ is the loss associated with the task of interest (it depends on both the feature encoder parameters θ_F and the final classifier's parameters θ_C), $\mathcal{L}_{Disc}$ is a loss measuring how well a discriminator model parametrized by θ_D can distinguish whether a feature belongs to source ($\mathbf{y}_d = 1$) or to target domain ($\mathbf{y}_d = 0$), and λ is a trade-off parameter. By alternatively training the discriminator C_{Disc} to distinguish between domains and the feature encoder F to *fool* it, one can learn domain agnostic features. Also, training the encoder and the final classifier C_{task} for the task of interest, guarantees that such features are not simply domain-invariant, but also discriminative.

An effective way to approach this minimax problem consists in introducing a Gradient Reversal Layer (GRL) (Ganin *et al.*, 2016) which reverses the gradient direction during the backward pass in backpropagation (in the forward pass, it is inactive). The GRL allows to train the discriminator and the encoder at the same time.

A related but different approach by Tzeng *et al.* (2017) brings adversarial training for UDA closer to the original GAN formulation (Goodfellow *et al.*, 2014). It splits the training procedure into two different phases: a fully discriminative one, where a module is trained on source samples, and a fully generative one, where a GAN loss is used to learn features for the target domain that mimic the source ones – or, more formally, that are projected into the same feature space, on which the original classifier is learned. This second step can be carried out by approaching the following minimax game

$$\mathcal{L}_{GAN} = \min_{\theta_{F_\mathcal{T}}, \theta_{F_\mathcal{S}}} \max_{\theta_D} \{\mathbb{E}_{\mathbf{x}_s \in \mathcal{X}_\mathcal{S}}[\log(C_{Disc}(F_S(\mathbf{x}_s)))] + \mathbb{E}_{\mathbf{x}_t \in \mathcal{X}_\mathcal{T}}[\log(1 - C_{Disc}(F_T(\mathbf{x}_t)))]\},$$

where C_{Disc} is the discriminator, and both F_S and F_T are initialized with the weights pre-trained by supervised learning on the source data.

2.2 Adapting SiS between Domains

Since the advent of representation learning solutions in most machine learning applications, UDA research has witnessed a shift towards end-to-end solutions to learn models that may perform well on target samples. In image classification, a very successful idea has been to learn a representation where the source and target samples get *aligned* – that is, the source and target distributions are close in the feature space under some statistical metrics.

This *alignment* is often achieved by means of a Siamese architecture (Bromley *et al.*, 1993) with two streams, each corresponding to a semantic segmentation model: one stream is aimed at processing source samples and the other at processing the target ones (as shown in Figure 2.1). The parameters of the two streams can be shared, partially shared or domain specific; generally, the backbone architectures of both streams are initialized with weights that are pre-trained on the source set. The Siamese network is typically trained with a loss comprising two terms. For what concerns SiS, one term is the standard *pixel-wise cross-entropy* loss (referred in this paper also as $\mathcal{L}_{Task}$), measuring performance on source samples for which the ground-truth annotations are available

$$\mathcal{L}_{ce} = -\mathbb{E}_{(\mathcal{X}_\mathcal{S}, \mathcal{Y}_\mathcal{S})} \left[\sum_{h,w,c} \mathbf{y}_s^{(h,w,c)} \cdot \log(p^{(h,w,c)}(F_S(\mathbf{x}_s))) \right],$$

where $p^{(h,w,c)}(F_S(\mathbf{x}_s))$ is a probability of class c at pixel $\mathbf{x}_s^{(h,w)}$ and $\mathbf{y}_s^{(h,w,c)}$ is 1 if c is the pixel's true class and 0 otherwise.

The second term is a *domain alignment* loss that measures the distance between source and target samples. The alignment can be addressed at different levels of the pipeline, as illustrated in Figure 2.1,

namely, network (feature), at image (pixel) and output (prediction) levels, as detailed in Sections 2.2.1, 2.2.2 and 2.2.3, respectively. Note that – as shown in Table 2.1 – many approaches apply alignment at multiple levels.

While aligning the marginal feature distributions tends to reduce the domain gap, it can be sub-optimal as it does not explicitly take the specific task of interest (in this case, SiS) into account during the domain alignment as discussed for example by Zhao *et al.* (2019a). To overcome these weaknesses, several works have been proposed to leverage the class predictions during the alignment, what we call output level alignment (see Section 2.2.3). Furthermore, there is a growing interest for adaptation at pixel-level (see Section 2.2.2), Indeed, the shift between two domains is often strongly related to visual appearance variations such as day *versus* night, seasonal changes, synthetic *versus* real. By exploiting the progress of image-to-image translation and style transfer brought by deep learning-based techniques (Huang and Belongie, 2017; Zhu *et al.*, 2017), several DASiS methods have been proposed to explicitly account for such stylistic domain shifts by performing an alignment at image level.

In the following, we discuss in details alignment solutions between source and target at various levels of the segmentation pipeline.

2.2.1 Feature-level adaptation

Generic DA solutions proposed for image classification perform domain alignment in a latent space by minimizing some distance metrics, – such as the maximum mean discrepancy (MMD) (Long *et al.*, 2015b) between feature distributions of source and target data, – or by adversarially training a domain discriminator to increase *domain confusion* (Ganin *et al.*, 2016; Tzeng *et al.*, 2015; Tzeng *et al.*, 2017). Both approaches scale up to semantic segmentation problems. In particular, adversarial training has been largely and successfully applied and combined with other techniques.

In DASiS, we consider more complex models to tackle the SiS task. We recall that adaptation in SiS is more challenging than in image classification, due to the structural complexity and the scale factor of

the task. Indeed it is rather difficult and sub-optimal to fully capture and handle the DASiS problem by simple alignment of the latent global representations (activation layers) between domains. Therefore, the domain alignment is often carried out at different layers of the network (see Figure 2.2). The alignment is done either by minimizing the feature distribution discrepancy (Bermúdez-Chacón *et al.*, 2018) or by adversarial training via a domain classifier to increase domain confusion (Hoffman *et al.*, 2016; Huang *et al.*, 2018a; Shen *et al.*, 2019).

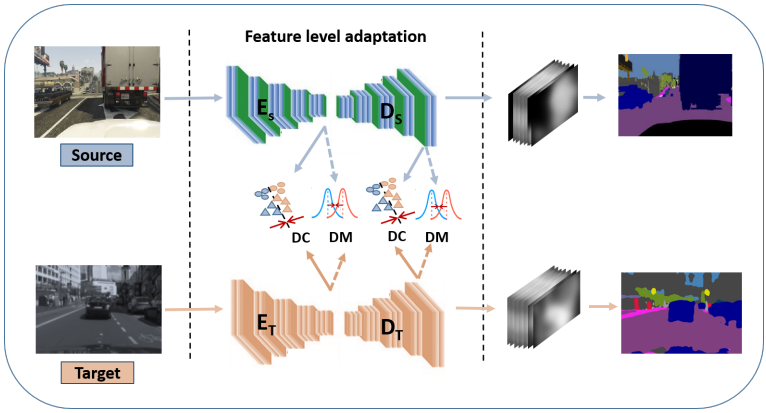

Figure 2.2: In generic DA, domain alignment is often performed in a single latent representation space. In DASiS, the alignment is often done at multiple layers, by discrepancy minimization between feature distributions or by adversarial learning relying on a domain classifier (DC) to increase domain confusion. Encoders and decoders of the segmentation network are often shared: $E_S = E_T$, $D_S = D_T$.

While some works consider the alignment simply a global representation of the image (Huang *et al.*, 2018a) – by flattening or pooling the activation map – most often pixel-wise (Hoffman *et al.*, 2016), grid-wise (Chen *et al.*, 2017a) or region-wise (Zhang *et al.*, 2020c) representations are used. Furthermore, to improve the model performance on the target data, such methods are often combined with some prior knowledge or specific tools as discussed below (and also in Section 2.3).

In their seminal work, Hoffman *et al.* (2016) combine the distribution alignment with the class-aware constrained multiple instance loss used to transfer the spatial layout. Chen *et al.* (2017a) consider global and class-wise domain alignment and address it via adversarial training. In particular, they rely on local class-wise domain predictions over image

grids assuming that the composition/proportion of object classes across domains – different urban environments in their case – is similar.

Hong *et al.* (2018a) rely on a conditional generator that transforms the source features into target-like features, using a multi-layer perceptron as domain discriminator. Assuming that decoding these target-like feature maps preserve the semantics, they are used with the corresponding source labels within an additional cross-entropy loss to make the model more suitable for the target data.

The Pivot Interaction Transfer (Lv *et al.*, 2020) consists in optimizing a semantic consistency loss between image-level and pixel-level semantic information. This is achieved by training the model with both a fine-grained component producing pixel-level segmentation and coarse-grained components generating class activation maps obtained by multiple region expansion units, trained with image-level category information independently. Zhang *et al.* (2020c), to improve alignment, explore three label-free constraints as model regularizer, enforcing *patch-level*, *cluster-level* and *context-level* semantic prediction consistencies at different levels of image formation (see Figure 2.3).

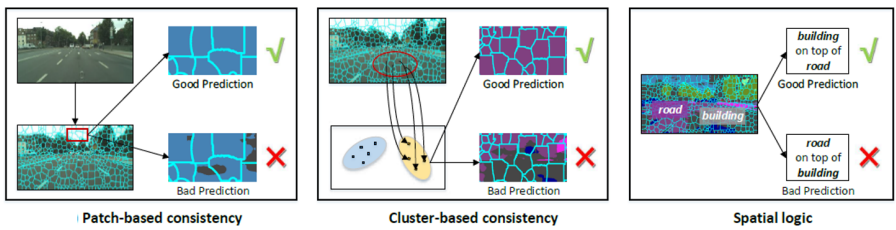

Figure 2.3: To improve the domain alignment, Zhang *et al.* (2020c) propose to reduce patch-level, cluster-level and context-level inconsistencies. Figure based on Zhang *et al.* (2020c).

2.2.2 Image-level adaptation

This class of methods relies on image style transfer (IST) methods, where the main idea is to transfer the *domain style* (appearance) from target to source, from source to target, or considering both (see illustration in Figure 2.4). The *style transferred* source images maintain the semantic

content of the source, and therefore its pixel-level labeling too, while their appearance results more similar to the target style – helping the network to learn a model more suitable for the target domain (Csurka *et al.*, 2017; Thomas and Kovashka, 2019).

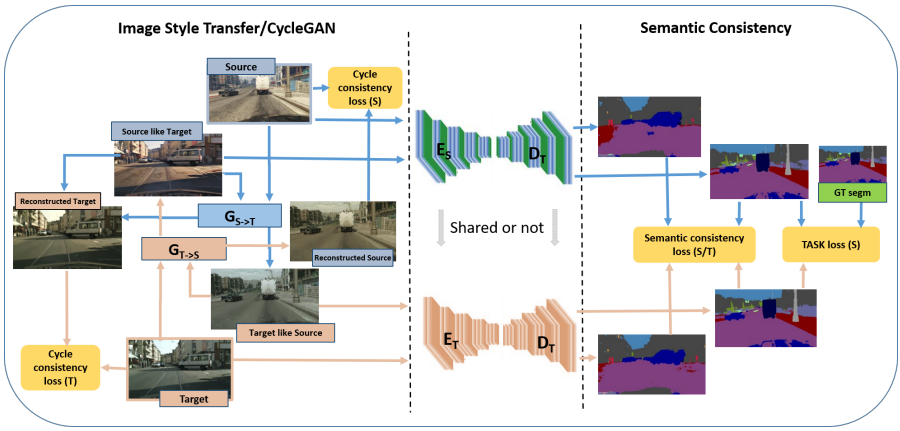

Figure 2.4: In image-level adaptation that relies on image style transfer (IST), the main idea is to translate the *style* of the target domain to the source data and/or the source style to the target domain. In order to improve the style transfer, often the *style transferred* image is translated back to the original domain allowing to use a *cyclic consistency* reconstruction loss. The style transferred source images *inherit* the semantic content of the source and thus its pixel-level labeling, that allows the segmentation network to learn a model suitable for the target domain. On the other hand, the target and the *source-like* target image share the content and therefore imposing that their predicted segmentation should match, – using the *semantic consistency* loss as a regularization, – which helps improving the model performance in the target domain.

Image-to-image translation for UDA has been pioneered within the context of image classification (Bousmalis *et al.*, 2017; Liu and Tuzel, 2016; Taigman *et al.*, 2017); typically, such methods employ GANs (Goodfellow *et al.*, 2014) to transfer the target images' style into one that resembles the source style. This approach has been proved to be a prominent strategy also within DASiS (Chang *et al.*, 2019a; Chen *et al.*, 2019c; Hoffman *et al.*, 2018b; Murez *et al.*, 2018; Toldo *et al.*, 2020b; Sankaranarayanan *et al.*, 2018; Wu *et al.*, 2018; Yang *et al.*, 2020d). Still, as in the case of feature alignment, for a better adaptation most methods combine the image translation with other ingredients (see

also Table 2.1), most often with self-training and different consistency regularization terms (detailed in Section 2.3).

The most used regularization terms in IST based DASiS are the *cycle consistency* loss and the *semantic consistency* loss proposed by Hoffman *et al.* (2018b). The proposed CyCADA is one of the first model that adopted image-to-image translation – and in particular the consistency losses pioneered by Cycle-GAN (Zhu *et al.*, 2017) – for the DASiS problem. The cycle consistency loss is defined as follows

$$\begin{aligned}\mathcal{L}_{cycle} = &\mathbb{E}_{\mathbf{x}_s \sim \mathcal{X}_S}\left[\|G_{T\to S}\left(G_{S\to T}\left(\mathbf{x}_s\right)\right) - \mathbf{x}_s\|_k\right] \\ &+ \mathbb{E}_{\mathbf{x}_T \sim \mathcal{X}_T}\left[\|G_{S\to T}\left(G_{T\to S}\left(\mathbf{x}_t\right)\right) - \mathbf{x}_t\|_k\right].\end{aligned}$$

where $G_{S\to T}$ and $G_{T\to S}$ are the image generators that learn to map the style from source to target and target to source, respectively and $\|\cdot\|_k$ is the Lk loss, where most often the L1 or the L2 loss is used. In short, this loss encourages the preservation of structural properties during the style transfer, while the semantic consistency loss

$$\begin{aligned}\mathcal{L}_{SemCons} = &\mathcal{L}_{Task}\left(F_S(G_{S\to T}(\mathbf{x}_s)), p(F_S(\mathbf{x}_s))\right) \\ &+ \mathcal{L}_{Task}\left(F_S(G_{T\to S}(\mathbf{x}_t)), p(F_S(\mathbf{x}_t))\right),\end{aligned}$$

enforces an image to be labeled identically before and after translation. The task loss $\mathcal{L}_{Task}$ here is the source pixel-wise cross-entropy, but instead of using GT label maps, it is used with the pseudo-labeled predicted maps $p(F_S(\mathbf{x}_s)) = \text{argmax}(F_S(\mathbf{x}_s))$ and $p(F_S(\mathbf{x}_t)) = \text{argmax}(F_S(\mathbf{x}_t))$, respectively.

Inspired by CyCADA, several approaches tried to refine IST for the DASiS problem. Murez *et al.* (2018) propose a method that simultaneously learns domain specific reconstruction with cycle consistency and domain agnostic feature extraction, and learn to predict the segmentation from these agnostic features. In the IST based method proposed by Zhu *et al.* (2018) the classical cross-entropy loss is replaced by a so-called Conservative Loss that penalizes the extreme cases, – for which performance is very good or very bad – enabling the network to find an equilibrium between its discriminative power and its domain-invariance.

Toldo *et al.* (2020b) perform image-level domain adaptation with Cycle-GAN (Zhu *et al.*, 2017) and feature-level adaptation with a consistency loss between the semantic maps. Furthermore, they consider

as backbone a lightweight MobileNet-v2 architecture which allows the model's deployment on devices with limited computational resources such as the ones used in autonomous vehicles.

Li *et al.* (2018d) propose a semantic-aware Grad-GAN that aims at transferring personalized styles for distinct semantic regions. This is achieved by a *soft gradient-sensitive* objective for keeping semantic boundaries, and a *semantic-aware discriminator* for validating the fidelity of personalized adaptions with respect to each semantic region.

The method introduced by Wu *et al.* (2018) jointly synthesizes images and, to preserve the spatial structure, performs segmentation by fusing channel-wise distribution alignment with semantic information in both the image generator and the segmentation network. In particular, the generator synthesizes new images *on-the-fly* to appear target-like and the segmentation network refines the high level features before predicting semantic maps by leveraging feature statistics of sampled images from the target domain.

Chen *et al.* (2019c) rely on both image-level adversarial loss to learn image translation and feature-level adversarial loss to align feature distributions. Furthermore, they propose a bi-directional cross-domain consistency loss based on KL divergence, – to provide additional supervisory signals for the network training, – and show that this yields more accurate and consistent predictions in the target domain.

The Domain Invariant Structure Extraction (DISE) method (Chang *et al.*, 2019a) combines image translation with the encoder-decoder based image reconstruction, where a set of shared and private encoders are used to disentangle high-level, *domain-invariant* structure information from *domain-specific* texture information. Domain adversarial losses and perceptual losses ensure the perceptual similarities between the translated images and their counterparts in the source or target domains. Furthermore, an adversarial loss in the output space ensures domain alignment and therefore generalization to the target.

The approach by Li *et al.* (2019c) relies on *bi-directional* learning, proposing to move from a sequential pipeline – where the SiS model benefits from the image-to-image translation network – to a closed loop, where the two modules help each other. Essentially, the idea is to propagate information from semantic segmentation back to the image transformation network as a semantic consistent regularization.

Cheng *et al.* (2021) consider two image translation and segmentation pipelines from opposite domains to alleviate visual inconsistencies raised by image translation and to promote each other in an interactive manner. The source path assists the target path to learn precise supervision from source data, while the target path guides the source path to generate high quality pseudo-labels for self-training the target segmentation network. Musto and Zinelli (2020) propose a source to target translation model guided by the source semantic map using Spatially-Adaptive (De)normalization (SPADE) (Park *et al.*, 2019) and Instance Normalization layers (Ulyanov *et al.*, 2016).

Yang *et al.* (2020b) introduce a reconstruction network that relies on conditional GANs, which learn to reconstruct the source or the source-like target image from their respective predicted semantic label map. Furthermore, a perceptual loss and a discriminator feature matching loss are used to enforce the semantic consistency between the reconstructed and the original image features.

Some recent works propose IST solutions that do not rely primarily on GANs for image translation. For instance, the innovative approach by Yang and Soatto (2020) relies on the Fourier Transform and its inverse to map the target style into that of the source images, by swapping the low-frequency component of the spectrum of the images from the two domains. The same research team proposes to exploit the phase of the Fourier transform within a consistency loss (Yang *et al.*, 2020d); this guarantees to have an image-to-image translation network that preserves semantics.

2.2.3 Output-level adaptation

To avoid the complexity of high-dimensional feature space adaptation, several papers propose to perform instead adversarial adaptation on the low-dimensional label prediction output space, – defined by the class-likelihood maps (see Figure 2.5). In this case, the pixel-level representations corresponds to the class predictions (forming a C dimensional vector), and in the derived feature space, – similarly to the approaches described in Section 2.2.1, – domain confusion between the domains can be achieved by learning a corresponding domain discriminator. Such

adversarial learning in the output space has been initially proposed by Tsai *et al.* (2019) where they learn a discriminator to distinguish whether the segmentation predictions come from the source or from the target domain. To make the model adaptation more efficient, auxiliary pixel-level semantic and domain classifiers are added at multiple layers of the network, and trained jointly.

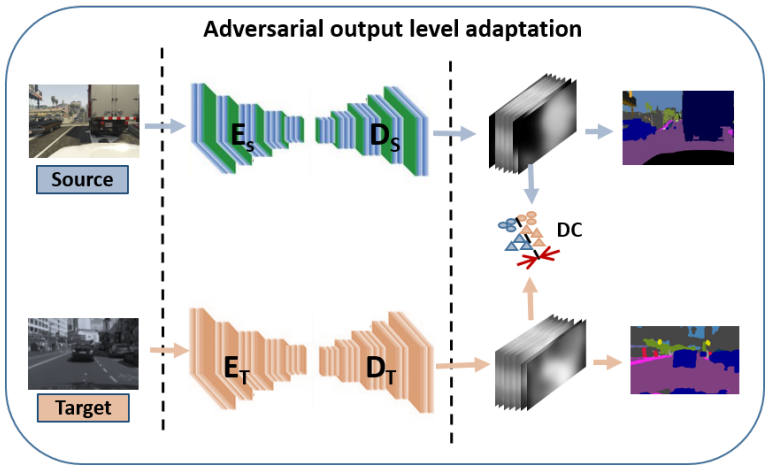

Figure 2.5: Adversarial adaptation on the label prediction output space, where pixel-level representations are derived from the class-likelihood map and used to train the domain classifier.

Vu *et al.* (2019a) first derive the so called *weighted self-information* maps (wSIM) defined as

$$I_x^{(h,w,c)} = -p^{(h,w,c)} \log p^{(h,w,c)}$$

and perform adversarial adaptation on the features derived from these maps. Furthermore, they show that minimizing the sum of these wSIMs is equivalent to direct entropy minimization and train the model jointly with these two complementary entropy-based losses (the direct entropy and the corresponding adversarial loss). Pan *et al.* (2020) instead train a domain classifier on the entropy maps $E_x^{(h,w)} = -\sum_c I_x^{(h,w,c)}$ to reduce the distribution shift between the source and target data.

Output level adversarial learning has often been used in combination with image-level style transfer and self-training (Chang *et al.*, 2019a; Li

et al., 2019c; Wang *et al.*, 2021d) and curriculum learning (Pan *et al.*, 2020) (see also Table 2.1).

2.3 Complementary Techniques

In the previous section, we mainly focused on the core issue of domain alignment; in this section, we discuss other techniques that can be coupled with the DASiS methods previously presented. Generally, they are not explicitly focused on domain alignment, but rather on improving the segmentation model accuracy on the target data. As a part of the transfer learning, DASiS – and UDA in general – possesses characteristics (domain separation, unlabeled target instances, etc.) that encourage researchers to integrate techniques from ensemble, semi- and self-supervised learning, often resulting in their mutual empowering. While being extensively used in UDA research, the methodologies detailed below originated from other branches of machine learning; for example, *self-training with pseudo-labels* (Lee, 2013) and *entropy minimization* (Grandvalet and Bengio, 2004) have been originally formulated for semi-supervised learning; *curriculum learning* has been devised as a stand-alone training paradigm (Bengio *et al.*, 2009); *model distillation* and *self-ensembling* are recent deep learning techniques that allow training more accurate models.

2.3.1 Pseudo-labelling and self-training (SelfT)

Originated from semi-supervised learning, the idea is to generate pseudo-labels for the target data and to refine (self-train) the model over iterations, by using the most confident labels from the target set (Li *et al.*, 2019c; Zou *et al.*, 2018; Kim and Byun, 2020; Li *et al.*, 2019c) (see illustration in Figure 2.6). Indeed, pseudo-labels are often error-prone, so it is important to select the most reliable ones and to progressively increase the set of pseudo-labels as the training progresses. To this end, different works have been proposed that extrapolate the pseudo-label confidence relying on the maximum class probability (MCP) of the model output (Li *et al.*, 2020a; Wang *et al.*, 2020f; Zou *et al.*, 2018) or on the entropy of the softmax predictions (Saporta *et al.*, 2020).

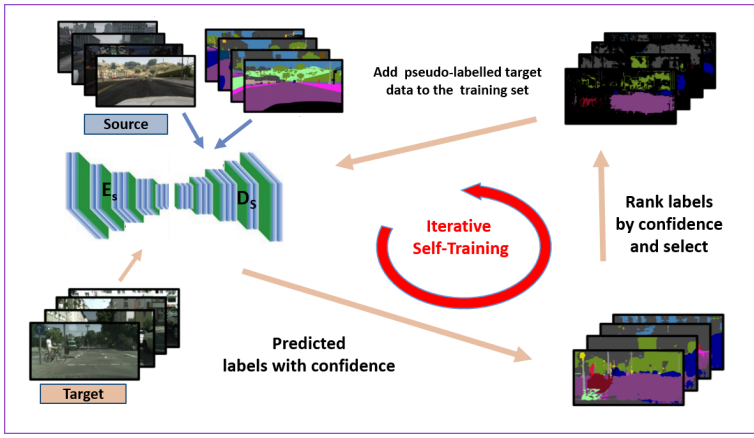

Figure 2.6: Pseudo-labeling. Originated from semi-supervised learning, the idea is to generate pseudo-labels for the target data and to refine the model over iterations, by using the most confident labels from the target set.

Pseudo-labels for which the prediction is above a certain threshold are assumed to be reliable; vice-versa, values below the threshold are not trusted. As shown in Table 2.1, self-training is one of the most popular complementary methods combined with domain alignment techniques.

In the following, we list a few of such methods with their particularities, aimed at further improving the effectiveness of self-learning. To avoid the gradual dominance of large classes on pseudo-label generation, Zou *et al.* (2018) propose a class balanced self-training framework and introduce spatial priors to refine the generated labels. Chen *et al.* (2017a) rely on static-object priors estimated for the city of interest by harvesting the Time-Machine of Google Street View to improve the soft pseudo-labels required by the proposed class-wise adversarial domain alignment. Kim and Byun (2020) introduce a texture-invariant pre-training phase; in particular, the method relies on image-to-image translation to learn a better performing model at a first stage, which is then adapted via the pseudo-labeling in a second stage.

In order to regularize the training procedure, Zheng and Yang (2020) average predictions of two different sets from the same network. Relatedly, Shen *et al.* (2019) combine the output of different discriminators with the confidence of a segmentation classifier, in order to increase the

reliability of the pseudo-labels. To rectify the pseudo-labels, Zheng and Yang (2021) propose to explicitly estimate the prediction uncertainty during training. They model the uncertainty via the prediction variance and integrate the uncertainty into the optimization objective.

Corbière *et al.* (2021) propose an auxiliary network to estimate the *true-class probability map* for semantic segmentation and integrate it into an adversarial learning framework to cope with the fact that the predicted true-class probabilities might suffer from the domain shift. The confidence branch has a multi-scale architecture based on ASPP, allowing the network to better cope with semantic regions of variable size in the image.

Differently, Mei *et al.* (2020) propose an *instance adaptive* framework where pseudo-labels are generated via an *adaptive selector*, namely a confidence-based selection strategy with a confidence threshold that is adaptively updated throughout training. Regularization techniques are also used to respectively smooth and sharpen the pseudo-labeled and non-pseudo-labeled regions.

In order to make the self-training less sensitive to incorrect pseudo-labels, Zou *et al.* (2019) rely on *soft* pseudo-labels in the model regularization, forcing the network output to be smooth. Shin *et al.* (2020) propose a pseudo-label densification framework where a sliding window voting scheme is used to propagate confident neighbor predictions. In a second phase, a confidence-based easy-hard classifier selects images for self-training, while a hard-to-easy adversarial learning pushes hard samples to be like easy ones.

Zhang *et al.* (2019) propose a strategy where pseudo-labels are used in both a cross-entropy loss and a *category-wise distance loss*, where class-dependent centroids are used to assign pseudo-labels to training samples. Li *et al.* (2020a) select source images that are most similar to the target ones via *semantic layout matching* and to retain some pixels for the adaptation via *pixel-wise similarity matching*. These pixels are used together with pseudo-labeled target samples to refine the model. Furthermore, entropy regularization is imposed on all the source and target images.

To mitigate low-quality pseudo-labels arising from the domain shift, Tranheden *et al.* (2021) propose to *mix* images from the two domains along with the corresponding labels and pseudo-labels. While training

the model, they enforce consistency between predictions of images in the target domain and images mixed across domains.

Guo *et al.* (2021b) propose to improve the reliability of pseudo-labels via a *meta-correction* framework; they model the noise distribution of the pseudo-labels by introducing a *noise transaction matrix* that encodes inter-class noise transition relationship. The meta-correction loss is further exploited to improve the pseudo-labels via a meta-learning strategy to adaptively distill knowledge from all samples during the self-training process.

Alternatively, pseudo-labels can also be used to improve the model without necessarily using them in a self-training cross-entropy loss. For example, Wang *et al.* (2020f) use pseudo-labels to disentangle source and target features by taking into account regions associated with *things* and *stuff* (Caesar *et al.*, 2018).

Du *et al.* (2019) use separate semantic features according to the downsampled pseudo-labels to build *class-wise confidence map* needed to reweigh the adversarial loss. A progressive confidence strategy is used to obtain reliable pseudo-labels and, in turn, class-wise confidence maps.

2.3.2 Entropy minimization of target predictions (TEM)

Originally devised for semi-supervised learning (Grandvalet and Bengio, 2004), entropy minimization has received a broad recognition as an alternative or complementary technique for domain alignment. Different DASiS/UDA methods extend simple entropy minimization on the target data by applying it jointly with adversarial losses (Du *et al.*, 2019; Vu *et al.*, 2019a) or square losses (Chen *et al.*, 2019a; Toldo *et al.*, 2021).

Vu *et al.* (2019a) propose to enforce structural consistency across domains by minimizing both the conditional entropy of pixel-wise predictions and an adversarial loss that ensures the distribution matching in terms of weighted entropy maps (as discussed in Section 2.2.3). The main advantage of their approach is that computation of the pixel-wise entropy does not depend on any network and entails no overhead.

Similarly, Huang *et al.* (2020a) design an entropy-based minimax adversarial learning scheme to align local contextual relations across

domains. The model learns to enforce the prototypical local contextual relations explicitly in the feature space of a labeled source domain, while transferring them to an unlabeled target domain via backpropagation-based adversarial learning using a Gradient Reversal Layer (GRL) (Ganin *et al.*, 2016).

Chen *et al.* (2019a) show that entropy minimization based UDA methods often suffer from the probability imbalance problem. To prevent the adaptation process from being dominated by the easiest to adapt samples, they propose instead a *class-balanced weighted* maximum squares loss with a linear growth gradient. Furthermore, they extend the model with self-training on low-level features guided by pseudo-labels obtained by averaging the output map at different levels of the network. Toldo *et al.* (2021) integrate this image-wise class-balanced entropy-minimization loss to regularize their feature clustering-based DASiS method. To further enhance the discriminative clustering performance, they introduce an *orthogonality loss* – which force individual representations to be orthogonal, – and a *sparsity loss* to reduce class-wise the number of active feature channels.

The Bijective Maximum Likelihood (BiMaL) loss (Truong *et al.*, 2021) is a generalized form of the adversarial entropy minimization, without any assumption about pixel independence. The BiMaL loss is formed using a maximum-likelihood formulation to model the global structure of a segmentation input, and a bijective function, to map that segmentation structure to a deep latent space. Additionally, an *unaligned domain score* is introduced to measure the efficiency of the learned model on a target domain in an unsupervised fashion.

2.3.3 Curriculum learning (CurrL)

Several papers apply *curriculum* strategy to DASiS; the main idea is to apply simpler, intermediate tasks to determine certain properties of the target domain which allow to improve performance on the main segmentation task. In this regard, Zhang *et al.* (2020b) propose to use *image-level label distribution* to guide the pixel-level target segmentation. Furthermore, they use the *label distributions of anchor super-pixels* to indicate the network where to update. Learning these easier tasks

improves the predicted pseudo-labels for the target samples and therefore can be used to effectively regularize the fine-tuning of the SiS network.

Similarly, Sakaridis *et al.* (2019) propose a curriculum learning approach where models are adapted from *day-to-night* learning with progressively increasing the level of darkness. They exploit the correspondences of images captured across different daytime to improve pixel predictions at inference time. Lian *et al.* (2019) adopt the *easy-to-hard* curriculum learning approach by predicting labels first at image level, then at region level and finally at pixel level (the main task).

To further improve model performance in the target domain, Pan *et al.* (2020) separate the target data into *easy* and *hard* samples – relying on the entropy – and try to diminish the gap between those predictions by so called intra-domain adversarial training on the corresponding entropy maps (see also Section 2.2.3).

2.3.4 Co-training (CoT)

Another set of UDA/DASiS methods has been inspired by *co-training* (Zhou and Li, 2005) where the idea is to have two distinct classifiers enforced to be diverse, in order to capture different views of the data while predicting the same labels. The main idea behind such methods is that *diversifying the classifiers* in terms of learned parameters – while at the same time *maximizing the consensus* on their predictions – will encourage the model to output more discriminative feature maps for the target domain. The rationale is that the target samples near the class boundaries are likely to be misclassified by the source classifier and using the disagreement of two classifiers on the prediction for target samples can implicitly detect such cases and, in turn, improve the class boundaries.

The first such UDA model maximizing the classifier discrepancy has been proposed by Saito *et al.* (2018b) where the adversarial model alternates between 1) maximizing the discrepancy between two classifiers on the target sample while keeping the feature generator fixed, and 2) training the feature encoder to minimize discrepancy while keeping the classifiers fixed. As an alternative, to encourage the encoder to output more discriminative features for the target domain, Saito *et al.* (2018a) rely on adversarial dropout and Luo *et al.* (2019b) enforce the weights

of the two classifiers to be diverse while using self-adaptive weights in the adversarial loss to improve local semantic consistency. Finally, Lee *et al.* (2019a) consider the sliced Wasserstein discrepancy to capture the dissimilarity between the predicted probability measures that provides a geometrically meaningful guidance to detect target samples that lie far from the support of the source.

2.3.5 Self-ensembling

Another popular method for semi-supervised learning is to use an ensemble of models and to exploit the consistency between predictions under some perturbations. While Laine and Aila (2016) propose the temporal ensembling by taking the per-sample moving average of predictions, Tarvainen and Valpola (2017) replace the averaging predictions with an *exponential moving average* (EMA) of the model weights. In the latter case the Mean Teacher framework is used, represented by a second, non-trainable model whose weights are updated with the EMA over the actual trainable weights.

Such self-ensembling models can also be considered for UDA and DASiS, where the model is generally composed of a *teacher* and a *student* network, encouraged to produce consistent predictions. The teacher is often an ensembled model that averages the student's weights and therefore the predictions from the teacher can be interpreted as pseudo-labels for the student model. Indeed, French *et al.* (2018) extend the model proposed by Tarvainen and Valpola (2017) to UDA considering a separate path for source and target, and sampling independent batches making the Batch Normalization (BN) (Ioffe and Szegedy, 2015) domain specific during the training process. Perone *et al.* (2019) apply self-ensembling to adapt medical image segmentation. The Self-ensembling Attention Network by Xu *et al.* (2019b) aims at extracting attention aware features for domain adaptation (see Figure 2.7).

In contrast to the above mentioned ensemble models, which are effective but require heavily-tuned manual data augmentation for successful domain alignment, Choi *et al.* (2019) propose a self-ensembling framework which deploys a target-guided GAN-based data augmentation with spectral normalization. To produce semantically accurate prediction

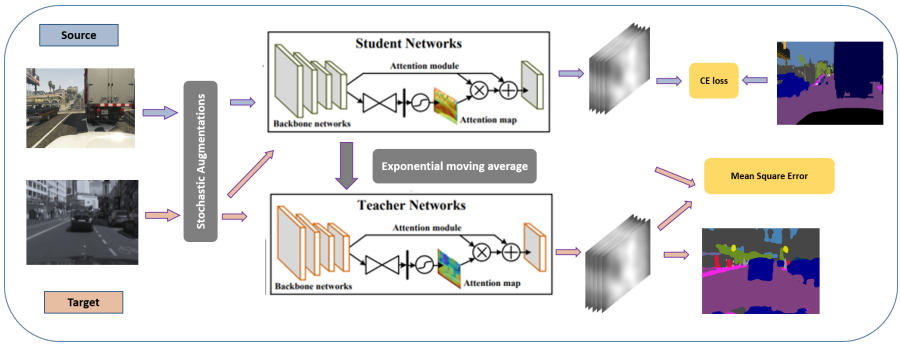

Figure 2.7: The self-ensembling attention network (Xu *et al.*, 2019b) consists of a student network and a teacher network. The two networks share the same architecture with an embedded attention module. The *student network* is jointly optimized with a supervised segmentation loss for the source domain and an unsupervised consistency loss for the target domain. The *teacher network* is excluded from the back-propagation, it is updated with an exponential moving average. In the test phase, the target-domain images are sent to the teacher network to accomplish the SiS task. Image based on Xu *et al.* (2019b).

for the source and augmented samples, a semantic consistency loss is used. More recently, Wang *et al.* (2021b) proposed a method that relies on AdaIN (Huang and Belongie, 2017) to convert the style of source images into that of the target images, and vice-versa. The stylized images are exploited in a training pipeline that exploits self-training, where pseudo-labels are improved via the usage of self-ensembling.

2.3.6 Model distillation

In machine learning, *model distillation* (Hinton *et al.*, 2015) has been introduced as a way to transfer the knowledge from a large model to a smaller one – for example, compressing the discriminative power of an ensemble of models into a single, lighter one. In the context of DASiS, it has been exploited to guide the learning of more powerful features for the target domain, transferring the discriminative power gained on the source samples.

Chen *et al.* (2018c) propose to tackle the distribution alignment in DASiS by using a distillation strategy to learn the target style convolutional filters (see Figure 2.8). Furthermore, taking advantage of

the intrinsic spatial structure presented in urban scene images (that they focus on), they propose to partition the images into non-overlapping grids, and the domain alignment is performed on the pixel-level features from the same spatial region using GRL (Ganin *et al.*, 2016). The Domain Adaptive Knowledge Distillation model (Kothandaraman *et al.*, 2021) consists of a multi-level strategy to effectively distill knowledge at different levels – feature space and output space – using a combination of KL divergence and MSE losses.

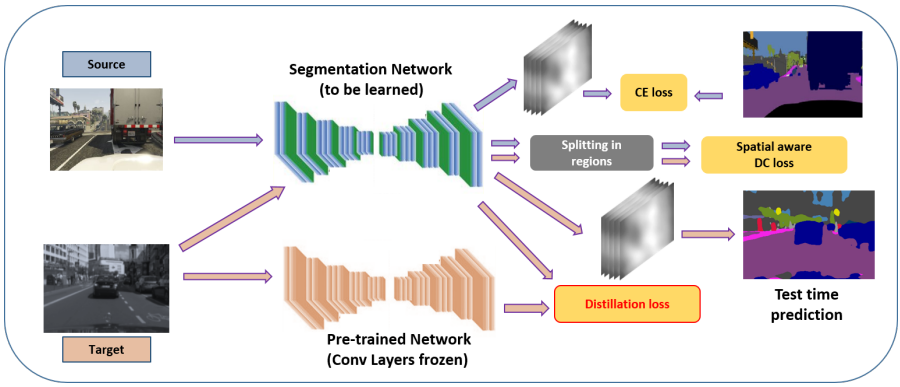

Figure 2.8: Chen *et al.* (2018c) propose to incorporate a target guided distillation module to learn the target (real) style convolutional filters from the (synthetic) source ones and to combine it with a spatial-aware distribution adaptation module. Figure based on Chen *et al.* (2018c).

Chen *et al.* (2022) formalize the self-training as knowledge distillation where the target network is learned by knowledge distillation from the source teacher model. They analyze failures when adapting Swin Transformer (Liu *et al.*, 2021d) based segmentation model to new domains, and suggest that these failures are due to the severe high-frequency components generated during both the pseudo-label construction and feature alignment for target domains. As a solution, they introduce a low-pass filtering mechanism integrated into a *momentum network* which smooths the learning dynamics of target domain features and their pseudo labels. Then a dynamic adversarial training strategy is used to align the distributions, where the *dynamic weights* are used to evaluate the importance of the samples. In a similar spirit, Hoyer *et al.* (2022) exploit the strengths of the transformers in a knowledge distillation-

based self-training framework. The proposed DAFormer architecture is based on a Transformer encoder and a context-aware fusion decoder. To overcome adaptation instability and overfitting to the source domain, they propose *Rare Class Sampling*, which takes into account the long-tail distribution of the source domain. They further distill ImageNet knowledge through the *Thing-Class ImageNet Feature Distance*.

2.3.7 Adversarial attacks

The aim of adversarial attacks (Szegedy *et al.*, 2014) is *to perturb* examples in a way that makes deep neural networks fail when processing them. The model trained with both clean and perturbed samples in an adversarial manner, have been shown to learn more robust models for the given task. While the connection between adversarial robustness and generalization is not fully explained yet (Gilmer *et al.*, 2019), adversarial training has been successfully applied to achieve different goals than adversarial robustness; for instance, it has been used to mitigate overfitting in supervised and semi-supervised learning (Zheng *et al.*, 2016), to tackle domain generalization tasks (Volpi *et al.*, 2019), or to fill in the gap between the source and target domains by adapting the classification decision boundaries (as discussed in Section 2.2).

Concerning adversarial attack in the case of DASiS, Yang *et al.* (2020a) propose pointwise perturbations to generate adversarial features that capture the vulnerability of the model – for example the tendency of the classifier to collapse into the classes that are more represented, in contrast with the long tail of the most under-represented ones – and conduct adversarial training on the segmentation network to improve its robustness.

Yang *et al.* (2021) study the adversarial vulnerability of existing DASiS methods and propose the adversarial self-supervision UDA, where the objective is to maximize – by using a contrastive loss – the proximity between clean images and their adversarial counterparts in the output space. Huang *et al.* (2021) propose a Fourier adversarial training method, where the pipeline is to generate adversarial samples – by perturbing certain high frequency components that do not carry significant semantic information – and use them to train the model. This

training technique allows reaching an area with a flat loss landscape, which yields a more robust domain adaptation model.

2.3.8 Self-supervised learning

Self-supervised learning approaches (see also Section 1.3.5) have found their place in UDA research (Sun *et al.*, 2019c; Bucci *et al.*, 2021; Xu *et al.*, 2019a). For what concerns DASiS, Araslanov and Roth (2021) propose a lightweight self-supervised training scheme, where the consistency of the semantic predictions across image transformations such as photometric noise, mirroring and scaling is ensured. The model is trained end-to-end using co-evolving pseudo-labels – using a momentum network, which is a copy of the original model that evolves slowly – and maintaining an exponentially moving class prior. The latter is used to discount the confidence thresholds for classes with few samples, in order to increase their relative contribution to the training loss.

Similarly, Yang *et al.* (2021) – as mentioned in the previous paragraph – exploit self-supervision in DASiS by minimizing the distance between clean and adversarial samples in the output space via a contrastive loss.

2.4 Beyond Classical DASiS

Typical DASiS methods assume that both source and target domains consist of samples drawn from single data distributions, both available, and that there is a shift between the two distributions. Yet, these assumptions may not hold in the real world and therefore several methods have been proposed that tackle specific problem formulations where some of these assumptions are relaxed or additional constraints added (see Figure 2.9 for an illustration of different scenarios related to different data availability assumptions).

For instance, in *multi-source* domain adaptation (MSDA) the goal is learning from an arbitrary number of source domains (Section 2.4.1), and in *multi-target* domain adaptation (MTDA) the aim is to learn from a single source for several unlabeled target domains simultaneously (Section 2.4.2). Instead of having a well defined set of target domains, one

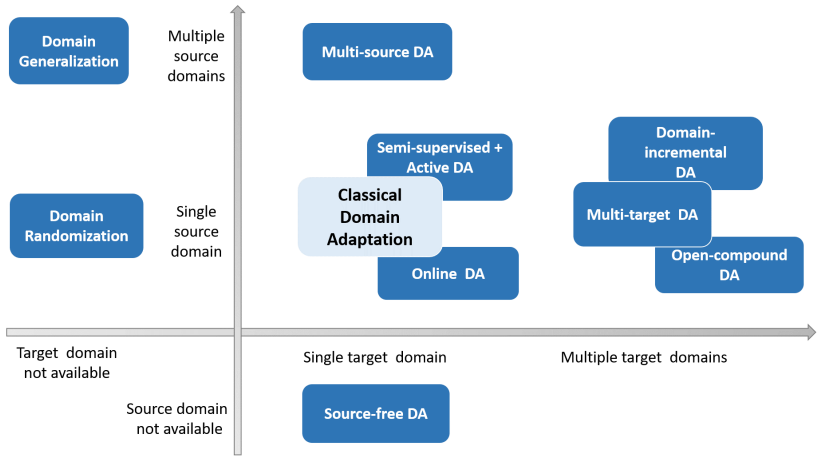

Figure 2.9: Illustration of different scenarios based on source and target data availability.

can address the problem of a target distribution which is assumed to be a compound of multiple, unknown, homogeneous domains (Section 2.4.3).

Alternatively to adapting the model simultaneously to several new target domains, the learning can be done incrementally when the access to new domains is in a sequential manner (Section 2.4.4), or considering a single target domain, but the access to the target data is continuous and online (Section 2.4.5). One could make the assumption that the source model is available, but the source data on which it was trained on is not – the *source-free domain adaptation* problem (Section 2.4.6);

Another scenario is *domain generalization*, where the model learns from one or multiple source domains, but has no access to any target sample, nor hints on the target distribution (Section 2.4.7). On the other end, different methods tackle the semi-supervised domain adaptation problem, where one even assumes that a few target samples are annotated (Section 2.4.8), or can be actively annotated (Section 2.4.9).

Besides the number of domains and the amount of labeled/unlabeled samples available in the source/target domains, another important axis of variation for domain adaptation strategies is the overlap between source and target labels. Indeed, the class of semantic labels in the source and the target domains is not necessarily the same and, therefore, several methods have been proposed that address this issue (Section 2.4.10).

2.4.1 Multi-source DASiS

The simplest way to exploit multiple source domains is to combine all the available data in a single source and train a classical UDA model. While this can, in some cases, provide a reasonable baseline, in other cases, it might yield poor results. This can be due to 1) the fact that there are several data distributions mixed in the combined source, making the adaptation process more difficult if this is not explicitly handled, and 2) in many cases this solution might yield to strong negative transfer as shown by Mansour *et al.* (2009). Alternatively, one can consider a weighted combination of multiple source domains for which theoretical analysis of error bounds has been proposed by Ben-David *et al.* (2010) and Crammer *et al.* (2008). One such algorithm with strong theoretical guarantees was proposed by Hoffman *et al.* (2018a), where they design a distribution-weighted combination for the cross-entropy loss and other similar losses. Cortes *et al.* (2021) propose instead a discriminative method which only needs conditional probabilities – that can be accurately estimated for the unlabeled target data, – relying only on the access to the source predictors and not the labeled source data. Russo *et al.* (2019) extend adversarial DASiS to deal with multiple sources and investigate such baselines, *i.e.* comparing models trained on the union of the source domains versus weighted combination of adaptive adversarial models trained on individual source-target pairs.

Further methods proposed for image classification (Li *et al.*, 2018g; Peng *et al.*, 2019; Peng *et al.*, 2020; Yang *et al.*, 2020c; Zhao *et al.*, 2018a; Zhao *et al.*, 2020; Zhou *et al.*, 2020b; Zhu *et al.*, 2019; Nguyen *et al.*, 2021) show that when the relationship between different source domains is appropriately exploited, it is possible to train a target model to perform significantly better than using just the union of source data or a weighted combination of individual models' outputs. These deep multi-source DA (MSDA) approaches have often focused on learning a common domain-invariant feature extractor that achieves a small error on several source domains, hoping that such representation can generalize well to the target domain.

Inspired by these approaches, several methods have been proposed that extend MSDA solutions from classification to semantic image seg-

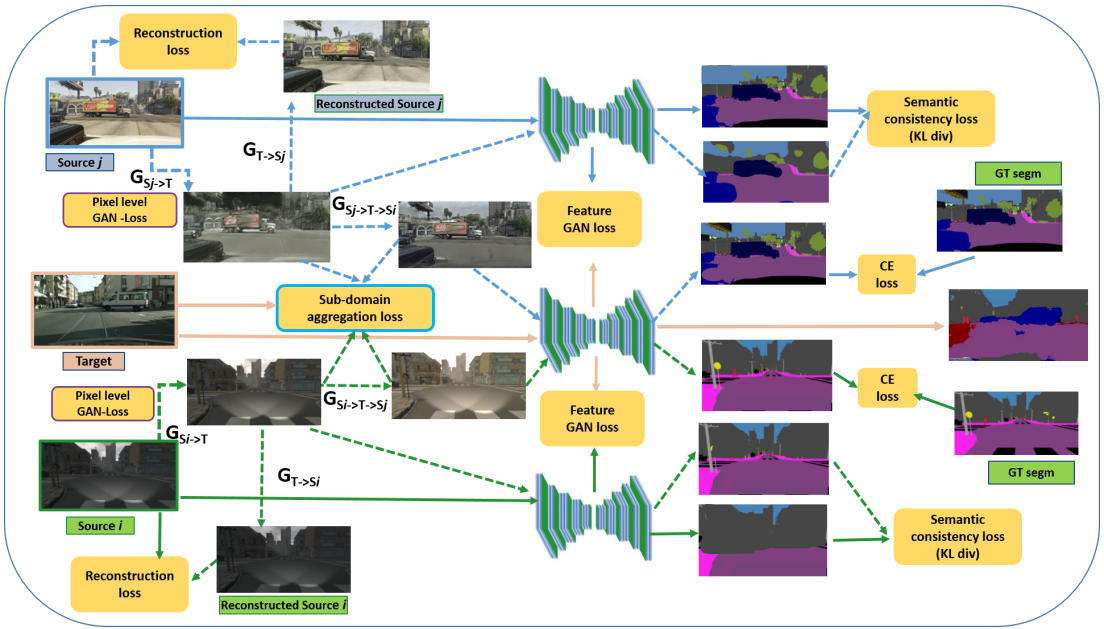

Figure 2.10: The multi-source DASiS framework proposed by Zhao *et al.* (2019b) consists in adversarial domain aggregation with two kinds of discriminators: a sub-domain aggregation discriminator, which is designed to directly make the different adapted domains indistinguishable, and a cross-domain cycle discriminator, discriminates between the images from each source and the images transferred from other sources. Figure based on Zhao *et al.* (2019b).

mentation. As such, the Multi-source Adversarial Domain Aggregation Network (Zhao *et al.*, 2019b) extends (Zhao *et al.*, 2018a) by combining it with CyCADA (Hoffman *et al.*, 2018b). The model, trained end-to-end, generates for each source an adapted style transferred domain with dynamic semantic consistency loss between the source predictions of a pre-trained segmentation model and the adapted predictions of a dynamic segmentation model. To make these adapted domains indistinguishable, a *sub-domain aggregation discriminator* and a *cross-domain cycle discriminator* is learned in an adversarial manner (see Figure 2.10).

Similarly, Tasar *et al.* (2020) propose StandardGAN, a data standardization technique based on GANs (style transfer) for satellite image segmentation, whose goal is to standardize the visual appearance of the source and the target domain with adaptive instance normalization (AdIN) (Huang and Belongie, 2017) and Least-square GAN (Mao *et al.*, 2017) to effectively process target samples. Then, they extend the single-source StandardGAN to multi-source by multi-task learning where an auxiliary classifier is added on top of the discriminator.

In contrast, He *et al.* (2021a) propose a collaborative learning approach. They first translate source domain images to the target style by aligning the different distributions to the target domain in the LAB color space. Then, the SiS network for each source is trained in a supervised fashion by relying on the GT annotations and additional soft supervision coming from other models trained on different source domains. Finally, the segmentation models associated with different sources collaborate with each other to generate more reliable pseudo-labels for the target domain, used to refine the models.

Gong *et al.* (2021b) consider the case where the aim is to learn from different source datasets with potentially different class sets, and formulate the task as a multi-source domain adaptation with *label unification*. To approach this, they propose a two-step solution: first, the knowledge is transferred form the multiple sources to the target; second, a unified label space is created by exploiting pseudo-labels, and the knowledge is further transferred to this representation space. To address – in the first step – the risk of making confident predictions for unlabeled samples in the source domains, three novel modules are proposed: *domain attention*, *uncertainty maximization* and *attention-guided adversarial alignment*.

2.4.2 Multi-target DASiS

In multi-target domain adaptation (MTDA) the goal is to learn from a single labeled source domain with the aim of performing well on multiple target domains at the same time. To tackle MTDA within an image classification context, standard UDA approaches were directly extended to multiple targets (Gholami *et al.*, 2020; Chen *et al.*, 2019d; Roy *et al.*, 2021; Nguyen-Meidine *et al.*, 2021).

Within the DASiS context, a different path was taken. Isobe *et al.* (2021) propose to train an expert model for every target domain where the models are encouraged to *collaborate via style transfer*. Such expert models are further exploited as teachers for a common student model that learns to imitate their output and serves as regularizer to bring the different experts closer to each other in the learned feature space. Instead, Saporta *et al.* (2021) propose to combine for each target domain $\mathcal{T}_i$ two adversarial pipelines: one that learns to discriminate between the domain $\mathcal{T}_i$ and the source, and one between $\mathcal{T}_i$ and the union of the other target domains. Then, to reduce the instability that the multi-discriminator model training might cause, they propose a multi-target knowledge transfer by adopting a *multi-teacher/single-student distillation mechanism*, which leads to a model that is agnostic to the target domains.

2.4.3 Open-compound DASiS

The possibility of having multiple target domains is also addressed in the *open-compound domain adaptation* (OCDA) setting, where the target distribution is assumed to be a compound of multiple, unknown, homogeneous domains (see Figure 2.11). To face this problem, Liu *et al.* (2020b) rely on a *curriculum adaptive strategy*, where they schedule the learning of unlabeled instances in the compound target domain according to their individual gaps to the labeled source domain, approaching an incrementally harder and harder domain adaptation problem until the entire target domain is covered. The purpose is to learn a network that maintains its discriminative leverage on the classification or segmentation task at hand, while at the same time learning more robust features for the whole compound domain. To further prepare the model

for open domains during inference, a *memory module* is adopted to effectively augment the representations of an input when it is far away from the source.

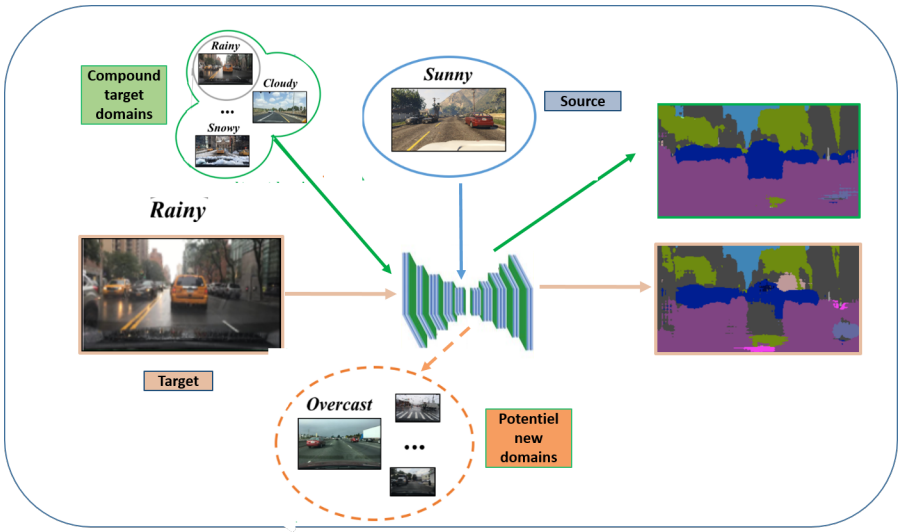

Figure 2.11: In open-compound domain adaptation (OCDA) setting, the target distribution is assumed to be a compound of multiple, unknown, homogeneous domains. Figure based on Liu *et al.* (2020b).

In contrast, Gong *et al.* (2021a) propose a *meta-learning-based framework* to approach OCDA. First, the target domain is clustered in an unsupervised manner into multiple sub-domains by image styles; then, different sub-target domains are split into independent branches, for which domain-specific BN parameters are learned as in Chang *et al.* (2019b). Finally, a meta-learner is deployed to learn to fuse sub-target domain-specific predictions, conditioned upon the style code, which is updated online by using the model-agnostic meta-learning algorithm that further improves its generalization ability.

2.4.4 Domain-incremental SiS

Domain incremental learning is a branch of continual learning, where the goal is extending the underlying knowledge of a machine learning system to new domains, in a sequence of different stages. These incremental

learning stages can be either supervised or unsupervised, according to the available annotations.

For what concerns general solutions under the assumption of *supervised* adaptation, where all data is labeled, a method, not specifically designed but successfully tested on SiS, was proposed by Volpi *et al.* (2021). In order to learn visual representations that are robust against catastrophic forgetting, they propose a meta-learning solution where artificial meta-domains are crafted by relying on domain randomization techniques and they are exploited to learn models that are more robust when transferred to new conditions. The model can benefit from such a solution also when only a few samples are stored under the form of an episodic memory instead of access.

For what concerns methods designed *ad hoc* for SiS, different works consider the problem of learning different domains over the lifespan of a model (Wu *et al.*, 2019; Porav *et al.*, 2019; Garg *et al.*, 2022). They face the domain-incremental problem by assuming that data from new domains come unlabeled – and, therefore, they are more connected to the DASiS literature where the typical task is *unsupervised* DA.

Wu *et al.* (2019) propose to generate data that resembles that of the current target domain and to update the model's parameters relying on such samples. They further propose a *memory bank* to store some domain-specific feature statistics, in order to quickly *restore* domain-specific performance in case the need arises. This is done by deploying the model on a previously explored domain, on which the model has been previously adapted already.

Porav *et al.* (2019) rely on a series of *input adapters* to convert the images processed by the computer vision model when they come from a domain that significantly differ from the source one. They build their method by using GANs, and the proposed approach does not require domain-specific fine-tuning. Instead, Garg *et al.* (2022) learn domain-specific parameters for each new domain – in their case corresponding to different geographical regions – whereas other parameters are assumed to be domain-invariant.

2.4.5 Online DASiS

In online learning (Cesa-Bianchi and Lugosi, 2006), the goal is taking decisions and improving the underlying knowledge of the model sample by sample – in contrast with offline learning, where typically one can process huge amount of data over multiple epochs. The problem of *online adaptation*, intimately connected to online learning, is essentially that of performing UDA as new samples arrive, in order to better perform on them. This problem has been recently re-branded as *test-time adaptation* (Sun *et al.*, 2020b; Wang *et al.*, 2021a; Schneider *et al.*, 2020), where the main focus has been mainly on image classification.

While online adaptation has been addressed in the case of object detection more than a decade ago (Roth *et al.*, 2009), it has been addressed only recently for SiS. In this context, Volpi *et al.* (2022) propose a benchmark to tackle the problem of online adaptation of SiS models (the OASIS benchmark) where the goal is to adapt pre-trained models to new, unseen domains, in a frame-by-frame fashion, Such domain shifts can be adversarial weather conditions met by an autonomous car (see examples in the ACDC dataset (Sakaridis *et al.*, 2021) and Section 3.2).

Different approaches from the continual learning and the test-time adaptation literature have been tailored by Volpi *et al.* (2022) to face this problem, and empirically shown to be helpful, in particular, self-training via pseudo-labels (Lee, 2013) and the application of the TENT algorithm (Wang *et al.*, 2021a). While adapting the model frame by frame, the main challenge is avoiding catastrophic forgetting of the pre-trained model that is adapted to the new sequences. Therefore two solutions are proposed to tackle this problem, The first one is *experience replay* where the test-time adaptation objective is regularized by optimizing a loss with respect to the original, labeled training samples. The second solution is a reset strategy that allows resetting the model to its original weights when catastrophic forgetting is detected.

Concurrently, Wang *et al.* (2022a) propose a continual Test-Time Domain Adaptation (CoTTA) model to limit the error accumulation by using predictions computed via averaging different weights and augmented copies of an image, which allows mitigating catastrophic forget-

ting. They propose to stochastically restore a small amount of network units to their source pre-trained values at each iteration and, in turn, enforcing the adapted model to preserve source knowledge over time.

2.4.6 Source-free domain adaptation

Source-free domain adaptation constitutes the problem of adapting a given source model to a target domain, but without access to the original source dataset. It has been introduced by Chidlovskii *et al.* (2016), who propose solutions for both supervised and unsupervised domain adaptation, testing them in a variety of machine learning problems (*e.g.* document analysis, object classification, product review classification).

More recently, Li *et al.* (2020c) propose to exploit the pre-trained source model as a starting component for an adversarial generative model that generates target-style samples, improving the classifier performance in the target domain, and in turn, improving the generation process. Liang *et al.* (2020) learn a target-specific feature extraction module by implicitly aligning target representations to the source hypothesis, with a method that exploits at the same time information maximization and self-training. Kurmi *et al.* (2021) treat the pre-trained source model as an energy-based function, in order to learn the joint distribution, and train a GAN that generates annotated samples that are used throughout the adaptation procedure. Xia *et al.* (2021) propose a learnable target classifier that improves the recognition ability on source-dissimilar target features, and perform adversarial domain-level alignment and contrastive matching at category level.

For semantic segmentation, Liu *et al.* (2021c) propose a *dual attention distillation mechanism* to help the generator to synthesize samples with meaningful semantic context used to perform efficient pixel-level domain knowledge transfer. They rely on an *entropy-based intra-domain* module to leverage the correctly segmented patches as supervision during the model adaptation stage (see Figure 2.12).

Sivaprasad and Fleuret (2021) propose a solution where the uncertainty of the target domain samples' predictions is minimized, while the robustness against noise perturbations in the feature space is max-

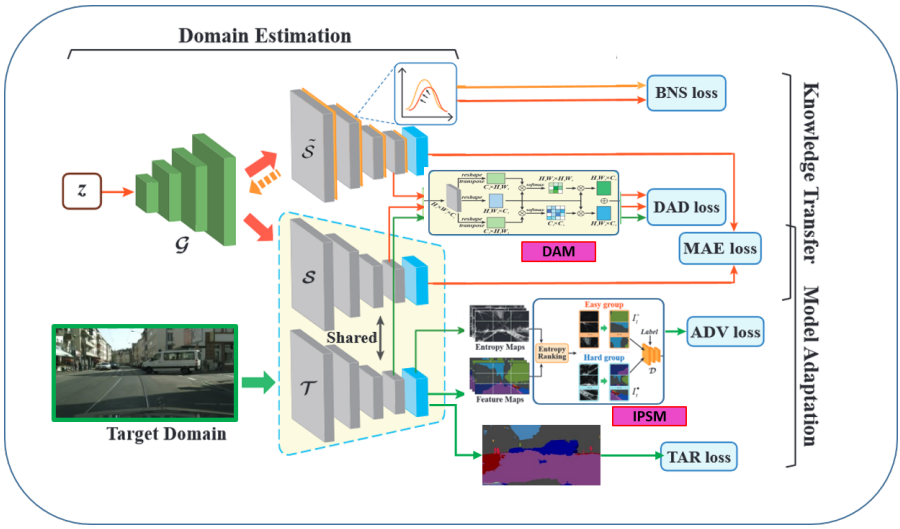

Figure 2.12: Illustration of the source-free DASiS method proposed by Liu *et al.* (2021c) where during training only a well-trained source model and the unlabeled target domain set are accessible but not the source data. To tackle the problem, they propose a *dual attention distillation mechanism* (DAM) to help the generator $\mathcal{G}$ to synthesize "source" samples with meaningful semantic context, beneficial to efficient pixel-level domain knowledge transfer. Furthermore, an *entropy-based intra-domain* patch-level supervision module (IPSM) leverages the correctly segmented patches during the model adaptation stage. Figure based on Liu *et al.* (2021c).

imized. Kundu *et al.* (2021) decompose the problem into performing first source-only domain generalization and then adapting the model to the target by self-training with reliable target pseudo-labels.

2.4.7 Domain generalization

In *domain generalization* (DG), the goal is to generalize a model trained on one or several source domains to new, unseen target domains. Therefore, its main goal is learning *domain-agnostic representations*. As the target domain is unknown at training time, most DG methods aim to minimize the average risk over all possible target domains. According to Wang *et al.* (2020b) and Zhou *et al.* (2020a), such DG methods can be categorized into multiple groups, those relying on data randomization (Tobin *et al.*, 2017), ensemble learning (Xu *et al.*, 2014; Seo

et al., 2020), meta-learning (Balaji *et al.*, 2019; Li *et al.*, 2019b; Rahman *et al.*, 2020) domain-invariant representation learning (Li *et al.*, 2018c; Motiian *et al.*, 2017), feature disentanglement (Chattopadhyay *et al.*, 2020), self-supervised learning (Carlucci *et al.*, 2019; Wang *et al.*, 2020c), invariant risk minimization (Arjovsky *et al.*, 2020) and others.

While not devised *ad hoc* for SiS, several data augmentation methods have been empirically shown to be well performing for the segmentation task. Indeed, Volpi and Murino (2019), Volpi *et al.* (2019), and Qiao *et al.* (2020) show that worst-case data augmentation strategies can improve robustness of segmentation models. Volpi *et al.* (2019) propose to create *fictitious visual domains* – that are hard for the model at hand – by leveraging adversarial training to augment the source domain, and use them to train the segmentation model. Qiao *et al.* (2020) extend this idea by relying on meta-learning; to encourage *out-of-domain augmentations*, the authors rely on a Wasserstein auto-encoder which is jointly learned with the segmentation and domain augmentation within a *meta-learning framework*. Volpi and Murino (2019) instead rely on standard image transformations, by using random and evolution search to find the worst-case perturbations that are further used as data augmentation rules.

Concerning DG methods specifically designed for SiS (DGSiS), different techniques have been explored. Gong *et al.* (2019) propose to learn domain-invariant representations via *domain flow generation*. The main idea is to generate a continuous sequence of intermediate domains between the source and the target, in order to bridge the gap between them. To translate images from the source domain into an arbitrary intermediate domain, an adversarial loss is used to control how the intermediate domain is related to the two original ones (source and target). Several intermediate domains of this kind are generated, such that the discrepancy between the two domains is gradually reduced in a manifold space.

Yue *et al.* (2019) rely on domain randomization, where – using auxiliary datasets – the synthetic images are translated with multiple real image styles to effectively learn domain-invariant and scale-invariant representations. Instead, Jin *et al.* (2020) consider *style normalization* and *restitution* module to enhance the generalization capabilities, while

preserving the discriminative power of the networks. The style normalization is performed by instance normalization to filter out the style variations and therefore foster generalization. To ensure high discriminative leverage, a restitution step adaptively distills task-relevant discriminative features from the residual (*i.e.* the difference between original and style normalized features), which are then exploited to learn the network.

Liu *et al.* (2020a) extend domain-specific BN layers proposed by Seo *et al.* (2020) for MRI image segmentation, where at inference time an ensemble of prediction is generated and their confidence-weighted average is considered as the final prediction. Choi *et al.* (2021) propose an instance selective whitening loss which disentangles domain-specific and domain-invariant properties from higher-order statistics of the feature representation, selectively suppressing the domain-specific ones. Lee *et al.* (2022) learn domain-generalized semantic features by leveraging a variety of contents and styles from the wild, where they diversify the styles of the source features with the help of wild styles. This is carried out by adding several AdaIN (Huang and Belongie, 2017) layers to the feature extractor during the learning process and increasing the intra-class content variability with content extension to the wild in the latent embedding space.

Closely related with DGSiS, Lengyel *et al.* (2021) propose *zero-shot day-to-night domain adaptation* to improve performance on unseen illumination conditions without the need of accessing target samples. The proposed method relies on task agnostic physics-based illumination priors where a trainable Color Invariant Convolution layer is used to transform the input to a domain-invariant representation. It is shown that this layer allows reducing the day-night domain shift in the feature map activations throughout the network and, in turn, improves SiS on samples recorded at night.

2.4.8 Semi-supervised domain adaptation

Semi-supervised learning (SSL) methods exploit at training time accessibility to both a small amount of labeled data and a large amount of unlabeled data. After gaining traction for more standard classifica-

tion tasks, recently several semi-supervised methods have emerged that address SiS problems (see Section 1.3.1).

The standard UDA setting shares with semi-supervised learning the availability at training time of labeled and unlabeled data; the core difference is that in the semi-supervised framework both sets are drawn from the same domain (i.i.d. assumption), whereas in UDA they are drawn from different data distributions (source and target). In Section 2.3 we have discussed how several strategies from the SSL literature such as pseudo-labeling, self-training, entropy minimization, self-ensembling, have been inherited by DASiS and tailored for cross-domain tasks. Semi-supervised domain adaptation can be seen as a particular case of them, where on the one hand we can see part of pseudo-labels replaced by GT target labels, or on the other hand we can see the source labeled data extended with labeled target samples.

To address such a scenario, Wang *et al.* (2020e) leverage a few labeled images from the target domain to supervise the segmentation task and the adversarial semantic-level feature adaptation. They show that the proposed strategy improves also over a target domain's oracle. Chen *et al.* (2021a) tackle the semi-supervised DASiS problem with a method that relies on a variant of CutMix (Yun *et al.*, 2019) and a *student-teacher* approach based on self-training. Two kinds of data mixing methods are proposed: on the one hand, directly mixing labeled images from two domains from holistic view; on the other hand, region-level data mixing is achieved by applying two masks to labeled images from the two domains. The latter encourages the model to extract domain-invariant features about semantic structure from partial view. Then, a student model is trained by distilling knowledge from the two complementary *domain-mixed teachers* – one obtained by direct mixing and another obtained by region-level data mixing – and which is refined in a self-training manner for another few rounds of teachers trained with pseudo-labels.

Zhu *et al.* (2021) first train an ensemble of student models with various backbones and network architectures using both labeled source data and pseudo labeled target data where the labels are obtained with a teacher model trained on the labeled source. This model is further finetuned for the target domain using a small set of labeled samples not

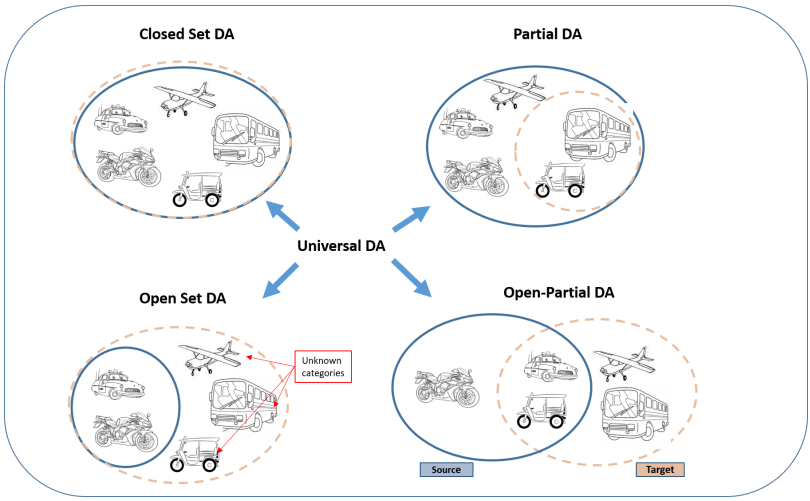

Figure 2.13: A summary of the standard domain adaptation (also known as closed set), partial DA, open set DA and open-partial DA with respect to the overlap between the label sets of the source and the target domains. Universal DA tends to address all cases simultaneously.

only to further adapt to the new domain but also to address class-label mismatch across domains (see also Section 2.4.10).

2.4.9 Active DASiS

Active DASiS (Ning *et al.*, 2021; Shin *et al.*, 2021) is related to semi-supervised DASiS. While for the latter we assume that a small set of target samples are already labeled, in the former, an algorithm selects itself the images or pixels to be annotated by human annotators, and use them to update the segmentation model over iterations. Ning *et al.* (2021) propose a multi-anchor based active learning strategy to identify the most complementary and representative samples for manual annotation by exploiting the feature distributions across the target and source domains. Shin *et al.* (2021) – inspired by the maximum classifier discrepancy (Saito *et al.*, 2018b) – propose a method that selects the regions to be annotated based on the mismatch in predictions across the two classifiers.

More recently, Xie *et al.* (2022a) proposed a new region-based acquisition strategy for active DASiS, which relies on both region impurity and prediction uncertainty, in order to identify the image regions that are both diverse in spatial adjacency and uncertain in terms of output predictions.

2.4.10 Class-label mismatch across domains

Another way of sub-dividing domain adaptation approaches is by considering the mismatch between source and target class sets. Specifically, in *partial domain adaptation* (Zhang *et al.*, 2018a; Cao *et al.*, 2018; Cao *et al.*, 2019) the class set of the source is a super-set of the target one, while *open set domain adaptation* (Panareda Busto and Gall, 2017; Saito *et al.*, 2018c; Rakshit *et al.*, 2020; Jing *et al.*, 2021) assumes that extra private classes exist in the target domain. Finally, *universal domain adaptation* (Fu *et al.*, 2020; Li *et al.*, 2021a; Saito and Saenko, 2021; Ma *et al.*, 2021) integrates both open set and partial DA (see different cases in Figure 2.13).

For what concerns segmentation, Gong *et al.* (2021b) propose an MSDA strategy where the label space of the target domain is defined as the union of the label spaces of all the different source domains and the knowledge in different label spaces is transferred from different source domains to the target domain, where the missing labels are replaced by pseudo-labels.

Liu *et al.* (2021b) propose an optimization scheme which alternates between 1) conditional distribution alignment with adversarial UDA relying on estimated *class-wise balancing* in the target, and 2) target *label proportion estimates* with Mean Matching (Gretton *et al.*, 2009), assuming conditional distributions alignment between the domains.

3

Datasets and Benchmarks

In this section, we discuss datasets and evaluation protocols commonly adopted in SiS (Section 3.1), DASiS (Section 3.2) and related problems – such as class-incremental SiS (Section 3.1.4) and online adaptation (Section 3.2.2). We further cover the main evaluation metrics used in SiS (Section 3.1.1), also discussing more recent alternatives. Furthermore, we emphasize that segmentation performance in terms of accuracy or mIoU is only one of the aspects one should consider when assessing the effectiveness of an SiS approach, discussing the trade-off between accuracy and efficiency in Section 3.1.2 and the vulnerability of SiS models in Section 3.1.3.

3.1 SiS Datasets and Benchmarks

In SiS, we can mainly distinguish the following groups of datasets and benchmarks that we call *object segmentation* (Obj) datasets, *image parsing* (IP) datasets and *scene understanding in autonomous driving* (AD) scenarios. Note that the separation between these datasets are not strict, for example the AD is a particular case of IP. There is also a large set of *medical image* (Med) segmentation datasets and benchmarks (Liu *et al.*, 2021a) that we do not discuss here.

PASCAL Visual Object Classes (VOC) (Everingham *et al.*, 2010) is one of the first and most popular object segmentation datasets. It contains 20 classes to be segmented plus the background. Several versions are available, the most used ones being the Pascal-VOC 2007 (9,963 images) and Pascal-VOC 2012 (11,5K images). MS COCO (Lin *et al.*, 2014) is another challenging object segmentation dataset containing complex everyday scenes with objects in their natural contexts. It contains 328K images with segmentations of 91 object class.

Image or scene parsing datasets contain both *things* (objects) and *stuff* classes. One of the first such dataset is MSRC-21 (Shotton *et al.*, 2009), containing 21 categories and 591 images. The Pascal Context (Mottaghi *et al.*, 2014), extends to IP the segmented images from Pascal-VOC 2010 by labeling the background. It has 10,1K images and 400 classes, however mainly a subset of 59 classes is used, ignoring the others as they have rather low frequency in the dataset. SiftFlow (Liu *et al.*, 2009) includes 2,688 images from the LabelMe database (Russell *et al.*, 2008) annotated with 33 semantic classes. The Stanford background dataset (Gould *et al.*, 2009) contains 715 outdoor images from LabelMe, MSRC and Pascal-VOC where the aim is to separate the foreground (single class) from the background, identifying the seven following semantic *stuff* regions: "sky", "tree", "road", "grass", "water", "mountain" and "buildings". The most used IP dataset is ADE20K (Zhou *et al.*, 2019a) which contains 20K images with 150 semantic categories.

There exists a large set of image parsing datasets proposed in the literature specifically built for urban scene understanding, targeting autonomous driving (AD) scenarios. One of the most popular datasets used to compare SiS methods is Cityscapes (Cordts *et al.*, 2016), but with the increased interest for the AD scenarios, recently a large set of labeled urban scene datasets have been proposed, both real and synthetically rendered with game-engines. In Table 3.1 we provide a summary of such AD oriented SiS datasets with their most important characteristics: the number of classes, the number of annotated samples, whether images are real or rendered, whether the dataset contains video sequences (and not only temporally uncorrelated images), the geographical location (for what concerns simulated datasets, we report the simulated area indicated, if available), and whether the dataset

Table 3.1: Datasets commonly used in urban scene SiS. From left to right columns: dataset's name, number of annotated samples, whether images are simulated or real, whether samples are temporally correlated, recording/simulated location, whether the used can set different visual conditions.

Dataset name (ref. paper)	# Classes	# Annotated samples	Real or sim.	Video seq.	Environment/ geography	Visual conditions
Cityscapes (Cordts *et al.*, 2016)	30	5,000*	Real	Yes	Germany; Zurich	–
BDD100K (Yu *et al.*, 2020)	19	10,000	Real	No	United States	–
KITTI (Geiger *et al.*, 2012)	28	400	Real	Yes	Germany	–
CamVid (Brostow *et al.*, 2009)	32	701	Real	Yes	Cambridge (UK)	–
Mapillary (Neuhold *et al.*, 2017)	66	25,000	Real	No	Worldwide	–
IDD (Varma *et al.*, 2019)	34	10,004	Real	Yes	India	–
RainCityscapes (Hu *et al.*, 2019a)	32	10,620	Real	Yes	Germany	Artificial rain
FoggyCityscapes (Sakaridis *et al.*, 2018)	32	15,000	Real	Yes	Germany	Artificial fog
ACDC (Sakaridis *et al.*, 2021)	19	4,006	Real	Yes	Switzerland	Daytime; Weather
FoggyZurich (Sakaridis *et al.*, 2018)	19	40**	Real	Yes	Zurich	Fog
GTA-5 (Richter *et al.*, 2016)	19	24,966	Sim	Yes	–	–
SYNTHIA (Ros *et al.*, 2016)	13	200,000	Sim	Yes	Highway; NYC; EU	Season; Daytime; Weather
SYNTHIA-RAND (Ros *et al.*, 2016)	11	13,407	Sim	No	–	–
KITTI-v2 (Cabon *et al.*, 2020)	15	21,260	Sim	Yes	Germany	Daytime; Weather
Synscapes (Wrenninge and Unger, 2018)	19	25,000	Sim	No	-	Daytime; Overcast; Scene param.

allows setting arbitrary conditions (seasonal, weather, daylight, *etc.*). In addition, in Table 3.2 we present a summary of the classes available in these different datasets, to ease the comprehension of the compatibility between different models. They are also interesting in the light of incremental SiS (see Section 1.3.4) and new DASiS problems where the sets of semantic classes in the source and target sets do not coincide (shortly discussed in Section 2.4.10).

Table 3.2: Categories of which annotation is provided in different SiS datasets. We report classes available in several (at least three) distinct datasets: some datasets, *e.g.* CamVid (Brostow *et al.*, 2009), contain a variety of other categories.

Classes	Cityscapes	BDD100K	CamVid	IDD	ACDC	GTA-5	SYNTHIA-R	SYNTHIA	KITTI-v2	FoggyZurich	Synscapes
Bicycle	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓
Bridge	✓		✓	✓							
Building	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Bus	✓	✓	✓	✓	✓	✓	✓			✓	✓
Car	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Caravan	✓			✓					✓		
Fence	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓
Guard rail	✓			✓					✓		
Lane marking			✓				✓	✓			
Motorcycle	✓	✓	✓	✓	✓	✓	✓			✓	✓
Parking	✓		✓	✓			✓				
Person	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓
Pole	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Rail track	✓			✓							
Rider	✓	✓		✓	✓	✓	✓			✓	✓
Road	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Sky	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Sidewalk	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓
Terrain	✓	✓			✓	✓	✓		✓	✓	✓
Train	✓	✓	✓		✓	✓	✓			✓	✓
Traffic light	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓
Traffic sign	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Truck	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓
Tunnel	✓		✓	✓							
Vegetation	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Wall	✓	✓	✓	✓	✓	✓	✓			✓	✓

Annotating SiS datasets. Generally, software tools that allow us to annotate images are based on an interface where the user can manipulate polygons that are shaped according to the image's instances; such polygons are further processed into segmentation maps. Some examples of popular, open-source annotation tools are LabelMe,[1] Label Studio,[2] and VIA.[3]

Initially taking an hour or more per image (Cordts *et al.*, 2016), recent semi-automatic tools manage to reduce the annotation time for common urban classes ("people", "road" "surface" or "vehicles") by relying, *e.g.* on pre-trained models for object detection,[4]– however they still require manual verification and validation. For an up-to-date collection of annotation tools,please refer to the link in the footnote.[5]

3.1.1 Evaluating SiS performance

To evaluate SiS, the *overall pixel accuracy* and the *per-class accuracy* have been proposed by Shotton *et al.* (2009). The former computes the proportion of correctly labeled pixels, while the latter calculates the proportion of correctly labeled pixels for each class and then averages over the classes. The Jaccard Index (JI), more popularly known as *intersection over the union* (IoU), takes into account both the false positives and the missed values for each class. It measures the intersection over the union of the labeled segments for each class and reports the average. This measure became the standard to evaluate SiS models, after having been introduced in the Pascal-VOC challenge (Everingham *et al.*, 2010) in 2008. Long *et al.* (2015a) propose, in addition, a *frequency weighted IoU* measure where the IoU for each class is weighted by the frequency of GT pixels corresponding to that class.

We schematize these main metrics below, following the notation used by Long *et al.* (2015a). Let n_{ij} be the number of pixels from the i^{th} class that are classified as belonging to the j^{th} class where $i, j \in \{1, \cdots, C\}$, C being the number of different semantic classes. Let $t_i = \sum_j n_{ij}$ be the total number of pixels of the i^{th} class. The metrics introduced above are defined as follows:

[1] https://github.com/wkentaro/labelme
[2] https://github.com/heartexlabs/label-studio
[3] https://gitlab.com/vgg/via
[4] https://github.com/virajmavani/semi-auto-image-annotation-tool
[5] https://github.com/heartexlabs/awesome-data-labeling.

- **Mean IoU**: $\frac{1}{C}\sum_i \frac{n_{ii}}{(t_i+\sum_j n_{ji}-n_{ii})}$
- **Frequency weighted IoU**: $\frac{1}{\sum_k t_k}\sum_i \frac{t_i \cdot n_{ii}}{(t_i+\sum_j n_{ji}-n_{ii})}$
- **Pixel accuracy**: $\frac{\sum_i n_{ii}}{\sum_i t_i}$
- **Mean accuracy**: $\frac{1}{C}\sum_i \frac{n_{ii}}{t_i}$.

The above measures are generally derived from the confusion matrix computed over the whole dataset having the main advantage that there is no need to handle the absent classes in each image. While these metrics are the most used to evaluate and compare SiS and DASiS models, we would like to mention below a few other metrics that have been introduced in the literature to evaluate SiS models, and could also be interesting for evaluating DASiS.

Instead of relying on the confusion matrix computed over the whole dataset, Csurka *et al.* (2013) propose to evaluate the pixel accuracy, the mean accuracy and the IoU for each image individually, where the IoU is computed by averaging only over the classes present in the GT segmentation map of the image. The main rationale behind this is that the measures computed over the whole dataset do not enable to distinguish an algorithm that delivers a medium score on all images from an algorithm that performs very well on some images and very poorly on others (they could yield very similar averages). To better assess such differences, Csurka *et al.* (2013) propose to measure the percentage of images with a performance higher than a given threshold. Then, given a pair of approaches, the percentage of images for which one of the method outperforms the other one is analyzed, *e.g.* considering the statistical difference of two segmentation algorithms with t-test. Finally, it has also been noticed by Csurka *et al.* (2013) that per-image scores reduce the bias w.r.t. large objects, as missing or incorrectly segmented small objects have low impact on the global confusion matrix.

Another important aspect of semantic segmentation is the accurate semantic border detection. To evaluate the accuracy of boundary segmentation, Kohli *et al.* (2009) propose Trimap that defines a narrow band around each contour and computes pixel accuracies within

the given band. Instead, to measure the quality of the segmentation boundary, Csurka *et al.* (2013) extend the Berkeley contour matching (BCM) score (Martin *et al.*, 2004) – proposed to evaluate similarity between unsupervised segmentation and human annotations – to SiS, where a BCM score is computed between the GT and predicted contours corresponding to each semantic class (after binarizing first both segmentation maps). The scores are averaged over the classes present in the GT map.

3.1.2 Trade-off between accuracy and efficiency

The segmentation accuracy is not a unique metric when evaluating and comparing segmentation models. Indeed, SiS can be extremely demanding for high computational resources, particularly due to the fact that it is a pixel-level task, as opposed to image-level tasks. In real applications where latency is crucial, one needs to trade-off accuracy for efficiency. Indeed, as previously discussed, being a key element of scene understanding for autonomous driving, robotic applications or augmented reality, semantic segmentation models should accommodate real-time settings.

Historical methods, in order to achieve reasonable performance, often required a costly post-processing. While deep neural network models have significantly boosted the segmentation performance, in most cases this improvement came with a significant cost increase both on model parameters and computation, both at train and inference time.

Several solutions have been proposed to find a good trade-off between accuracy and efficiency. One possibility is to reduce the computational complexity by restricting the input size (Wu *et al.*, 2017; Zhao *et al.*, 2018b); yet, this comes with the loss of fine-grain details and, hence, accuracy drops – especially around the boundaries. An alternative solution is to boost the inference speed by pruning the channels of the network, especially in the early stages of the base model (Badrinarayanan *et al.*, 2017; Paszke *et al.*, 2016). Due to the fact that such solutions weaken the spatial capacity, Paszke *et al.* (2016) propose to abandon the downsampling operations in the last stage, at the cost of diminishing

the receptive field of the model. To further overcome the loss of spatial details, these methods often use U-shape architectures to gradually increase the spatial resolution and to fill some missing details that however introduces additional computational cost.

Instead, Yu *et al.* (2018a) propose the Bilateral Segmentation Network (BiSeNet) where two components – the *Spatial Path* and the *Context Path* – are devised to confront with the loss of spatial information and shrinkage of receptive field respectively.

The segmentation accuracy obtained with Deep Convolutional Networks has further been improved by Transformer-based SiS models (see some examples in Section 1.2.9). These networks rely on high-performing attention-based modules which have linear complexity with respect to the embedding dimension, but a quadratic complexity with respect to the number of tokens. In vision applications, the number of tokens is typically linearly correlated with the image resolution – yielding a quadratic increase in complexity and memory usage in models strictly using self-attention, such as ViT (Dosovitskiy *et al.*, 2021). To alleviate this increase, local attention modules were proposed such as Swin (Liu *et al.*, 2021d). Furthermore, Vaswani *et al.* (2021) found that a combination of local attention blocks and convolutions result in the best trade-off between memory requirements and translational equivariance. Instead Hassani *et al.* (2022) propose the *Neighborhood Attention Transformer*, which limits each query token's receptive field to a fixed-size neighborhood around its corresponding tokens in the *key-value* pair, controlling the receptive fields in order to balance between translational invariance and equivariance. Zhang *et al.* (2022) propose a mobile-friendly architecture named Token Pyramid Vision Transformer (TopFormer) which takes tokens from various scales as input to produce scale-aware semantic features with very light computation cost.

Finally, the recent ConvNeXt architecture proposed by Liu *et al.* (2022) competes favorably with Transformers in terms of accuracy, scalability and robustness across several tasks including SiS, while maintaining the efficiency of standard ConvNets.

3.1.3 Vulnerability of SiS models

While very effective when handling samples from the training distribution, it is well known that deep learning-based models can suffer when facing *corrupted* samples (Hendrycks and Dietterich, 2019). Crucially, these models suffer from perturbations that are imperceptible to the human eye, but cause severe prediction errors (Szegedy *et al.*, 2014). Modern SiS models are also vulnerable in this sense, therefore increasing their robustness against natural or adversarial perturbations is an active research area. Finally, models with a finite set of classes, including SiS models, can suffer when instances of previously unseen categories appear in a scene.

Adversarial perturbations. Xie *et al.* (2017) and Metzen *et al.* (2017) concurrently show for the first time that semantic segmentation models can also be fooled by perturbations that are imperceptible to the human eye. Metzen *et al.* (2017) show that it is possible to craft *universal* perturbations (Moosavi-Dezfooli *et al.*, 2017), namely perturbations that are sample-agnostic, that can make the network consistently miss-classify a given input. In particular, they show how to craft perturbations to 1) make the SiS model provide always the same output, and 2) make the model avoid predicting "cars" or "pedestrians". Xie *et al.* (2017) instead focus on sample-specific adversarial perturbations, proposing the "Dense Adversary Generation" algorithm. Both works raise security issues on the reliability of SiS models, and therefore the overall systems they are embedded into.

Corruptions. Hendrycks and Dietterich (2019) showed that deep neural network models for image classification are extremely brittle against simple input miss-specification, such as Gaussian and salt-and-pepper noises, but also to artificial corruptions and contrast or brightness modifications such as simulated fog and snow. Kamann and Rother (2020) extend this analysis to SiS models and show that the same conclusions hold: the models are very vulnerable against simple corruptions, which – even though perceptible – would not cause particular difficulties to a human eye.

Unseen classes. The out-of-distribution (OOD) detection (Hendrycks

and Gimpel, 2017) literature is a very active topic in computer vision: given that the number of classes a model can predict is finite, it is important to be able to handle images with unknown instances. In the case of SiS models, this results in being able to determine when *some pixels* in an image are related to a class the model had never been trained on.

Blum *et al.* (2019) and Chan *et al.* (2021) propose the "Fishyscapes" and the "SegmentMeIfYouCan" benchmarks that evaluate and compare SiS models on the task of determining which pixels are related to unknown classes. The latter further introduces a new problem where the task is to determine pixels associated with road obstacles (from known and unknown classes). For what concerns methods for the task of determining pixels from unknown classes, most of them are derived from the OOD literature (Hendrycks and Gimpel, 2017; Liang *et al.*, 2018a) and the uncertainty literature (Kendall and Gal, 2017). While methods in both fields are typically designed for classification tasks, they can be extended to SiS by applying them at pixel level instead of image level.

3.1.4 Class-incremental SiS protocols

In Section 1.3.4 we formulated the problem of class-incremental learning – in the context of SiS. In the following lines, we review the main protocols used to evaluate such class-incremental SiS algorithms. For reference, the first protocols for this task have been proposed by Cermelli *et al.* (2020) and Michieli and Zanuttigh (2021).

The learning procedure, as typical in continual learning, is divided in a sequence of different tasks. In the context of class-incremental SiS, solving a task means learning to segment novel classes, given images where the classes of interest are annotated with GT, and the others are considered as "background". The first task is defined as a learning procedure over a multitude of different classes (as generally happens during model's pre-training). In the following tasks, one or more classes are learned, but generally in inferior numbers with respect to a number of categories learned during the first task.

Formally, given a dataset D with N classes, we will indicate the benchmark as $M - K$, which means that the model is first trained

on M classes, then it learns K new classes at the time (resulting in $1 + (N - M)/K$ consecutive learning steps). Current class-incremental SiS approaches were evaluated mainly on Pascal VOC'12 (Everingham *et al.*, 2010) (20 classes) and ADE20K (Zhou *et al.*, 2019a) (150 classes) datasets. Following the notations above, the following benchmarks have been considered by the community: for Pascal-VOC 2012, $19 - 1$ (2 tasks), $15 - 5$ (5 tasks) and $15 - 1$ (2 tasks) and for ADE20K, $150 - 50$ (2 tasks), $150 - 50$ (2 tasks) and $50 - 50$ (3 tasks).

Furthermore, two different setups are considered by Cermelli *et al.* (2020): the *Disjoint* one, where each task is defined by images that are unique for that task only – which cannot contain classes associated with classes that will be seen in the future; and the *Overlapped* one, where future classes may be present, and images can be replicated across different tasks.

3.2 DASiS Benchmarks

Understanding traffic scene images taken from vehicle mounted cameras is important for such advanced tasks as autonomous driving and driver assistance. It is a challenging problem due to large variations under different weather or illumination conditions (Di *et al.*, 2018) or when a model needs to cope with different environments such as city, countryside and highway.

Even though relying on real samples (such as the datasets listed in Table 3.1) allows assessing model performance in conditions that are more similar to deployment ones, manually annotating an image at pixel level for SiS is a very tedious and costly operation. Recent progress in computer graphics and modern graphics platforms such as game engines raise the prospect of easily obtaining labeled, synthetic datasets. Some examples in this direction are SYNTHIA (Ros *et al.*, 2016) and GTA-5 (Richter *et al.*, 2016) (see examples in Figure 3.1, the middle and right sides).

However, models learned on such datasets might not be optimal due to the domain shift between synthetic and real data. To tackle this problem, a large set of DASiS methods have been proposed, most of which we surveyed in Section 2. These methods start with a model pre-

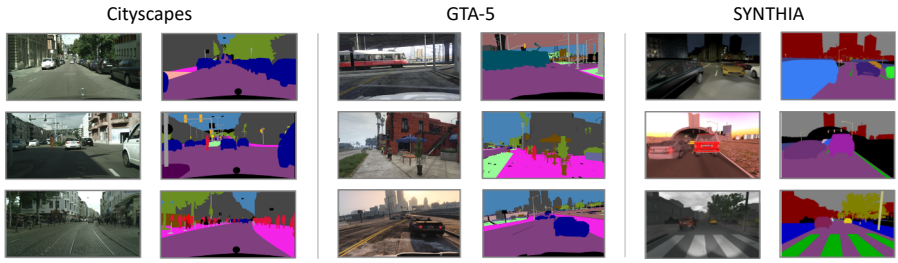

Figure 3.1: Left: Samples from Cityscapes (Cordts *et al.*, 2016) recorded in the real world. They allow to evaluate the model performance on images that resemble the ones an agent will cope with at deployment; the difficulty of collecting real, large-scale datasets is the huge cost required to obtain fine annotations. **Middle:** Synthetic data from GTA-5 (Richter *et al.*, 2016), obtained with high quality game engines, which makes easy the pixel-wise annotation for SiS and scene understanding. However, if the domain shift between real and synthetic data is not addressed, models trained on GTA-5 perform poorly on Cityscapes. **Right:** An autonomous car must cope with large variations, such as day vs. night, weather condition changes, or structural differences, which might affect the image appearance even when the image is taken from the same viewpoint. Simulation engines allow generating large number of samples from urban environments in different conditions, as for example in the SYNTHIA (Ros *et al.*, 2016) dataset.

trained on the simulated source data (typically GTA-5 or SYNTHIA) which is adapted to real target data, for which it is assumed no access to ground-truth annotations. Typically, the Cityscapes (Cordts *et al.*, 2016) dataset is considered in most papers (see examples in Figure 3.1 (left)), however more recent methods started to provide results on newer datasets (listed in Table 3.1). This scenario mimics the realistic conditions such that a large database of simulated, labeled samples is available for training, and the model needs to be adapted to real world conditions without having access to ground-truth annotations.

We summarize the most common settings used in the DASiS research in Table 3.3. They have been introduced in the pioneering DASiS study by Hoffman *et al.* (2016). As the first row in the table indicates, the most widely used benchmark is GTA-5 (Richter *et al.*, 2016) $\rightarrow$ Cityscapes (Cordts *et al.*, 2016) task. It represents a sim-to-real adaptation problem, since GTA-5 was conceived to be consistent with Cityscapes annotations. Following the notation from Section 2.1, the source dataset $\mathcal{D}_{\mathcal{S}}$ is defined by GTA-5 (Richter *et al.*, 2016) annotated

Table 3.3: The most widely used benchmarks within the DASiS community. The first column indicates the source dataset (labeled images available); the second column indicates the target dataset (unlabeled images available); the third column indicates the type of adaptation problem.

Main benchmarks for DASiS

Source domain	Target domain	Adaptation type
GTA-5	Cityscapes	Sim-to-real
SYNTHIA-RAND	Cityscapes	Sim-to-real
Cityscapes (Train)	Cityscapes (Val)	Cross-city (real)
SYNTHIA (Fall)	SYNTHIA (Winter)	Cross-weather (sim)

samples, and the target dataset $\mathcal{D}_\mathcal{T}$ is defined by Cityscapes (Cordts *et al.*, 2016) (non-annotated) samples.

Naturally, datasets generated with the help of simulation engines are significantly larger, as they are able to generate synthetic data under a broad set of conditions (the only exception is GTA-5 (Richter *et al.*, 2016), that is considerably large but does not allow the user to set different visual conditions). Still, in order to evaluate how the models will perform in the real environment on various real conditions, these synthetic datasets might be not sufficient. Therefore, an important contribution to the semantic segmentation landscape is the real-image ACDC dataset (Sakaridis *et al.*, 2021), that is both reasonably large (slightly smaller than Cityscapes (Cordts *et al.*, 2016)) and flexible in terms of visual conditions: researchers can indeed choose between *foggy*, *dark*, *rainy* and *snowy* scenarios. More importantly, samples are recorded from the same streets in such different conditions, allowing to properly assess the impact of adverse weather/daylight on the models (see examples in Figure 3.2 (left)). RainCityscape (Hu *et al.*, 2019a) and FoggyCityscape (Sakaridis *et al.*, 2018) (see examples in Figure 3.2 (right)) are also extremely valuable in this direction, but in this case the weather conditions are simulated (on top of the real Cityscapes images). We think that these datasets are better suited than the currently used Cityscapes dataset and we expect that in the future DASiS methods will be also evaluated on these or similar datasets.

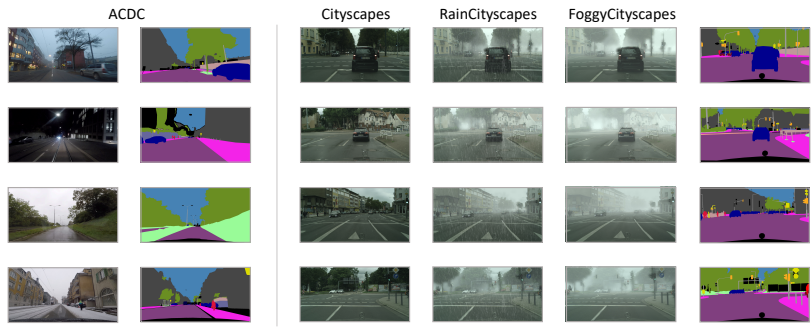

Figure 3.2: Left: Example images from ACDC dataset (Sakaridis *et al.*, 2021) which permits to assess the model performance on real-world weather condition changes (*fog*, *night*, *snow*, *rain*). **Right:** Example images from RainCityscape (Hu *et al.*, 2019a) and FoggyCityscape (Sakaridis *et al.*, 2018), which provide Cityscapes (Cordts *et al.*, 2016) images with simulated *rain* and *fog*, respectively.

3.2.1 DA and DASiS evaluation protocols

There exist two main evaluation protocols in DA, namely, *transductive* and *inductive*. Transductive DA aims to learn prediction models that directly assign labels to the target instances available during training. In other words, the model aims to perform well on the sample set $\mathcal{D}_{\mathcal{T}}$ used to learn the model. Instead, the inductive UDA measures the performance of the learned models on held-out target instances that are sampled from the same target distribution, $\widehat{\mathcal{D}_{\mathcal{T}}} \sim D_{\mathcal{T}}$. While in classical DA most often the transductive protocol is considered, in the case of DASiS, the *inductive* setting is the preferred one.

Selecting the best models, hyper-parameter settings is rather challenging in practice. As described by Saito *et al.* (2021), many methods do hyper-parameter optimization using the risk computed on target domain's annotated samples, which contradicts the core assumption of UDA – *i.e.* not using any labels from the target set. Furthermore, in many papers, a clear description about how the final model has been selected for evaluation is often missing, making the comparisons between different methods rather questionable. Even if in the inductive evaluation protocol a different set is used to select the model, an obvious question arises: *If the model has access to target labels for evaluation,*

why not use those labeled target samples to improve the model in a semi-supervised DA fashion?

Fairer strategies such as transfer cross-validation (Zhong *et al.*, 2010), reverse cross-validation (Ganin *et al.*, 2016), importance-weighted cross-validation (Long *et al.*, 2018) and deep embedded validation (You *et al.*, 2019) rely on source labels, evaluating the risk in the source domain and/or exploiting the data distributions. However, these strategies remain sub-optimal due to the fact that they still rely on the source risk which is not necessarily a good estimator of the target risk in the presence of a large domain gap (Saito *et al.*, 2021).

Instead, Saito *et al.* (2021) revisit the unsupervised validation criterion based on the classifier entropy and show that when the classification model produces confident and low-entropy outputs on target samples the target features are discriminative and the predictions likely reliable. However, they claim that such criterion is unable to detect when a DA method falsely align target samples with the source and incorrectly changes the neighborhood structure. To overcome this limitation, they propose a model selection method based on soft neighborhood density measure to evaluate the discriminability of target features.

3.2.2 Online adaptation for SiS protocols

In Section 2.4.5, we had introduced the problem of online adaptation for SiS for which Volpi *et al.* (2022) propose a three-stage benchmark to train, validate and test corresponding algorithms (the OASIS benchmark). In general, the three steps are 1) pre-train a model on simulated data; 2) validate the adaptation algorithm on simulated *sequences* of *temporally correlated* samples; 3) test the validated model/method on real sequences (see illustration in Figure 3.3). In practice, they propose to use GTA-5 dataset (Richter *et al.*, 2016) in 1), the SYNTHIA dataset (Ros *et al.*, 2016) in 2), and Cityscapes (Cordts *et al.*, 2016) (original and with artificial weather conditions) and ACDC (Sakaridis *et al.*, 2021) datasets for final testing in 3). The proposed pipeline allows evaluating the algorithm performance on environments that are unseen, both at training and validation, mimicking real-world deployment in unfamiliar environments.

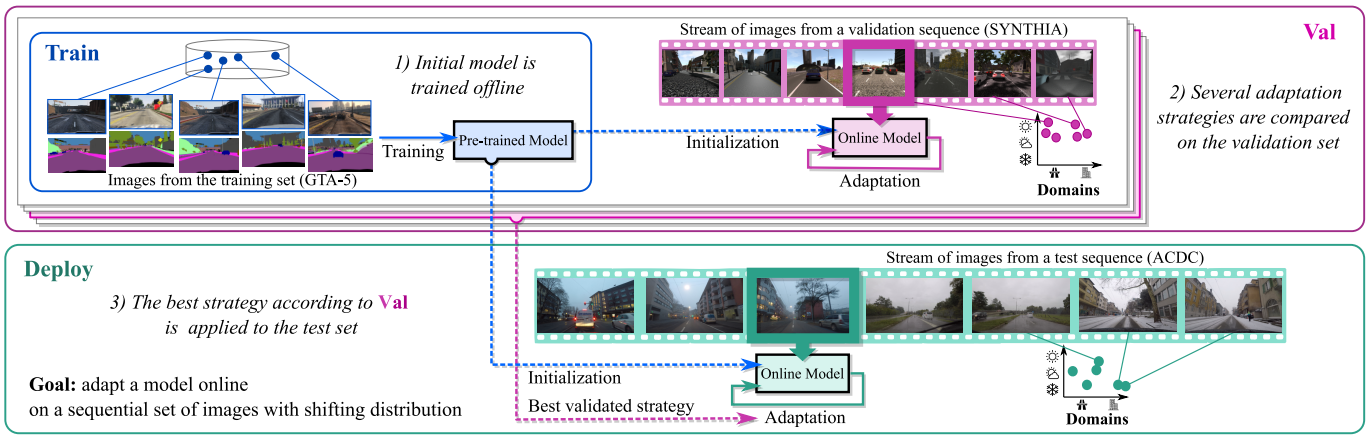

Figure 3.3: The OASIS benchmark addressing evaluation of the online, unsupervised adaptation of semantic segmentation three steps. The model is **trained** offline on simulated data (top-left), several adaptation strategies can be **validated** on simulated data organized in sequentially shifting domains (*e.g. sunny-to-rainy*, *highway-to-city*), to mimic deploy (top-right), and **tested** on real data (bottom).

4

Related Segmentation Tasks

In this section, we discuss briefly some tasks that are closely related to SiS such as instance segmentation (Section 4.1), panoptic segmentation (Section 4.2) and medical image segmentation (Section 4.3).

4.1 Instance Segmentation (InstS)

SiS is strongly related to Instance Segmentation (Yang *et al.*, 2012), which can be seen as a combination of object detection and semantic segmentation. The goal in InstS is indeed to detect and segment all instances of a category in a given image, while also ensuring that each instance is uniquely identified (see illustration in Figure 4.1 (middle)).

Early instance segmentation methods are based on complex graphical models (Silberman *et al.*, 2012; Zhang *et al.*, 2016b; Arnab and Torr, 2017), post-processing object detection (Yang *et al.*, 2012; Tighe *et al.*, 2014; Chen *et al.*, 2015), or models built on top of segment region proposals (Hariharan *et al.*, 2014; Pinheiro *et al.*, 2015).

Amongst more recent deep methods relying on object detectors, Mask R-CNN (He *et al.*, 2017) is one of the most successful ones. It employs an R-CNN object detector (Girshick *et al.*, 2014) and region of interest (RoI) operations – typically RoIPool or RoIAlign – to crop

Figure 4.1: Semantic image segmentation is related to Instance Segmentation (Yang *et al.*, 2012) and to Panoptic Segmentation (Kirillov *et al.*, 2019b). Instance Segmentation can be seen as a combination of object detection and semantic segmentation where the aim is to detect and segment all instances of a category in an image and such that each instance is uniquely identified. Panoptic Segmentation mixes semantic and instance segmentation, where for some *things* classes – countable objects such as "cars", "pedestrians", *etc.* – each instance is segmented individually, while for other classes especially those belonging to *stuff* – "road", "sky", "vegetation", "buildings" – all classes are labeled with a single class label.

the instance from the feature maps. Liu *et al.* (2018b) propose to further improve Mask R-CNN by 1) bottom-up path augmentation, which shortens the information path between lower layers and top most features, 2) by adaptive feature pooling, and 3) by including a complementary branch that captures different views for each proposal.

Novotny *et al.* (2018) extend Mask R-CNN with semi-convolutional operators, which mix information extracted from the convolutional network with information about the global pixel location. YOLACT (Bolya *et al.*, 2019) and BlendMask (Chen *et al.*, 2020a) can be seen as a reformulation of Mask R-CNN, which decouple RoI detection and feature maps used for mask prediction. MaskLab (Chen *et al.*, 2018a) builds on top of Faster-RCNN (Ren *et al.*, 2015) and for each RoI perform foreground/background segmentation by exploiting semantic segmentation and direction logits.

In contrast to the above *detect-then-segment* strategies, many recent methods build on the top of deep segmentation models reviewed in Section 1.2. FCN models, discussed in Section 1.2.3, are amongst the most popular ones. One of such models, InstanceCut (Kirillov *et al.*, 2017), combines the output of two pipelines – a FCN based SiS model and an instance-aware edge detection, processed independently, – with an image partitionning block that merges the super-pixels into

connected components with a class label assigned to each component. InstanceFCN (Dai *et al.*, 2016), instead of generating one score map per semantic class, computes 3×3 position-sensitive score maps where each pixel corresponds to a classifier prediction concerning its relative positions to an object instance. Li *et al.* (2017) propose a fully convolutional Instance-aware SiS model where position-sensitive inside/outside score maps are used to perform object segmentation and detection jointly and simultaneously. The SOLO models (Wang *et al.*, 2020d; 2020d) assign categories to each pixel within an instance according to the instance's location and size, converting instance segmentation into a single-shot classification-solvable problem using FCNs to output dense predictions.

Dilated Convolutional Models (discussed in Section 1.2.7), and in particular DeepLab-CRF-LargeFOV (Chen *et al.*, 2017b), are fine-tuned and refined for InstS by Liang *et al.* (2018b) and by Zhang *et al.* (2016b). The latter combines it with Densely Connected MRFs to improve instance boundaries (similarly to methods in Section 1.2.2).

Ren and Zemel (2017) propose an end-to-end RNN architecture with an attention mechanism (see also Section 1.2.8). This model combines a box proposal network responsible for localizing objects of interest with a DeconvNet (Noh *et al.*, 2015) to segment image pixels within the box. Arun *et al.* (2020) modify a UNet architecture where they explicitly model the uncertainty in the pseudo label generation process using a conditional distribution.

A transformer-based model (see Section 1.2.9) is applied by Xu *et al.* (2021) who propose a co-scale mechanism to image transformers, where encoder branches are maintained at separate scales while engaging attention across scales. They also design a Conv-attention module which performs relative position embeddings with convolutions in the factorized attention module.

Finally, as for SiS (Section 1.3.2), a large set of weakly supervised methods that rely on bounding box supervision have been proposed also for instance segmentation. For example, Tian *et al.* (2019) train jointly a Mask R-CNN detection and segmentation branches, estimate the object instance map inside each detected bounding box and then generate the positive and negative bags using the bounding box annotations. Tian *et al.* (2021) extends this architecture with CondInst (Tian *et*

al., 2020) employing dynamic instance-aware networks, conditioned on instances which eliminates the need for RoI operations. Lan *et al.* (2021) propose a self-ensembling framework where instance segmentation and semantic correspondences are jointly learned by a structured teacher and bounding box supervision. The teacher is a structured energy model incorporating a pairwise potential and a cross-image potential to model the pairwise pixel relationships both within and across the boxes.

4.2 Panoptic Segmentation (PanS)

Panoptic Segmentation (Kirillov *et al.*, 2019b) unifies semantic and instance segmentation, where for several *things* classes – countable objects such as "cars", "pedestrians", *etc.* – each instance is segmented individually, while for classes belonging to *stuff* – "road", "sky", "vegetation", "buildings" – all pixels are labeled with a single class label (see illustration in Figure 4.1 (right)).

Kirillov *et al.* (2019b), emphasizing the importance of tackling semantic and instance segmentation jointly, introduce the panoptic quality metric in order to evaluate jointly semantic and instance segmentation, and thus open the path to a new set of methods called Panoptic Segmentation. The key idea is that for the *things* classes the model has to predict both the belongings to the given *things* class as well as distinguish the instances within the class, while for *stuff* only the semantic class label needs to be assigned to the relevant pixels.

To solve PanS, Kirillov *et al.* (2019b) propose to combine PSPNet (Zhao *et al.*, 2017) with Mask R-CNN (He *et al.*, 2017), where the models process the inputs independently and then their outputs are combined in a post-processing step. de Geus *et al.* (2018) propose to jointly train two branches with a shared backbone, one being a Mask R-CNN for the InstS and a second one relying on an Augmented Pyramid Pooling module for SiS. The Attention Guided Unified Network (Li *et al.*, 2019a) combines a proposal attention module that selects regions potentially containing *things* with a mask attention module to refine the boundary between *things* and *stuff*.

Liu *et al.* (2019) propose an end-to-end occlusion aware pipeline, where 1) the instance segmentation and stuff segmentation branches –

sharing the backbone features – are optimized by the accumulated losses, and 2) the head branches are fine-tuned on the specific tasks. A spatial ranking module addresses the ambiguities of the overlapping relationship. Instead, Xiong *et al.* (2019) design a deformable convolution based SiS head and a Mask R-CNN based InstS head, and solve the two subtasks simultaneously. Sofiiuk *et al.* (2019) propose a fully differentiable end-to-end network for class-agnostic instance segmentation which, jointly trained with an SiS Branch, can perform panoptic segmentation.

Li *et al.* (2018e), building on top of the Dynamically Instantiated Network (Arnab and Torr, 2017), propose a weakly supervised model for PanS where *things* classes are weakly supervised by bounding boxes, and *stuff* classes with image-level tags.

Several PanS methods have been proposed on the top of DeepLab (Chen *et al.*, 2017b). For instance, Porzi *et al.* (2019) propose an architecture which seamlessly integrates multi-scale features generated by an FPN (Lin *et al.*, 2017b) with contextual information conveyed by a lightweight DeepLab-like module. Yang *et al.* (2019) adopt the encoder-decoder paradigm where SiS and InstS predictions are generated from the shared decoder output and then fused to produce the final image parsing result. This model has been extended by Cheng *et al.* (2020), by adding a dual-ASPP and a dual-decoder structure for each sub-task branch, and by Wang *et al.* (2020a) where axial-attention blocks are used instead of ASPP.

Gao *et al.* (2019) propose to jointly train semantic class labeling with a pixel-pair affinity pyramid that computes – in a hierarchical manner – the probability that two pixels belong to the same instance. Furthermore, they incorporate, with the learned affinity pyramid, a novel cascaded graph partition module to sequentially generate instances from coarse to fine. Yuan *et al.* (2020) proposed the Object-Contextual Representations (OCR) for SiS and generalized it to Panoptic Segmentation where the Panoptic-FPN (Kirillov *et al.*, 2019a) head computes soft object regions and then the OCR head predicts a refined semantic segmentation map.

The Efficient Panoptic Segmentation architecture (Mohan and Valada, 2021) combines a semantic head that aggregates fine and contextual features coherently with a Mask R-CNN-like instance head. The final panoptic segmentation output is obtained by the panoptic fusion module that congruously integrates the output logits from both heads.

Amongst recent transformer-based solutions we can mention the Masked-attention Mask Transformer (Mask2Former) (Cheng *et al.*, 2022) which extracts localized features by constraining cross-attention within predicted mask regions.

4.3 Medical Image Segmentation

Medical image segmentation has an important role in sustainable medical care. With the proliferation of medical imaging equipment, *i.e. computed tomography* (CT), *magnetic resonance imaging* (MRI), *positron-emission tomography*, *X-ray* and *ultrasound imaging* (UI), microscopy and fundus retinal images are widely used in clinics, and medical images segmentation can effectively help doctors in their diagnoses (Greenspan *et al.*, 2016; Ahuja, 2019; King Jr., 2018; Jan and Chen, 2020).

Here we only briefly mention a few works on medical image segmentation that heavily rely on architectures discussed in Sections 1 and 2; for a detailed survey on medical image segmentation we refer the interested reader to the survey by Liu *et al.* (2021a).

FCN and 3D-FCN based methods have been applied for segmenting brain tumors (Myronenko, 2017; Nie *et al.*, 2019) or pathological lung tissues in MRI (Novikov *et al.*, 2018; Anthimopoulos *et al.*, 2019), eye vessels in fundoscopy images (Edupuganti *et al.*, 2017), or skin lesions in dermatology images (Mirikharaji and Hamarneh, 2018).

3D-Unet has been used by Borne *et al.* (2019) to segment brain in MRI, by Ye *et al.* (2019a) to segment heart in CT, and by Zhang and Chung (2018) to segment eye vessel in fundoscopy images. Oktay *et al.* (2018) propose Attention U-Net to segment pancreas in CT.

A SegNet based network has been applied to segment musculoskeletal MRI images (Liu *et al.*, 2018a) and cells on microscopic images (Tran *et al.*, 2018). Different works rely on GAN-based models, in order to predict segmentation maps that are similar to humans' annotations. Such models have been used for MRI image segmentation (Rezaei *et al.*, 2017; Moeskops *et al.*, 2017; Han *et al.*, 2018) and in histopathology (Wang *et al.*, 2017).

DASiS solutions have been designed for MRI segmentation of liver and kidney (Valindria *et al.*, 2018), neuroanatomy (Novosad *et al.*, 2019), retinal vessel (Huang *et al.*, 2020b), white matter hyper-intensities (Or-

bes-Arteaga *et al.*, 2019), and multiple sclerosis lesions (Ackaouy *et al.*, 2020). Furthermore, Bermúdez-Chacón *et al.* (2018) apply DASiS to microscopic image segmentation; Dou *et al.* (2018) and Jiang *et al.* (2018) perform adaptation between CT and MRI images for cardiac structure segmentation and for lung cancer segmentation, respectively. Venkataramani *et al.* (2019) propose a continuous DA framework for X-ray lung segmentation. Cross-center adaptation results of multiple sclerosis lesions and brain tumor segmentation have been considered by Li *et al.* (2020b) and adaptation between gray matter segmentations Perone *et al.*, 2019.

Li *et al.* (2021b) insert a polymorphic transformer (polyformer) into a U-Net model which relying on prototype embeddings, dynamically transforms the target-domain features making them semantically compatible with the source domain. They showcase their model on optic disc/cup segmentation in fundus images and polyp segmentation in colonoscopy images.

5

Conclusive Remarks

5.1 Monograph Summary

In this monograph, we provide a comprehensive and up-to-date review of both semantic image segmentation (SiS) in general as well as the domain adaptation of semantic image segmentation (DASiS) literature. We describe in both cases the main trends and organize methods according to their most important characteristics.

We extend the discussions on the two topics with scenarios that depart from the classical setting. In the case of SiS, we overview methods exploiting unlabeled or weakly labeled data, curriculum or self-supervised strategies or methods learning the semantic classes incrementally. Concerning DASiS, we go beyond the typical single labeled source single unlabeled target and survey proposed methods for tasks such as multi-source or multi-target DA, domain incremental learning, source-free adaptation and domain generalization. We also discuss semi-supervised, active and online domain adaptation.

We complement the discussion around SiS and DASiS topics with an extensive list of the existing datasets, evaluation metrics and protocols – designed to compare different approaches. Finally, we conclude the monograph with a brief overview of three strongly related tasks: instance segmentation, panoptic segmentation and medical image segmentation.

As the survey shows, both SiS and DASiS are very active research fields, with an increasing number of approaches being developed by the community and actively integrated in advanced industrial applications and solutions for autonomous driving, robot navigation, medical imaging, remote sensing, *etc.* Therefore, we believe that the community can benefit from our survey – in particular, PhD students and researchers who are just beginning their work in these fields, but also developers from the industry, willing to integrate SiS or DASiS in their systems, can find answers to their numerous questions.

5.2 SiS with Additional Modalities

This monograph mainly focuses on SiS and DASiS, where raw images represent the only information available for scene understanding. However, both SiS and DASiS can benefit from additional visual information such as depth, 3D maps, text or other – when available. There exists already a large amount of work in this direction and we expect that this line of research will grow further. Though out of the monograph scope, for the sake of completeness we highlight here some of the key directions. The interested reader can find more details in the surveys by Zhang *et al.* (2021), Feng *et al.* (2021), and Zhou *et al.* (2019b).

Additional visual modalities include near-infrared images (Salamati *et al.*, 2014; Liang *et al.*, 2022), thermal images (Ha *et al.*, 2017; Sun *et al.*, 2019d), depth (Wang *et al.*, 2015; Qi *et al.*, 2017; Schneider *et al.*, 2017), surface-normals (Eigen and Fergus, 2015), 3D LiDAR point clouds (Kim *et al.*, 2018; Jaritz *et al.*, 2018; Caltagirone *et al.*, 2019), *etc.* Any of these modalities brings additional information about a scene and can be used to learn a better segmentation model.

One solution to address semantic segmentation with extra modalities is to deploy multi-modal fusion networks (Hazirbas *et al.*, 2016; Li *et al.*, 2016; Valada *et al.*, 2017; Schneider *et al.*, 2017; Caltagirone *et al.*, 2019; Sun *et al.*, 2019d) where multiple modalities are given as input to the system – both at training and inference time – and the model outputs pixel-level semantic labeling. To enhance the fusion between RGB and depth, Hu *et al.* (2019b) propose to add Attention Complementary Modules between the single modality branches and the Fusion branch

allowing the model to selectively gather features from the RGB and depth branches.

Alternatively, the second *modality* is considered as privileged information given at training time but not at test time. Most works on this direction focused on joint monocular depth estimation and semantic segmentation showing that joint training allows improving the performance of both tasks (Wang *et al.*, 2015; Mousavian *et al.*, 2016; Zhang *et al.*, 2018b; Kendall *et al.*, 2018; Chen *et al.*, 2018d; He *et al.*, 2021b). A multi-task guided Prediction-and-Distillation Network was designed by Xu *et al.* (2018), where the model first predicts a set of intermediate auxiliary tasks ranging from low to high level, and then such predictions are used as multi-modal input to a multi-modal distillation module, opted at learning the final tasks. Jiao *et al.* (2018) rely on a synergy network to automatically learn information propagation between the two tasks. Gao *et al.* (2022) use a shared attention block for the two tasks with contextual supervision and rely on a feature sharing module to fuse the task-specific features.

Similarly, extra modality was used as privileged information to improve the segmentation accuracy of DASIS methods, in particular using depth information available for the source data by Lee *et al.* (2019c), Vu *et al.* (2019b), and Chen *et al.* (2019b) and Mordan *et al.* (2020). Instead of using depth information as explicit supervision, Guizilini *et al.* (2021) infer and leverage depth in the target domain through self-supervision from geometric video-level cues, and use it as the primary source of domain adaptation.

5.3 Perspectives in SIS

Concerning the perspectives, the most important one comes from the introduction of *foundation models* (Yuan *et al.*, 2021b) aimed at gaining and applying knowledge with good transferability. They consider the lifecycle of multiple deep learning applications as divided into two stages: pre-training and fine-tuning. In the first stage, the deep model is pre-trained on an upstream task with large-scale data (labeled or unlabeled) for gaining transferable knowledge. In the second stage, the pre-trained model is adapted to a downstream task in the target

domain with labeled data. If the downstream task only has unlabeled data, then additional labeled data from another source domain of identical learning task but different data distribution can be used to improve the performance. Compared with supervised pre-training, self-supervised pre-training leads to competitive or sometimes even better performance on downstream tasks such as object detection and semantic segmentation (Yuan *et al.*, 2021b).

We believe that while these models provide good initialization for the methods discussed in this monograph, without undermining their value, we can foresee that future solutions will exploit and combine the strengths of both worlds.

As an example, we can mention *Language driven Semantic Segmentation* (Li *et al.*, 2022) and *Referring Image Segmentation* (Hu *et al.*, 2016) which are emerging and challenging segmentation problems. Their aim is to segment a target semantic region in an image by understanding a given natural linguistic expression. In early solutions, the models were trained on specific referring image segmentation datasets and where visual and linguistic features are simply concatenated (Liu *et al.*, 2017; Li *et al.*, 2018f) or combined with Cross-Modal Self-Attention (Ye *et al.*, 2019b), using linguistic features to choose amongst visual target regions (proposed by *e.g.* Mask R-CNN) (Yu *et al.*, 2018b), or in a multi-task setting by optimizing expression comprehension and segmentation simultaneously (Luo *et al.*, 2020). More recent solutions are vision-language transformer based architectures (Ding *et al.*, 2021), which build upon and exploit the inherited knowledge of transformer-based joint language and vision models pretrained in a self-supervised manner on very large datasets. It is worth mentioning the successful Contrastive Language-Image Pre-training (CLIP) model (Radford *et al.*, 2021), used by Wang *et al.* (2022b) for referring image segmentation.

Abbreviations

AD Autonomous Driving

ADAS Advanced Driving Assistance System

AdaIN Adaptive Instance Normalization

AdvF Adversarial Features

AAL Axial Attention Layer

BiMaL Bijective Maximum Likelihood

BN Batch Normalization

CAM Classification Activation Maps

CIM Corrupted Image Modeling

CoT Co-training

contrL Contrastive Loss

CNN Convolutional Neural Network

CurrL Curriculum Learning

CT Computed Tomography

DA Domain Adaptation

UDA Unsupervised Domain Adaptation

MSDA Multi-source DA

MTDA Multi-target DA

DASiS Domain Adaptation for Semantic Image Segmentation

OCDA Open-compound DA

DC Domain Classifier

DCN Deconvolution Network

sDCN stacked DCNs

DG Domain Generalization

DM Discrepancy Minimisation

DUC Dense Upsampling Convolutions

EM Expectation-Maximization

EMA Exponential Moving Average

FCN Fully Convolutional Network

RbFCN Resnet based FCN

dFCN Dilated FCN

FPA Feature Pyramid Attention

FPN Feature Pyramid Network

SFPN Semantic FPN

FRRN Full Resolution Residual Network

FSS Few-Shot Segmentation

iFSS Incremental FFS

GAN Generative Adversarial Network

GPA Global Attention Upsample

GSA Global Sub-sampled Attention

GRL Gradient Reversal Layer

GT Ground-truth

HDC Hybrid Dilated Convolutions

HR High-resolution

HTwinT Hybrid Twin Transformer

HTr Hierarchical Transformer

IoU Intersection-over-union

IP Image Parsing

IST Image Style Transfer

InstS Instance Segmentation

JI Jaccard Index

KL Kullback-Leibler

LGA Locally-grouped Attention

LocCons Local Consistency Loss

LSTM Long Short Term Memory

MaskT Mask Transformer

MAE Masked Autoencoder

MST Masked Self-Supervised Transformer

MCD Multi-classifier Discrepancy

MCG Multi-scale Combinatorial Grouping

MCP Maximum Class Probability

MHSA Multiheaded Self-attention

MMD Maximum Mean Discrepancy

MLP Multi Layer Perceptron

MRI Magnetic Resonance Imaging

MSE Mean Squared Error

Obj Object Segmentation

OOD Out-of-distribution

PanS Panoptic Segmentation

PSA Point-wise Spatial Attention

CPAM Position and Channel Attention Module

PUP Progressive Upsampling

PL Pseudo Labels

PFW Positional Feature Weight

PSPNet Pyramid Scene Parsing Network

PVT Pyramid Vision Transformer

RF Random Field

CRF Conditional Random Field

dCRF Dense CRF

MRF Markov Random Field

RCU Residual Convolutional Units

RKHS Reproducing Kernel Hilbert Space

RNN Recurrent Neural Network

RoI Region of Interest

SPP Spatial Pyramid Pooling

ASPP Atrous SPP

SRA Spatial Reduction Attention

SemCons Semantic Consistency

SIM Self-information Map

wSIM weighted SIM

SSL Semi-supervised Learning

SeMask Semantically Masked Transformer

SAB SeMask Attention Block

SEC Seed-expand-constrain

SelfEns Self-ensembling

SelfT Self-training

SGD Stochastic Gradient Descent

SiS Semantic Image Segmentation

SP Super-pixels

SPL Self-paced Learning

SVM Support Vector Machine

SwT Swin Transformer

sTrL Stacked Transformer Layers

TEM Target Entropy Minimisation

TTA Test-Time Adaptation

CoTTA Continual Test-Time Domain Adaptation

TENT Test-Time Adaptation by Entropy Minimization

ViT Visual Transformer

XCA Cross-covariance Attention

References

Ackaouy, A., N. Courty, E. Vallée, O. Commowick, C. Barillot, and F. Galassi. (2020). "Unsupervised Domain Adaptation With Optimal Transport in Multi-Site Segmentation of Multiple Sclerosis Lesions From MRI Data". *Frontiers in Computational Neuroscience*. 14.

Adams, R. and L. Bischof. (1994). "Seeded Region Growing". *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*. 16(6): 641–647.

Ahn, J. and S. Kwak. (2018). "Learning Pixel-level Semantic Affinity with Image-level Supervision for Weakly Supervised Semantic Segmentation". In: *CVPR*.

Ahuja, A. S. (2019). "The Impact of Artificial Intelligence in Medicine on the Future Role of the Physician". *PeerJ*. 7.

Ali, A., H. Touvron, M. Caron, P. Bojanowski, M. Douze, A. Joulin, I. Laptev, N. Neverova, G. Synnaeve, J. Verbeek, and H. Jegou. (2021). "XCiT: Cross-covariance Image Transformers". In: *NeurIPS*.

Alonso, I., A. Sabater, D. Ferstl, L. Montesano, and A. C. Murillo. (2021). "Semi-Supervised Semantic Segmentation with Pixel-Level Contrastive Learning from a Class-wise Memory Bank". In: *ICCV*.

Anthimopoulos, M., S. Christodoulidis, L. Ebner, T. Geiser, A. Christe, and S. Mougiakakou. (2019). "Semantic Segmentation of Pathological Lung Tissue with Dilated Fully Convolutional Networks". *IEEE Journal of Biomedical Health Information*. 23: 714–722.

Araslanov, N. and S. Roth. (2021). “Self-supervised Augmentation Consistency for Adapting Semantic Segmentation”. In: *CVPR*.

Arbelaez, P., M. Maire, C. Fowlkes, and J. Malik. (2011). “Contour Detection and Hierarchical Image Segmentation”. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*. 33(5): 898–916.

Arjovsky, M., L. Bottou, I. Gulrajani, and D. Lopez-Paz. (2020). “Invariant Risk Minimization”. arXiv:1907.02893.

Arnab, A. and P. H. S. Torr. (2017). “Pixelwise Instance Segmentation with a Dynamically Instantiated Network”. In: *CVPR*.

Arun, A., C. V. Jawahar, and M. P. Kumar. (2020). “Weakly Supervised Instance Segmentation by Learning Annotation Consistent Instances”. In: *ECCV*.

Bachmann, R., D. Mizrahi, A. Atanov, and A. Zamir. (2022). “MultiMAE: Multi-modal Multi-task Masked Autoencoders”. arXiv:2204.01678.

Badrinarayanan, V., A. Kendall, and R. Cipolla. (2017). “Segnet: a Deep Convolutional Encoder-Decoder Architecture for Image Segmentation”. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*. 39(12): 2481–2495.

Baevski, A., W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli. (2022). “data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language”. arXiv:2202.03555.

Balaji, Y., S. Sankaranarayanan, and R. Chellappa. (2019). “MetaReg: Towards Domain Generalization using Meta-Regularization”. In: *NeurIPS*.

Bao, H., L. Dong, S. Piao, and F. Wei. (2022). “BEiT: BERT Pre-Training of Image Transformers”. In: *ICLR*.

Bearman, A., O. Russakovsky, V. Ferrari, and L. Fei-Fei. (2016). “What’s the Point: Semantic Segmentation with Point Supervision”. In: *ECCV*.

Ben-David, S., J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. (2010). “A Theory of Learning from Different Domains”. *Machine Learning*. 79(5): 151–175.

Bengio, Y., J. Louradour, R. Collobert, and J. Weston. (2009). “Curriculum Learning”. In: *ICML*.

Berman, M., A. R. Triki, and M. B. Blaschko. (2018). "The Lovász-Softmax Loss: A Tractable Surrogate for the Optimization of the Intersection-Over-Union Measure in Neural Networks". In: *CVPR*.

Bermúdez-Chacón, R., P. M'arquez-Neila, M. Salzmann, and P. Fua. (2018). "A Domain-adaptive Two-stream U-Net for Electron Microscopy Image Segmentation". In: *International Symposium on Biomedical Imaging (ISBI)*.

Blei, D. M., A. Y. Ng, and M. I. Jordan. (2003). "Latent Dirichlet Allocation". *Journal of Machine Learning Research*. 3: 993–1022.

Blum, H., P.-E. Sarlin, J. Nieto, R. Siegwart, and C. Cadena. (2019). "Fishyscapes: A Benchmark for Safe Semantic Segmentation in Autonomous Driving". In: *ICCV Workshops*.

Bolya, D., C. Zhou, F. Xiao, and Y. J. Lee. (2019). "YOLACT: Real-time Instance Segmentation". In: *ICCV*.

Borenstein, E. and S. Ullman. (2004). "Learning to Segment". In: *ECCV*.

Borgwardt, K., A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola. (2006). "Integrating Structured Biological Data by Kernel Maximum Mean Discrepancy". *Bioinformatics*. 22: 49–57.

Borne, L., J.-F. Mangin, and D. Rivière. (2019). "Combining 3D U-Net and Bottom-up Geometric Constraints for Automatic Cortical Sulci Recognition". In: *Medical Imaging with Deep Learning*.

Borse, S., Y. Wang, Y. Zhang, and F. Porikli. (2021). "InverseForm: A Loss Function for Structured Boundary-Aware Segmentation". In: *CVPR*.

Bousmalis, K., N. Silberman, D. Dohan, D. Erhan, and D. Krishnan. (2017). "Unsupervised Pixel-Level Domain Adaptation With Generative Adversarial Networks". In: *CVPR*.

Boykov, Y. and M.-P. Jolly. (2001). "Interactive Graph Cuts for Optimal Boundary and Region Segmentation of Objects in N-D Images". In: *ICCV*.

Bromley, J., J. W. Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, E. Säckinger, and R. Shah. (1993). "Signature Verification Using a "Siamese" Time Delay Neural Network". *International Journal of Pattern Recognition and Artificial Intelligence*. 7(04): 669–688.

Brostow, G. J., J. Fauqueur, and R. Cipolla. (2009). "Semantic Object Classes in Video: a High-definition Ground Truth Database". *Pattern Recognition Letters*. 30(2): 88–89.

Bucci, S., A. D'Innocente, Y. Liao, F. M. Carlucci, B. Caputo, and T. Tommasi. (2021). "Self-Supervised Learning Across Domains". *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*.

Buló, S. R., G. Neuhold, and P. Kontschieder. (2017). "Loss Max-Pooling for Semantic Image Segmentation". In: *CVPR*.

Byeon, W., T. M. Breuel, F. Raue, and M. Liwicki. (2015). "Scene Labeling with LSTM Recurrent Neural Networks". In: *CVPR*.

Cabon, Y., N. Murray, and M. Humenberger. (2020). "Virtual KITTI 2". arXiv:2001.10773.

Caesar, H., J. Uijlings, and V. Ferrari. (2018). "COCO-Stuff: Thing and Stuff Classes in Context". In: *CVPR*.

Caesar, H., J. R. R. Uijlings, and V. Ferrari. (2015). "Joint Calibration for Semantic Segmentation". In: *BMVC*.

Caltagirone, L., M. Bellone, L. Svensson, and M. Wahde. (2019). "LIDAR-Camera Fusion for Road Detection Using Fully Convolutional Neural Networks". *Robotics and Autonomous Systems (RAS)*. 111: 125–131.

Cao, L. and L. Fei-Fei. (2007). "Spatially Coherent Latent Topic Model for Concurrent Object Segmentation and Classification". In: *ICCV*.

Cao, Z., L. Ma, M. Long, and J. Wang. (2018). "Partial Adversarial Domain Adaptation". In: *ECCV*.

Cao, Z., K. You, M. Long, J. Wang, and Q. Yang. (2019). "Learning to Transfer Examples for Partial Domain Adaptation". In: *CVPR*.

Carlucci, F. M., A. D'Innocente, S. Bucci, B. Caputo, and T. Tommasi. (2019). "Domain Generalization by Solving Jigsaw Puzzles". In: *CVPR*.

Cermelli, F., D. Fontanel, A. Tavera, M. Ciccone, and B. Caputo. (2022). "Incremental Learning in Semantic Segmentation from Image Labels". In: *CVPR*.

Cermelli, F., M. Mancini, S. Rota Bulò, E. Ricci, and B. Caputo. (2020). "Modeling the Background for Incremental Learning in Semantic Segmentation". In: *CVPR*.

Cermelli, F., M. Mancini, Y. Xian, Z. Akata, and B. Caputo. (2021). "Prototype-based Incremental Few-Shot Semantic Segmentation". In: *BMVC*.

Cesa-Bianchi, N. and G. Lugosi. (2006). *Prediction, Learning, and Games*. Cambridge University Press.

Cha, S., B. Kim, Y. Yoo, and T. Moon. (2021). "SSUL: Semantic Segmentation with Unknown Label for Exemplar-based Class-Incremental Learning". In: *NeurIPS*.

Chan, R., K. Lis, S. Uhlemeyer, H. Blum, S. Honari, R. Siegwart, P. Fua, M. Salzmann, and M. Rottmann. (2021). "SegmentMeIfYouCan: A Benchmark for Anomaly Segmentation". In: *NeurIPS*.

Chandra, S. and I. Kokkinos. (2016). "Fast, Exact and Multi-Scale Inference for Semantic Image Segmentation with Deep Gaussian CRFs". In: *ECCV*.

Chang, W.-L., H.-P. Wang, W.-H. Peng, and W.-C. Chiu. (2019a). "All about Structure: Adapting Structural Information across Domains for Boosting Semantic Segmentation". In: *CVPR*.

Chang, W.-G., T. You, S. Seo, S. Kwak, and B. Han. (2019b). "Domain-Specific Batch Normalization for Unsupervised Domain Adaptation". In: *CVPR*.

Chattopadhyay, P., Y. Balaji, and J. Hoffman. (2020). "Learning to Balance Specificity and Invariance for in and out of Domain Generalization". In: *ECCV*.

Chaurasia, A. and E. Culurciello. (2017). "LinkNet: Exploiting Encoder Representations for Efficient Semantic Segmentation". In: *VCIP*.

Chen, H., K. Sun, Z. Tian, C. Shen, Y. Huang, and Y. Yan. (2020a). "BlendMask: Top-Down Meets Bottom-Up for Instance Segmentation". In: *CVPR*.

Chen, Y.-H., W.-Y. Chen, Y.-T. Chen, B.-C. Tsai, Y.-C. F. Wang, and M. Sun. (2017a). "No More Discrimination: Cross City Adaptation of Road Scene Segmenters". In: *ICCV*.

Chen, L.-C., A. Hermans, G. Papandreou, F. Schroff, P. Wang, and H. Adam. (2018a). "MaskLab: Instance Segmentation by Refining Object Detection with Semantic and Direction Features". In: *CVPR*.

Chen, L.-C., G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. (2017b). "Deeplab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs". *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*. 40(4): 834–848.

Chen, L.-C., G. Papandreou, F. Schroff, and H. Adam. (2017c). "Rethinking Atrous Convolution for Semantic Image Segmentation". arXiv:1706.05587.

Chen, L.-C., Y. Yang, J. Wang, W. Xu, and A. L. Yuille. (2016). "Attention to Scale: Scale-aware Semantic Image Segmentation". In: *CVPR*.

Chen, L.-C., Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. (2018b). "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation". In: *ECCV*.

Chen, M., H. Xue, and D. Cai. (2019a). "Domain Adaptation for Semantic Segmentation with Maximum Squares Loss". In: *ICCV*.

Chen, R., Y. Rong, S. Guo, J. Han, F. Sun, T. Xu, and W. Huang. (2022). "Smoothing Matters: Momentum Transformer for Domain Adaptive Semantic Segmentation". arXiv:2203.07988.

Chen, S., X. Jia, J. He, Y. Shi, and J. Liu. (2021a). "Semi-Supervised Domain Adaptation Based on Dual-Level Domain Mixing for Semantic Segmentation". In: *CVPR*.

Chen, T., S. Kornblith, M. Norouzi, and G. Hinton. (2020b). "A Simple Framework for Contrastive Learning of Visual Representations". In: *ICML*.

Chen, Y.-T., X. Liu, and M.-H. Yang. (2015). "Multi-instance Object Segmentation with Occlusion Handling". In: *CVPR*.

Chen, X., A. Jain, A. Gupta, and L. S. Davis. (2011). "Piecing Together the Segmentation Jigsaw using Context". In: *CVPR*.

Chen, X., Y. Yuan, G. Zeng, and J. Wang. (2021b). "Semi-Supervised Semantic Segmentation with Cross Pseudo Supervision". In: *CVPR*.

Chen, Y., W. Li, X. Chen, and L. Van Gool. (2019b). "Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach". In: *CVPR*.

Chen, Y., W. Li, and L. Van Gool. (2018c). "Road: Reality Oriented Adaptation for Semantic Segmentation of Urban Scenes". In: *CVPR*.

Chen, Y.-C., Y.-Y. Lin, M.-H. Yang, and J.-B. Huang. (2019c). "CrDoCo: Pixel-level Domain Transfer with Cross-Domain Consistency". In: *CVPR*.

Chen, Z., V. Badrinarayanan, C.-Y. Lee, and A. Rabinovich. (2018d). "GradNorm: Gradient Normalization for Adaptive Loss Balancing in Deep Multitask Networks". In: *ICML*.

Chen, Z., J. Zhuang, X. Liang, and L. Lin. (2019d). "Blending-target Domain Adaptation by Adversarial Meta-Adaptation Networks". In: *CVPR*.

Cheng, B., M. Collins, Y. Zhu, T. Liu, T. S. Huang, H. Adam, and L.-C. Chen. (2020). "Panoptic-DeepLab: A Simple, Strong, and Fast Baseline for Bottom-Up Panoptic Segmentation". In: *CVPR*.

Cheng, B., I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar. (2022). "Masked-attention Mask Transformer for Universal Image Segmentation". In: *CVPR*.

Cheng, Y., F. Wei, J. Bao, D. Chen, F. Wen, and W. Zhang. (2021). "Dual Path Learning for Domain Adaptation of Semantic Segmentation". In: *ICCV*.

Chidlovskii, B., S. Clinchant, and G. Csurka. (2016). "Domain Adaptation in the Absence of Source Domain Data". In: *PKDD*.

Choi, J., T. Kim, and C. Kim. (2019). "Self-Ensembling with GAN-based Data Augmentation for Domain Adaptation in Semantic Segmentation". In: *ICCV*.

Choi, S., S. Jung, H. Yun, J. Kim, S. Kim, and J. Choo. (2021). "RobustNet: Improving Domain Generalization in Urban-Scene Segmentation via Instance Selective Whitening". In: *CVPR*.

Chu, X., Z. Tian, Y. Wang, B. Zhang, H. Ren, X. Wei, H. Xia, and C. Shen. (2021). "Twins: Revisiting the Design of Spatial Attention in Vision Transformers". In: *NeurIPS*.

Clinchant, S., J.-M. Renders, and G. Csurka. (2007). "XRCE's Participation to ImageCLEF". In: *CLEF Online Working Notes*.

Comanicu, D. and P. Meer. (2002). "Mean Shift: A Robust Approach Toward Feature Space Analysis". *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*. 24(5): 603–619.

Corbière, C., N. Thome, A. Saporta, T.-H. Vu, M. Cord, and P. Perez. (2021). "Confidence Estimation via Auxiliary Models". *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*.

Cordts, M., M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. (2016). "The Cityscapes Dataset for Semantic Urban Scene Understanding". In: *CVPR*.

Cortes, C., M. Mohri, A. T. Suresh, and N. Zhang. (2021). "A Discriminative Technique for Multiple-Source Adaptation". In: *ICML*.

Crammer, K., M. Kearns, and J. Wortman. (2008). "Learning from Multiple Sources". *Journal of Machine Learning Research*. 9.

Csurka, G. (2017). "A Comprehensive Survey on Domain Adaptation for Visual Applications". In: *Domain Adaptation in Computer Vision Applications*. Ed. by G. Csurka. *Advances in Computer Vision and Pattern Recognition*. Springer. 1–35.

Csurka, G. (2020). "Deep Visual Domain Adaptation". arXiv:2012.14176.

Csurka, G., F. Baradel, B. Chidlovskii, and S. Clinchant. (2017). "Discrepancy-Based Networks for Unsupervised Domain Adaptation: A Comparative Study". In: *ICCV Workshop on Transferring and Adapting Source Knowledge in Computer Vision (TASK-CV)*.

Csurka, G., C. R. Dance, L. Fan, J. Willamowski, and C. Bray. (2004). "Visual Categorization with Bags of Keypoints". In: *ECCV Workshop on Statistical Learning in Computer Vision (SLCV)*.

Csurka, G., D. Larlus, and F. Perronin. (2013). "What is a Good Evaluation Measure for Semantic Segmentation?" In: *BMVC*.

Csurka, G. and F. Perronnin. (2011). "An Efficient Approach to Semantic Segmentation". *International Journal of Computer Vision (IJCV)*. 95: 198–212.

Dai, J., K. He, Y. Li, S. Ren, and J. Sun. (2016). "Instance-sensitive Fully Convolutional Networks". In: *ECCV*.

Dai, J., K. He, and J. Sun. (2015). "BoxSup: Exploiting Bounding Boxes to Supervise Convolutional Networks for Semantic Segmentation". In: *ICCV*.

de Geus, D., P. Meletis, and G. Dubbelman. (2018). "Panoptic Segmentation with a Joint Semantic and Instance Segmentation Network". arXiv:1809.02110.

DeChicchis, J. (2020). "Semantic Understanding for Augmented Reality and Its Applications". *Tech. rep.* Duke University.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *NAACL HLT*.

Di, S., H. Zhang, C.-G. Li, X. Mei, D. Prokhorov, and H. Ling. (2018). "Cross-Domain Traffic Scene Understanding: A Dense Correspondence-Based Transfer Learning Approach". *IEEE Transactions on Intelligent Transportation Systems*. 19(3): 745–757.

Ding, H., C. Liu, S. Wang, and X. Jiang. (2021). "Vision-Language Transformer and Query Generation for Referring Segmentation". In: *CVPR*.

Dosovitskiy, A., L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. (2021). "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *ICLR*.

Dou, Q., C. Ouyang, C. Chen, H. Chen, and P.-A. Heng. (2018). "Unsupervised Cross-Modality Domain Adaptation of ConvNets for Biomedical Image Segmentations with Adversarial Loss". In: *IJCAI*.

Douillard, A., Y. Chen, A. Dapogny, and M. Cord. (2021). "PLOP: Learning Without Forgetting for Continual Semantic Segmentation". In: *CVPR*.

Du, L., J. Tan, H. Yang, J. Feng, X. Xue, Q. Zheng, X. Ye, and X. Zhang. (2019). "SSF-DAN: Separated Semantic Feature Based Domain Adaptation Network for Semantic Segmentation". In: *ICCV*.

Durand, T., T. Mordan, N. Thome, and M. Cord. (2017). "WILDCAT: Weakly Supervised Learning of Deep ConvNets for Image Classification, Pointwise Localization and Segmentation". In: *CVPR*.

Edupuganti, V. G., A. Chawla, and A. Kale. (2017). "Automatic Optic Disk and Cup Segmentation of Fundus Images Using Deep Learning". In: *ICIP*.

Eigen, D. and R. Fergus. (2015). "Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-Scale Convolutional Architecture". In: *CVPR*.

Everingham, M., L. Van Gool, C. Williams, J. Winn, and A. Zisserman. (2010). "The Pascal Visual Object Classes (VOC) Challenge". *International Journal of Computer Vision (IJCV)*. 88: 303–338.

Fan, R., Q. Hou, M.-M. Cheng, G. Yu, R. R. Martin, and S.-M. Hu. (2018). "Associating Inter-Image Salient Instances for Weakly Supervised Semantic Segmentation". In: *ECCV*.

Fang, K., Y. Bai, S. Hinterstoisser, S. Savarese, and M. Kalakrishnan. (2018). "Multi-Task Domain Adaptation for Deep Learning of Instance Grasping from Simulation". In: *ICRA*.

Fang, Y., L. Dong, H. Bao, X. Wang, and F. Wei. (2022). "Corrupted Image Modeling for Self-Supervised Visual Pre-Training". arXiv:2202.03382.

Farabet, C., C. Couprie, L. Najman, and Y. LeCun. (2012). "Scene Parsing with Multiscale Feature Learning, Purity Trees, and Optimal Covers". In: *ICML*.

Farabet, C., C. Couprie, L. Najman, and Y. LeCun. (2013). "Learning Hierarchical Features for Scene Labeling". *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*. 35(8): 1915–1929.

Feng, D., C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Glaeser, F. Timm, W. Wiesbeck, and K. Dietmayer. (2021). "Deep Multi-Modal Object Detection and Semantic Segmentation for Autonomous Driving: Datasets, Methods, and Challenges". *IEEE Intelligent Transportation Systems Conference*. 22(3): 1341–1360.

Feng, Z., Q. Zhou, X. Tan, G. Cheng, X. Lu, J. Shi, and L. Ma. (2020). "DMT: Dynamic Mutual Training for Semi-Supervised Learning". arXiv:2004.08514.

Fourure, D., R. Emonet, E. Fromont, D. Muselet, A. Tremeau, and C. Wolf. (2017). "Residual Conv-Deconv Grid Network for Semantic Segmentation". In: *BMVC*.

French, G., T. Aila, S. Laine, M. Mackiewicz, and G. Finlayson. (2019). "Semi-supervised Semantic Segmentation Needs Strong, High-dimensional Perturbations". In: *BMVC*.

French, G., S. Laine, T. Aila, and M. Mackiewicz. (2020). "Semi-supervised Semantic Segmentation Needs Strong, Varied Perturbations". In: *BMVC*.

French, G., M. Mackiewicz, and M. Fisher. (2018). "Self-ensembling for Visual Domain Adaptation". In: *ICLR*.

Fu, B., Z. Cao, M. Long, and J. Wang. (2020). "Learning to Detect Open Classes for Universal Domain Adaptation". In: *ECCV*.

Fu, J., J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu. (2019a). "Dual Attention Network for Scene Segmentation". In: *CVPR*.

Fu, J., J. Liu, Y. Wang, and H. Lu. (2017). "Densely Connected Deconvolutional Network for Semantic Segmentation". In: *ICIP*.

Fu, J., J. Liu, Y. Wang, J. Zhou, C. Wang, and H. Lu. (2019b). "Stacked Deconvolutional Network for Semantic Segmentation". *IEEE Transactions on Image Processing (TIP)*.

Galleguillos, C., A. Rabinovich, A. Rabinovich, and S. Belongie. (2008). "Weakly Supervised Object Localization with Stable Segmentations". In: *ECCV*.

Ganin, Y., E. Ustinova, P. Ajakan Hana And Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. S. Lempitsky. (2016). "Domain-Adversarial Training of Neural Networks". *Journal of Machine Learning Research*.

Gao, N., Y. Shan, Y. Wang, X. Zhao, Y. Yu, M. Yang, and K. Huang. (2019). "SSAP: Single-Shot Instance Segmentation With Affinity Pyramid". In: *ICCV*.

Gao, T., W. Wei, Z. Cai, Z. Fan, S. Xie, X. Wang, and Q. Yu. (2022). "CI-Net: Contextual Information for Joint Semantic Segmentation and Depth Estimation". *Applied Intelligence*. Open access.

Gao, W., F. Wan, X. Pan, Z. Peng, Q. Tian, Z. Han, B. Zhou, and Q. Ye. (2021). "TS-CAM: Token Semantic Coupled Attention Map for Weakly Supervised Object Localization". In: *ICCV*.

Garg, P., R. Saluja, V. N. Balasubramanian, C. Arora, A. Subramanian, and C. Jawahar. (2022). "Multi-Domain Incremental Learning for Semantic Segmentation". In: *WACV*.

Garg, S., N. Sünderhauf, F. Dayoub, D. Morrison, A. Cosgun, G. Carneiro, Q. Wu, T.-J. Chin, I. Reid, S. Gould, P. Corke, and M. Milford. (2020). "Semantics for Robotic Mapping, Perception and Interaction: A Survey". *Foundations and Trends® in Robotics*. 8(1-2): 1–224.

Gatta, C., A. Romero, and J. van de Veijer. (2014). "Unrolling Loopy Top-Down Semantic Feedback in Convolutional Deep Networks". In: *CVPR Workshops*.

Geiger, A., P. Lenz, and R. Urtasun. (2012). "Are We Ready for Autonomous Driving? the KITTI Vision Benchmark Suite". In: *CVPR*.

Ghiasi, G. and C. C. Fowlkes. (2016). "Laplacian Pyramid Reconstruction and Refinement for Semantic Segmentation". In: *ECCV*.

Gholami, B., P. Sahu, O. Rudovic, K. Bousmalis, and V. Pavlovic. (2020). "Unsupervised Multi-Target Domain Adaptation: An Information Theoretic Approach". *IEEE Transactions on Image Processing (TIP)*. 29(1): 3993–4002.

Gidaris, S., P. Singh, and N. Komodakis. (2018). "Unsupervised Representation Learning by Predicting Image Rotations". In: *ICLR*.

Gilmer, J., N. Ford, N. Carlini, and E. Cubuk. (2019). "Adversarial Examples Are a Natural Consequence of Test Error in Noise". In: *ICML*.

Girshick, R., J. Donahue, T. Darrell, and J. Malik. (2014). "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation". In: *CVPR*.

Gonfaus, J. M., X. Boix, J. van de Weijer, A. D. Bagdanov, J. Serrat, and J. Gonzàlez. (2010). "Harmony Potentials for Joint Classification and Segmentation". In: *CVPR*.

Gong, R., Y. Chen, D. P. Paudel, Y. Li, A. Chhatkuli, W. Li, D. Dai, and L. Van Gool. (2021a). "Cluster, Split, Fuse, and Update: Meta-Learning for Open Compound Domain Adaptive Semantic Segmentation". In: *CVPR*.

Gong, R., D. Dai, Y. Chen, W. Li, and L. Van Gool. (2021b). "mDALU: Multi-Source Domain Adaptation and Label Unification with Partial Datasets". In: *ICCV*.

Gong, R., W. Li, Y. Chen, and L. Van Gool. (2019). "DLOW: Domain Flow for Adaptation and Generalization". In: *CVPR*.

Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. (2014). "Generative Adversarial Nets". In: *NeurIPS*.

Gopalan, R., R. Li, V. M. Patel, and R. Chellappa. (2015). "Domain Adaptation for Visual Recognition". *Foundations and Trends® in Computer Graphics and Vision*. 8(4): 285–378.

Gould, S., R. Fulton, and D. Koller. (2009). "Decomposing a Scene into Geometric and Semantically Consistent Regions". In: *ICCV*.

Gould, S., J. Rodgers, D. Cohen, G. Elidan, and D. Koller. (2008). "Multi-Class Segmentation with Relative Location Prior". *International Journal of Computer Vision (IJCV)*. 80: 300–316.

Grandvalet, Y. and Y. Bengio. (2004). "Semi-supervised Learning by Entropy Minimization". In: *NeurIPS*.

Grangier, D., L. Bottou, and R. Collobert. (2009). "Deep Convolutional Networks for Scene Parsing". In: *ICML*.

Greenspan, H., B. van Ginneken, and R. M. Summers. (2016). "Guest Editorial Deep Learning in Medical Imaging: Overview and Future Promise of an Exciting New Technique". *IEEE Transactions on Medical Imaging*. 35(5): 1153–1159.

Gretton, A., A. Smola, J. Huang, M. Schmittfull, K. Borgwardt, and B. Schölkopf. (2009). "Covariate Shift by Kernel Mean Matching". In: *Dataset Shift in Machine Learning*. Ed. by J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. The MIT Press.

Gu, C., J. J. Lim, P. Arbeláez, and J. Malik. (2009). "Recognition using Regions". In: *CVPR*.

Gu, Q., Q. Zhou, M. Xu, Z. Feng, G. Cheng, X. Lu, J. Shi, and L. Ma. (2021). "PIT: Position-Invariant Transform for Cross-FoV Domain Adaptation". In: *ICCV*.

Guizilini, V., J. Li, R. Ambruş, and A. Gaidon. (2021). "Geometric Unsupervised Domain Adaptation for Semantic Segmentation". In: *ICCV*.

Guo, R., D. Niu, L. Qu, and Z. Li. (2021a). "SOTR: Segmenting Objects With Transformers". In: *ICCV*.

Guo, X., C. Yang, B. Li, and Y. Yuan. (2021b). "MetaCorrection: Domain-Aware Meta Loss Correction for Unsupervised Domain Adaptation in Semantic Segmentation". In: *CVPR*.

Ha, Q., K. Watanabe, T. Karasawa, Y. Ushiku, and T. Harada. (2017). "MFNet: Towards Real-time Semantic Segmentation for Autonomous Vehicles with Multi-spectral Scenes". In: *IROS*.

Han, Z., B. Wei, A. Mercado, S. Leung, and S. Li. (2018). "Spine-GAN: Semantic Segmentation of Multiple Spinal Structures". *Medical Image Analyses*. 79: 2379–2391.

Hariharan, B., P. Arbeláez, R. Girshick, and J. Malik. (2014). "Simultaneous Detection and Segmentation". In: *ECCV*.

Hariharan, B., P. Arbeláez, R. Girshick, and J. Malik. (2015). "Hypercolumns for Object Segmentation and Fine-grained Localization". In: *CVPR*.

Hassani, A., S. Walton, J. Li, S. Li, and H. Shi. (2022). "Neighborhood Attention Transformer". arXiv:2204.07143.

Hazirbas, C., L. Ma, C. Domokos, and D. Cremers. (2016). "FuseNet: Incorporating Depth into Semantic Segmentation via Fusion-Based CNN Architecture". In: *ACCV*.

He, J., X. Jia, S. Chen, and J. Liu. (2021a). "Multi-Source Domain Adaptation With Collaborative Learning for Semantic Segmentation". In: *CVPR*.

He, J., Z. Deng, L. Zhou, Y. Wang, and Y. Qiao. (2019). "Adaptive Pyramid Context Network for Semantic Segmentation". In: *CVPR*.

He, K., X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. (2022). "Masked Autoencoders are Scalable Vision Learners". In: *CVPR*.

He, K., G. Gkioxari, P. Dollár, and R. Girshick. (2017). "Mask R-CNN". In: *ICCV*.

He, K., X. Zhang, S. Ren, and J. Sun. (2014). "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition". In: *ECCV*.

He, L., J. Lu, G. Wang, S. Song, and J. Zhou. (2021b). "SOSD-Net: Joint Semantic Object Segmentation and Depth Estimation from Monocular images". *Neurocomputing*. 440: 251–263.

He, R., J. Yang, and X. Qi. (2021c). "Re-distributing Biased Pseudo Labels for Semi-supervised Semantic Segmentation: A Baseline Investigation". In: *CVPR*.

He, X. and R. S. Zemel. (2009). "Learning Hybrid Models for Image Annotation with Partially Labeled Data". In: *NeurIPS*.

He, X., R. S. Zemel, and M. Á. Carreira-Perpiñán. (2004). "Multiscale Conditional Random Fields for Image Labeling". In: *CVPR*.

He, X., R. S. Zemel, and D. Ray. (2006). "Learning and Incorporating Top-Down Cues in Image Segmentation". In: *ECCV*.

Hendrycks, D. and T. Dietterich. (2019). "Benchmarking Neural Network Robustness to Common Corruptions and Perturbations". In: *ICLR*.

Hendrycks, D. and K. Gimpel. (2017). "A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks". In: *ICLR*.

Hinton, G., O. Vinyals, and J. Dean. (2015). "Distilling the Knowledge in a Neural Network". arXiv:1503.02531.

Hoffman, J., M. Mohri, and N. Zhang. (2018a). "Algorithms and Theory for Multiple-Source Adaptation". In: *NeurIPS*.

Hoffman, J., E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. A. Efros, and T. Darrel. (2018b). "CyCADA: Cycle-Consistent Adversarial Domain Adaptation". In: *ICML*.

Hoffman, J., D. Wang, F. Yu, and T. Darrell. (2016). "FCNs in the Wild: Pixel-level Adversarial and Constraint-based Adaptation". arXiv:1612.02649.

Hofmann, T. (2001). "Unsupervised Learning by Probabilistic Latent Semantic Analysis". *Machine Learning*. 42(1-2): 177–196.

Hofmarcher, M., T. Unterthiner, J. Arjona-Medina, G. Klambauer, S. Hochreiter, and B. Nessler. (2019). "Visual Scene Understanding for Autonomous Driving Using Semantic Segmentation". In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Ed. by W. Samek, G. Montavon, A. Vedaldi, L. Hansen, and K.-R. Müller. *Advances in Computer Vision and Pattern Recognition*. Springer. 285–296.

Hoi, S. C., D. Sahoo, J. Lu, and P. Zhao. (2018). "Online Learning: A Comprehensive Survey". arXiv:1802.02871.

Hong, S., H. Noh, and B. Han. (2015). "Decoupled Deep Neural Network for Semi-supervised Semantic Segmentation". In: *NeurIPS*.

Hong, S., J. Oh, B. Han, and H. Lee. (2016). "Learning Transferrable Knowledge for Semantic Segmentation with Deep Convolutional Neural Network". In: *CVPR*.

Hong, S., D. Yeo, S. Kwak, H. Lee, and B. Han. (2017). "Weakly Supervised Semantic Segmentation using Web-Crawled Video". In: *CVPR*.

Hong, W., Z. Wang, M. Yang, and J. Yuan. (2018a). "Conditional Generative Adversarial Network for Structured Domain Adaptation". In: *CVPR*.

Hong, Z.-W., Y.-M. Chen, H.-K. Yang, S.-Y. Su, T.-Y. Shann, Y. H. Chang, B. Hsi-Lin Ho, C.-C. Tu, T.-C. Hsiao, H.-W. Hsiao, S.-P. Lai, Y.-C. Chang, and C.-Y. Lee. (2018b). "Virtual-to-Real: Learning to Control in Visual Semantic Segmentation". In: *IJCAI*.

Hoyer, L., D. Dai, and L. Van Gool. (2022). "DAFormer: Improving Network Architectures and Training Strategies for Domain-Adaptive Semantic Segmentation". In: *CVPR*.

Hu, H., F. Wei, H. Hu, Q. Ye, J. Cui, and L. Wang. (2021). "Semi-Supervised Semantic Segmentation via Adaptive Equalization Learning". In: *NeurIPS*.

Hu, R., M. Rohrbach, and T. Darrell. (2016). "Segmentation from Natural Language Expressions". In: *ECCV*.

Hu, R., D. Larlus, and G. Csurka. (2012). "On the Use of Regions for Semantic Image Segmentation". In: *ICCVGIP*.

Hu, X., C.-W. Fu, L. Zhu, and P.-A. Heng. (2019a). "Depth-Attentional Features for Single-Image Rain Removal". In: *CVPR*.

Hu, X., K. Yang, L. Fei, and K. Wang. (2019b). "ACNET: Attention Based Network to Exploit Complementary Features for RGBD Semantic Segmentation". In: *ICIP*.

Huang, H., Q. Huang, and P. Krähenbühl. (2018a). "Domain Transfer Through Deep Activation Matching". In: *ECCV*.

Huang, J., D. Guan, A. Xiao, and S. Lu. (2021). "RDA: Robust Domain Adaptation via Fourier Adversarial Attacking". In: *ICCV*.

Huang, J., S. Lu, D. Guan, and X. Zhang. (2020a). "Contextual-Relation Consistent Domain Adaptation for Semantic Segmentation". In: *ECCV*.

Huang, S. and S. Belongie. (2017). "Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization". In: *ICCV*.

Huang, Z., H. Mao, N. Jiang, and X. Wang. (2020b). "DAPR-Net: Domain Adaptive Predicting-refinement Network for Retinal Vessel Segmentation". In: *MICCAI Workshops*.

Huang, Z., X. Wang, J. Wang, W. Liu, and J. Wang. (2018b). "Weakly-Supervised Semantic Segmentation Network with Deep Seeded Region Growing". In: *CVPR*.

Hung, W.-C., Y.-H. Tsai, Y.-T. Liou, Y.-Y. Lin, and M.-H. Yang. (2018). "Adversarial Learning for Semi-supervised Semantic Segmentation ". In: *BMVC*.

Ioffe, S. (2021). "Batch Renormalization: Towards Reducing Minibatch Dependence in Batch-Normalized Models". In: *NeurIPS*.

Ioffe, S. and C. Szegedy. (2015). "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift". In: *ICML*.

Isobe, T., X. Jia, S. Chen, J. He, Y. Shi, J. Liu, H. Lu, and S. Wang. (2021). "Multi-Target Domain Adaptation with Collaborative Consistency Learning". In: *CVPR*.

Jain, J., A. Singh, N. Orlov, Z. Huang, J. Li, S. Walton, and H. Shi. (2021). "SeMask: Semantically Masked Transformers for Semantic Segmentation". arXiv:2112.12782.

Jan, E. and X. Chen, eds. (2020). *Computer-Aided Oral and Maxillofacial Surgery: Developments, Applications, and Future Perspectives*. Academic Press.

Jaritz, M., R. de Charette, E. Wirbel, X. Perrotton, and F. Nashashibi. (2018). "Sparse and Dense Data with CNNs: Depth Completion and Semantic Segmentation". In: *3DV*.

Jesson, A., N. Guizard, S. H. Ghalehjegh, D. Goblot, F. Soudan, and N. Chapados. (2017). "CASED: Curriculum Adaptive Sampling for Extreme Data Imbalance". In: *MICCAI*.

Ji, Z. and O. Veksler. (2021). "Weakly Supervised Semantic Segmentation: From Box to Tag and Back". In: *BMVC*.

Jiang, J., Y.-C. Hu, N. Tyagi, P. Zhang, A. Rimner, G. S. Mageras, J. O. Deasy, and H. Veeraraghavan. (2018). "Tumor-Aware, Adversarial Domain Adaptation from CT to MRI for Lung Cancer Segmentation". In: *MICCAI*.

Jiao, J., Y. Cao, Y. Song, and R. Lau. (2018). "Look Deeper into Depth: Monocular Depth Estimation with Semantic Booster and Attention-Driven Loss". In: *ECCV*.

Jin, B., M. V. Ortiz Segovia, and S. Susstrunk. (2017). "Webly Supervised Semantic Segmentation". In: *CVPR*.

Jin, X., C. Lan, W. Zeng, and Z. Chen. (2020). "Style Normalization and Restitution for Domain Generalization and Adaptation". arXiv:2101.00588.

Jing, T., H. Liu, and Z. Ding. (2021). "Towards Novel Target Discovery Through Open-Set Domain Adaptation". In: *ICCV*.

Jurie, F. and B. Triggs. (2005). "Creating Efficient Codebooks for Visual Recognition". In: *ICCV*.

Kamann, C. and C. Rother. (2020). "Benchmarking the Robustness of Semantic Segmentation Models". In: *CVPR*.

Kato, Z. and J. Zerubia. (2012). "Markov Random Fields in Image Segmentation". *Foundations and Trends® in Signal Processing*. 5(1-2): 1–155.

Kendall, A. and Y. Gal. (2017). "What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?" In: *NeurIPS*.

Kendall, A., Y. Gal, and R. Cipolla. (2018). "Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics". In: *CVPR*.

Khan, S. H., M. Hayat, M. Bennamoun, F. A. Sohel, and R. Togneri. (2018). "Cost-Sensitive Learning of Deep Feature Representations from Imbalanced Data". *IEEE Transactions on Neural Networks and Learning Systems*. 29(8): 3573–3587.

Khoreva, A., R. Benenson, J. Hosang, M. Hein, and B. Schiele. (2017). "Simple does it: Weakly Supervised Instance and Semantic Segmentation". In: *CVPR*.

Kim, D.-K., D. Maturana, M. Uenoyama, and S. Scherer. (2018). "Season-Invariant Semantic Segmentation with a Deep Multimodal Network". In: *Field and Service Robotics*. Ed. by M. Hutter and R. Siegwart. *Advanced Robotics*. Springer. 255–270.

Kim, M. and H. Byun. (2020). "Learning Texture Invariant Representation for Domain Adaptation of Semantic Segmentation". In: *CVPR*.

King Jr., B. F. (2018). "Artificial Intelligence and Radiology: What will the Future Hold?" *Journal of the American College of Radiology*. 15(3 Part B): 501–503.

Kirillov, A., K. He, R. Girshick, C. Rother, and P. Dollár. (2019a). "Panoptic Feature Pyramid Networks". In: *CVPR*.

Kirillov, A., K. He, R. Girshick, C. Rother, and P. Dollár. (2019b). "Panoptic Segmentation". In: *CVPR*.

Kirillov, A., E. Levinkov, B. Andres, B. Savchynskyy, and C. Rother. (2017). "InstanceCut: from Edges to Instances with MultiCut". In: *CVPR*.

Kohli, P., L. Ladický, and P. H. Torr. (2009). "Robust Higher Order Potentials for Enforcing Label Consistency". *International Journal of Computer Vision (IJCV)*. 82: 302–324.

Kolesnikov, A. and C. H. Lampert. (2016). "Seed, Expand and Constrain: Three Principles for Weakly-Supervised Image Segmentation". In: *ECCV*.

Kothandaraman, D., A. Nambiar, and A. Mittal. (2021). "Domain Adaptive Knowledge Distillation for Driving Scene Semantic Segmentation". In: *WACV*.

Kouw, W. M. and M. Loog. (2021). "A Review of Domain Adaptation without Target Labels". *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*. 43(3): 766–785.

Krähenbühl, P. and V. Koltun. (2011). "Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials". In: *NeurIPS*.

Kulharia, V., S. Chandra, A. Agrawal, P. Torr, and A. Tyagi. (2020). "Box2Seg: Attention Weighted Loss and Discriminative Feature Learning for Weakly Supervised Segmentation". In: *ECCV*.

Kumar, M. P., P. H. S. Torr, and A. Zisserman. (2005). "Obj Cut". In: *CVPR*.

Kumar, M. P., H. Turki, D. Preston, and D. Koller. (2011). "Learning Specific-Class Segmentation from Diverse Data". In: *ICCV*.

Kumar, S. and M. Hebert. (2005). "A Hierarchical Field Framework for Unified Context-Based Classification". In: *ICCV*.

Kundu, J. N., A. Kulkarni, A. Singh, V. Jampani, and R. Babu. (2021). "Generalize Then Adapt: Source-Free Domain Adaptive Semantic Segmentation". In: *ICCV*.

Kurmi, V. K., V. K. Subramanian, and V. P. Namboodiri. (2021). "Domain Impression: A Source Data Free Domain Adaptation Method". In: *WACV*.

Kwak, S., S. Hong, and B. Han. (2017). "Weakly Supervised Semantic Segmentation Using Superpixel Pooling Network". In: *AAAI*.

Ladický, L., C. Russell, P. Kohli, and P. H. S. Torr. (2009). "Associative Hierarchical CRFs for Object Class Image Segmentation". In: *ICCV*.

Lai, X., Z. Tian, L. Jiang, S. Liu, H. Zhao, L. Wang, and J. Jia. (2021). "Semi-supervised Semantic Segmentation with Directional Context-aware Consistency". In: *CVPR*.

Laine, S. and T. Aila. (2016). "Temporal Ensembling for Semisupervised Learning". arXiv:1610.02242.

Lan, S., Z. Yu, C. Choy, S. Radhakrishnan, G. Liu, Y. Zhu, L. S. Davis, and A. Anandkumar. (2021). "DiscoBox: Weakly Supervised Instance Segmentation and Semantic Correspondence From Box Supervision". In: *ICCV*.

Larlus, D., J. Verbeek, and F. Jurie. (2010). "Category Level Object Segmentation by Combining Bag-of-Words Models with Dirichlet Processes and Random Fields". *International Journal of Computer Vision (IJCV)*. 88: 238–253.

Lee, C.-Y., T. Batra, M. H. Baig, and D. Ulbricht. (2019a). "Sliced Wasserstein Discrepancy for Unsupervised Domain Adaptation". In: *CVPR*.

Lee, D.-H. (2013). "Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks". In: *ICML Workshops*.

Lee, J., E. Kim, S. Lee, J. Lee, and S. Yoon. (2019b). "FickleNet: Weakly and Semi-supervised Semantic Image Segmentation using Stochastic Inference". In: *CVPR*.

Lee, K.-H., J. Li, A. Gaidon, and G. Ros. (2019c). "SPIGAN: Privileged Adversarial Learning from Simulation". In: *ICLR*.

Lee, K., H. Lee, and J. Y. Hwang. (2021). "Self-Mutating Network for Domain Adaptive Segmentation in Aerial Images". In: *ICCV*.

Lee, S., H. Seong, S. Lee, and E. Kim. (2022). "WildNet: Learning Domain Generalized Semantic Segmentation from the Wild". In: *CVPR*.

Leibe, B., A. Leonardis, and B. Schiel. (2004). "Combined Object Categorization and Segmentation with an Implicit Shape Model". In: *ECCV Workshop on Statistical Learning in Computer Vision (SLCV)*.

Lempitsky, V., A. Vedaldi, and A. Zisserman. (2011). "A Pylon Model for Semantic Segmentation". In: *NeurIPS*.

Lempitsky, V. S., P. Kohli, C. Rother, and T. Sharp. (2009). "Image Segmentation with A Bounding Box Prior". In: *ICCV*.

Lengyel, A., S. Garg, M. Milford, and J. C. van Gemert. (2021). "Zero-Shot Day-Night Domain Adaptation with a Physics Prior". In: *ICCV*.

Li, B., Y. Shi, Z. Qi, and Z. Chen. (2018a). "A Survey on Semantic Segmentation". In: *ICDM Workshops*.

Li, B., K. Q. Weinberger, S. Belongie, V. Koltun, and R. Ranftl. (2022). "Language-driven Semantic Segmentation". In: *ICLR*.

Li, G., G. Kang, W. Liu, Y. Wei, and Y. Yang. (2020a). "Content-Consistent Matching for Domain Adaptive Semantic Segmentation". In: *ECCV*.

Li, G., G. Kang, Y. Zhu, Y. Wei, and Y. Yang. (2021a). "Domain Consensus Clustering for Universal Domain Adaptation". In: *CVPR*.

Li, H., P. Xiong, J. An, and L. Wang. (2018b). "Pyramid Attention Network for Semantic Segmentation ". In: *BMVC*.

Li, H., S. Jialin Pan, S. Wang, and A. C. Kot. (2018c). "Domain Generalization with Adversarial Feature Learning". In: *CVPR*.

Li, H., T. Löhr, A. Sekuboyina, J. Zhang, B. Wiestler, and B. Menze. (2020b). "Domain Adaptive Medical Image Segmentation via Adversarial Learning of Disease-Specific Spatial Patterns". arXiv:2001.09313.

Li, L.-J., R. Socher, and L. Fei-Fei. (2009). "Towards Total Scene Understanding: Classification, Annotation and Segmentation in an Automatic Framework". In: *CVPR*.

Li, P., X. Liang, D. Jia, and E. P. Xing. (2018d). "Semantic-aware Grad-GAN for Virtual-to-Real Urban Scene Adaption ". In: *BMVC*.

Li, Q., A. Arnab, and P. H. S. Torr. (2018e). "Weakly- and Semi-Supervised Panoptic Segmentation". In: *ECCV*.

Li, R., Q. Jiao, W. Cao, H.-S. Wong, and S. Wu. (2020c). "Model Adaptation: Unsupervised Domain Adaptation Without Source Data". In: *CVPR*.

Li, R., K. Li, Y.-C. Kuo, M. Shu, X. Qi, X. Shen, and J. Jia. (2018f). "Referring Image Segmentation via Recurrent Refinement Networks". In: *CVPR*.

Li, S., X. Sui, J. Fu, H. Fu, X. Luo, Y. Feng, X. Xu, Y. Liu, D. Ting, R. Siow, and M. Goh. (2021b). "Few-Shot Domain Adaptation with Polymorphic Transformers". In: *MICCAI*.

Li, Y., X. Chen, Z. Zhu, L. Xie, G. Huang, D. Du, and X. Wang. (2019a). "Attention-Guided Unified Network for Panoptic Segmentation". In: *CVPR*.

Li, Y., H. Qi, J. Dai, X. Ji, and Y. Wei. (2017). "Fully Convolutional Instance-aware Semantic Segmentation". In: *CVPR*.

Li, Y., M. Murias, S. Major, G. Dawson, and D. E. Carlson. (2018g). "Extracting Relationships by Multi-Domain Matching". In: *NeurIPS*.

Li, Y., Y. Yang, W. Zhou, and T. M. Hospedales. (2019b). "Feature-Critic Networks for Heterogeneous Domain Generalization". In: *ICML*.

Li, Y., L. Yuan, and N. Vasconcelos. (2019c). "Bidirectional Learning for Domain Adaptation of Semantic Segmentation". In: *CVPR*.

Li, Z., Z. Chen, F. Yang, W. Li, Y. Zhu, C. Zhao, R. Deng, L. Wu, R. Zhao, M. Tang, and J. Wang. (2021c). "MST: Masked Self-Supervised Transformer for Visual Representation". In: *NeurIPS*.

Li, Z., Y. Gan, X. Liang, Y. Yu, H. Cheng, and L. Lin. (2016). "LSTM-CF: Unifying Context Modeling and Fusion with LSTMs for RGB-D Scene Labeling". In: *ECCV*.

Lian, Q., F. Lv, L. Duan, and B. Gong. (2019). "Constructing Self-motivated Pyramid Curriculums for Cross-Domain Semantic Segmentation: A Non-Adversarial Approach". In: *ICCV*.

Liang, J., D. Hu, and J. Feng. (2020). "Do We Really Need to Access the Source Data? Source Hypothesis Transfer for Unsupervised Domain Adaptation". In: *ICML*.

Liang, S., Y. Li, and R. Srikant. (2018a). "Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks". In: *ICLR*.

Liang, X., X. Shen, J. Feng, L. Lin, and S. Yan. (2016). "Semantic Object Parsing with Graph LSTM". In: *ECCV*.

Liang, X., Y. Wei, X. Shen, J. Yang, L. Lin, and S. Yan. (2018b). "Proposal-free Network for Instance-level Object Segmentation". *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*. 40(12): 2978–2991.

Liang, Y., R. Wakaki, S. Nobuhara, and K. Nishino. (2022). "Multimodal Material Segmentation". In: *CVPR*.

Lin, D., J. Dai, J. Jia, K. He, and J. Sun. (2016). "ScribbleSup: Scribble-Supervised Convolutional Networks for Semantic Segmentation". In: *CVPR*.

Lin, G., A. Milan, C. Shen, and I. Reid. (2017a). "RefineNet: Multi-Path Refinement Networks for High-Resolution Semantic Segmentation". In: *CVPR*.

Lin, T.-Y., P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. (2017b). "Feature Pyramid Networks for Object Detection". In: *CVPR*.

Lin, T.-Y., M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. (2014). "Microsoft COCO: Common Objects in Context". In: *ECCV*.

Liu, C., J. Yuen, and A. Torralba. (2009). "Nonparametric Scene Parsing: Label Transfer via Dense Scene Alignment". In: *CVPR*.

Liu, C., Z. Lin, X. Shen, J. Yang, X. Lu, and A. Yuille. (2017). "Recurrent Multimodal Interaction for Referring Image Segmentation". In: *ICCV*.

Liu, F., Z. Zhou, H. Jang, A. Samsonov, G. Zhao, and R. Kijowski. (2018a). "Deep Convolutional Neural Network and 3D Deformable Approach for Tissue Segmentation in Musculoskeletal Magnetic Resonance Imaging". *Magnetic Resonance in Medicine*. 79: 2379–2391.

Liu, H., C. Peng, C. Yu, J. Wang, X. Liu, G. Yu, and W. Jiang. (2019). "An End-to-End Network for Panoptic Segmentation". In: *CVPR*.

Liu, M.-Y. and O. Tuzel. (2016). "Coupled Generative Adversarial Networks". In: *NeurIPS*.

Liu, Q., Q. Dou, L. Yu, and P. A. Heng. (2020a). "MS-Net: Multi-Site Network for Improving Prostate Segmentation with Heterogeneous MRI Data". *IEEE Transactions on Medical Imaging*. 39(9): 2713–2724.

Liu, S., L. Qi, H. Qin, J. Shi, and J. Jia. (2018b). "Path Aggregation Network for Instance Segmentation". In: *CVPR*.

Liu, W., A. Rabinovich, and A. C. Berg. (2016). "ParseNet: Looking Wider to See Better". In: *ICLR Workshops*.

Liu, X., L. Song, S. Liu, and Y. Zhang. (2021a). "A Review of Deep-Learning-Based Medical Image Segmentation Methods". *Sustainability*. 13(3): 1224.

Liu, X., Z. Guo, S. Li, F. Xing, J. You, C.-C. J. Kuo, G. El Fakhri, and J. Woo. (2021b). "Adversarial Unsupervised Domain Adaptation With Conditional and Label Shift: Infer, Align and Iterate". In: *ICCV*.

Liu, Y., W. Zhang, and J. Wang. (2021c). "Source-Free Domain Adaptation for Semantic Segmentation". In: *CVPR*.

Liu, Z., Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. (2021d). "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows". In: *ICCV*.

Liu, Z., H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. (2022). "A ConvNet for the 2020s". In: *CVPR*.

Liu, Z., Z. Miao, X. Pan, X. Zhan, D. Lin, S. X. Yu, and B. Gong. (2020b). "Open Compound Domain Adaptation". In: *CVPR*.

Long, J., E. Shelhamer, and T. Darrell. (2015a). "Fully Convolutional Networks for Semantic Segmentation". In: *CVPR*.

Long, M., Y. Cao, J. Wang, and M. I. Jordan. (2015b). "Learning Transferable Features with Deep Adaptation Networks". In: *ICML*.

Long, M., Z. Cao, J. Wang, and M. I. Jordan. (2018). "Conditional Adversarial Domain Adaptation". In: *NeurIPS*.

Lowe, D. G. (2004). "Distinctive Image Features from Scale-invariant Keypoints". *International Journal of Computer Vision (IJCV)*. 60: 91–110.

Lucchi, A., Y. Li, X. Boix, K. Smith, and P. Fua. (2011). "Are Spatial and Global Constraints Really Necessary for Segmentation?" In: *ICCV*.

Luo, G., Y. Zhou, X. Sun, L. Cao, C. Wu, C. Deng, and R. Ji. (2020). " Multi-task Collaborative Network for Joint Referring Expression Comprehension and Segmentation". In: *CVPR*.

Luo, W. and M. Yang. (2020). " Semi-supervised Semantic Segmentation via Strong-weak Dual-branch Network". In: *ECCV*.

Luo, Y., P. Liu, T. Guan, J. Yu, and Y. Yang. (2019a). "Significance-aware Information Bottleneck for Domain Adaptive Semantic Segmentation". In: *ICCV*.

Luo, Y., L. Zheng, T. Guan, J. Yu, and Y. Yang. (2019b). "Taking A Closer Look at Domain Shift: Category-level Adversaries for Semantics Consistent Domain Adaptation". In: *CVPR*.

Lv, F., T. Liang, X. Chen, and L. Guosheng. (2020). "Cross-Domain Semantic Segmentation via Domain-Invariant Interactive Relation Transfer ". In: *CVPR*.

Ma, L., Y. Liu, X. Zhang, Y. Yed, G. Yin, and B. A. Johnson. (2019). "Deep Learning in Remote Sensing Applications: A Meta-analysis and Review". *ISPRS Journal of Photogrammetry and Remote Sensing*. 152(6): 166–177.

Ma, X., J. Gao, and C. Xu. (2021). "Active Universal Domain Adaptation". In: *ICCV*.

Mansour, Y., M. Mohri, and A. Rostamizadeh. (2009). "Domain Adaptation with Multiple Sources". In: *NeurIPS*.

Mao, X., Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley. (2017). "Least Squares Generative Adversarial Networks". In: *ICCV*.

Maracani, A., U. Michieli, M. Toldo, and P. Zanuttigh. (2021). "RECALL: Replay-Based Continual Learning in Semantic Segmentation". In: *ICCV*.

Martin, D. R., C. C. Fowlkes, and J. Malik. (2004). "Learning to Detect Natural Image Boundaries using Local Brightness, Color, and Texture Cues". *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*. 26(5): 530–549.

Mei, K., C. Zhu, J. Zou, and S. Zhang. (2020). "Instance Adaptive Self-Training for Unsupervised Domain Adaptation". arXiv:2008.12197.

Metzen, J., M. C. Kumar, T. Brox, and V. Fischer. (2017). "Universal Adversarial Perturbations Against Semantic Image Segmentation". In: *ICCV*.

Michieli, U. and P. Zanuttigh. (2021). "Knowledge Distillation for Incremental Learning in Semantic Segmentation". *Computer Vision and Image Understanding (CVIU)*. 205.

Minaee, S., Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos. (2020). "Image Segmentation Using Deep Learning: A Survey". arXiv:2001.05566.

Mirikharaji, Z. and G. Hamarneh. (2018). "Star Shape Prior in Fully Convolutional Networks for Skin Lesion Segmentation". In: *MICCAI*.

Mittal, S., M. Tatarchenko, and T. Brox. (2021). "Semi-Supervised Semantic Segmentation With High- and Low-Level Consistency". *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*. 43(4): 1369–1379.

Moeskops, P., M. Veta, M. W. Lafarge, K. A. J. Eppenhof, and J. P. W. Pluim. (2017). "Adversarial Training and Dilated Convolutions for Brain MRI Segmentation". In: *MICCAI Workshops*.

Mohan, R. and A. Valada. (2021). "EfficientPS: Efficient Panoptic Segmentation". *International Journal of Computer Vision (IJCV)*. 129: 1551–1579.

Moosavi-Dezfooli, S.-M., A. Fawzi, O. Fawzi, and P. Frossard. (2017). "Universal Adversarial Perturbations". In: *CVPR*.

Mordan, T., A. Saporta, A. Alahi, M. Cord, and P. Pérez. (2020). "Bilinear Multimodal Discriminator for Adversarial Domain Adaptation with Privileged Information". In: *Symposium of the European Association for Research in Transportation (hEART)*.

Mostajabi, M., P. Yadollahpour, and G. Shakhnarovich. (2015). "Feedforward Semantic Segmentation with Zoom-out Features". In: *CVPR*.

Motiian, S., M. Piccirilli, D. A. Adjeroh, and G. Doretto. (2017). "Unified Deep Supervised Domain Adaptation and Generalization". In: *ICCV*.

Mottaghi, R., X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille. (2014). "The Role of Context for Object Detection and Semantic Segmentation in the Wild". In: *CVPR*.

Mousavian, A., H. Pirsiavash, and J. Košecká. (2016). "Joint Semantic Segmentation and Depth Estimation with Deep Convolutional Networks". In: *3DV*.

Murez, Z., S. Kolouri, D. Kriegman, R. Ramamoorthi, and K. Kim. (2018). "Image to Image Translation for Domain Adaptation". In: *CVPR*.

Musto, L. and A. Zinelli. (2020). "Semantically Adaptive Image-to-image Translation for Domain Adaptation of Semantic Segmentation". In: *BMVC*.

Myronenko, A. (2017). "3D MRI Brain Tumor Segmentation using Autoencoder Regularization". In: *MICCAI Workshops*.

Neuhold, G., T. Ollmann, S. R. Bulò, and P. Kontschieder. (2017). "The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes". In: *ICCV*.

Nguyen, V.-A., T. Nguyen, T. Le, Q. H. Tran, and D. Phung. (2021). "STEM: An Approach to Multi-Source Domain Adaptation With Guarantees". In: *ICCV*.

Nguyen-Meidine, L. T., A. Belal, M. Kiran, J. Dolz, L.-A. Blais-Morin, and E. Granger. (2021). "Unsupervised Multi-Target Domain Adaptation Through Knowledge Distillation". In: *WACV*.

Nie, D., L. Wang, E. Adeli, C. Lao, W. Lin, and D. Shen. (2019). "3-D Fully Convolutional Networks for Multimodal Isointense Infant Brain Image Segmentation". *IEEE Transactions on Computers*. 49(3): 1123–1136.

Ning, M., D. Lu, D. Wei, C. Bian, C. Yuan, S. Yu, K. Ma, and Y. Zheng. (2021). "Multi-Anchor Active Domain Adaptation for Semantic Segmentation". In: *ICCV*.

Noh, H., S. Hong, and B. Han. (2015). "Learning Deconvolution Network for Semantic Segmentation". In: *ICCV*.

El-Nouby, A., G. Izacard, H. Touvron, I. Laptev, H. Jegou, and E. Grave. (2021). "Are Large-scale Datasets Necessary for Self-Supervised Pre-training?" arXiv:2112.10740.

Novikov, A. A., D. Lenis, D. Major, J. Hladuvka, M. Wimmer, and K. Bühler. (2018). "Fully Convolutional Architectures for Multi-Class Segmentation in Chest Radiographs". *IEEE Transactions on Medical Imaging*. 37: 1865–1876.

Novosad, P., V. Fonov, and D. L. Collins. (2019). "Unsupervised Domain Adaptation for the Automated Segmentation of Neuroanatomy in MRI: a Deep Learning Approach". bioRxiv:845537.

Novotny, D., S. Albanie, D. Larlus, and A. Vedaldi. (2018). "Semi-convolutional Operators for Instance Segmentation". In: *ECCV*.

Oktay, O., J. Schlemper, L. Le Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert. (2018). "Attention U-Net: Learning Where to Look for the Pancreas". arXiv:1804.0399.

Olsson, V., W. Tranheden, J. Pinto, and L. Svensson. (2021). "ClassMix: Segmentation-based Data Sugmentation for Semi-supervised Learning". In: *WACV*.

Oquab, M., L. Bottou, I. Laptev, and J. Sivic. (2015). "Is Object Localization for Free?-weakly-supervised Learning with Convolutional Neural Networks". In: *CVPR*.

Orbes-Arteaga, M., T. Varsavsky, C. H. Sudre, Z. Eaton-Rosen, L. Haddow Lewis J. Sorensen, M. Nielsen, A. Pai, S. Ourselin, M. Modat, P. Nachev, and M. J. Cardoso. (2019). "Multi-Domain Adaptation in Brain MRI through Paired Consistency and Adversarial Learning". In: *MICCAI Workshops*.

Orbanz, P. and J. M. Buhmann. (2006). "Smooth Image Segmentation by Nonparametric Bayesian Inference". In: *ECCV*.

Ouali, Y., C. Hudelot, and M. Tami. (2020). "Semi-supervised Semantic Segmentation with Cross-consistency Training". In: *CVPR*.

Pan, F., I. Shin, F. Rameau, S. Lee, and I. Kweon. (2020). "Unsupervised Intra-domain Adaptation for Semantic Segmentation through Self-Supervision". In: *CVPR*.

Panareda Busto, P. and J. Gall. (2017). "Open Set Domain Adaptation". In: *ICCV*.

Pantofaru, C., C. Schmid, and M. He. (2008). "Object Recognition by Integrating Multiple Image Segmentations". In: *ECCV*.

Papandreou, G., L.-C. Chen, K. P. Murphy, and A. L. Yuille. (2015). "Weakly-and Semi-Supervised Learning of a Deep Convolutional Network for Semantic Image Segmentation". In: *ICCV*.

Park, T., M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu. (2019). "Semantic Image Synthesis with Spatially-Adaptive Normalization". In: *CVPR*.

Paszke, A., A. Chaurasia, S. Kim, and E. Culurciello. (2016). "ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation". arXiv:1606.02147.

Pathak, D., P. Krüahenbühl, and T. Darrell. (2015). "Constrained Convolutional Neural Networks for Weakly Supervised Segmentation". In: *ICCV*.

Peng, X., Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang. (2019). "Moment Matching for Multi-Source Domain Adaptation". In: *ICCV*.

Peng, X., Y. Li, and K. Saenko. (2020). "Domain2Vec: Domain Embedding for Unsupervised Domain Adaptation". In: *ECCV*.

Perone, C. S., P. Ballester, R. C. Barros, and J. Cohen-Adad. (2019). "Unsupervised Domain Adaptation for Medical Imaging Segmentation with Self-Ensembling". *NeuroImage*. 194: 1–11.

Perronnin, F. and C. R. Dance. (2007). "Fisher Kernels on Visual Vocabularies for Image Categorization". In: *CVPR*.

Pinheiro, P. H. O. and R. Collobert. (2014). "Recurrent Convolutional Neural Networks for Scene Parsing". In: *ICML*.

Pinheiro, P. O., R. Collobert, and P. Dollár. (2015). "Learning to Segment Object Candidates". In: *NeurIPS*.

Plath, N., M. Toussaint, and S. Nakajima. (2009). "Multi-class Image Segmentation using Conditional Random Fields and Global Classfication". In: *ICML*.

Pohlen, T., A. Hermans, M. Mathias, and B. Leibe. (2017). "Full-Resolution Residual Networks for Semantic Segmentation in Street Scenes". In: *CVPR*.

Pont-Tuset, J., P. Arbeláez, J. T. Barron, F. Marques, and J. Malik. (2016). "Multiscale Combinatorial Grouping for Image Segmentation and Object Proposal Generation". *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*. 39(1): 128–140.

Porav, H., T. Bruls, and P. Newman. (2019). "Don't Worry About the Weather: Unsupervised Condition-Dependent Domain Adaptation". In: *IEEE Intelligent Transportation Systems Conference*.

Porzi, L., S. R. Buló, A. Colovic, and P. Kontschieder. (2019). "Seamless Scene Segmentation". In: *CVPR*.

Qi, X., R. Liao, J. Jia, S. Fidler, and R. Urtasun. (2017). "3D Graph Neural Networks for RGBD Semantic Segmentation". In: *ICCV*.

Qi, X., Z. Liu, J. Shi, H. Zhao, and J. Jia. (2016). "Augmented Feedback in Semantic Segmentation under Image Level Supervision". In: *ECCV*.

Qiao, F., L. Zhao, and X. Peng. (2020). "Learning to Learn Single Domain Generalization". In: *CVPR*.

Radford, A., J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, A. Sastry Girishand Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. (2021). "Learning Transferable Visual Models From Natural Language Supervision". In: *ICML*.

Rahman, M. A. and Y. Wang. (2017). "Optimizing Intersection-Over-Union in Deep Neural Networks for Image Segmentation". In: *International Symposium on Visual Computing*.

Rahman, M. M., C. Fookes, M. Baktashmotlagh, and S. Sridharan. (2020). "Correlation-aware Adversarial Domain Adaptation and Generalization". *Pattern Recognition*. 100(107124).

Rakshit, S., D. Tamboli, P. S. Meshram, B. Banerjee, G. Roig, and S. Chaudhuri. (2020). "Multi-source Open-Set Deep Adversarial Domain Adaptation". In: *ECCV*.

Ranftl, R., A. Bochkovskiy, and V. Koltun. (2021). "Vision Transformers for Dense Prediction". In: *ICCV*.

Redondo-Cabrera, C., M. Baptista-Ríos, and R. López-Sastre. (2018). "Learning to Exploit the Prior Network Knowledge for Weakly-Supervised Semantic Segmentation". *IEEE Transactions on Image Processing (TIP)*. 28: 3649–3661.

Ren, M. and R. S. Zemel. (2017). "End-to-End Instance Segmentation with Recurrent Attention". In: *CVPR*.

Ren, S., K. He, R. Girshick, and J. Sun. (2015). "Faster R-CNN: Towards Real-time Object Detection with Region Proposal Networks". In: *NeurIPS*.

Rezaei, M., K. Harmuth, W. Gierke, T. Kellermeier, M. Fischer, H. Yang, and C. Meinel. (2017). "Conditional Adversarial Network for Semantic Segmentation of Brain Tumor". In: *MICCAI Workshops*.

Richter, S. R., V. Vineet, S. Roth, and K. Vladlen. (2016). "Playing for Data: Ground Truth from Computer Games". In: *ECCV*.

Ronneberger, O., P. Fischer, and T. Brox. (2015). "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *MICCAI*.

Ros, G., L. Sellart, J. Materzyńska, D. Vázquez, and A. M. López. (2016). "The SYNTHIA Dataset: a Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes". In: *CVPR*.

Roth, P. M., S. Sternig, H. Grabner, and H. Bischof. (2009). "Classifier grids for robust adaptive object detection". In: *CVPR*.

Rother, C., V. Kolmogorov, and A. Blake. (2004). "GrabCut — Interactive Foreground Extraction using Iterated Graph Cuts". *IEEE Transactions on Graphics*. 23(3): 309–314.

Roy, A. and S. Todorovic. (2017). "Combining Bottom-Up, Top-Down, and Smoothness Cues for Weakly Supervised Image Segmentation". In: *CVPR*.

Roy, S., E. Krivosheev, Z. Zhong, N. Sebe, and E. Ricci. (2021). "Curriculum Graph Co-Teaching for Multi-Target Domain Adaptation". In: *CVPR*.

Ru, L., Y. Zhan, B. Yu, and B. Du. (2022). "Learning Affinity from Attention: End-to-End Weakly-Supervised Semantic Segmentation with Transformers". In: *CVPR*.

Russell, B. C., A. Torralba, K. P. Murphy, and W. T. Freeman. (2008). "LabelMe: a Database and Web-based Tool for Image Annotation". *International Journal of Computer Vision (IJCV)*. 77: 157–173.

Russo, P., T. Tommasi, and B. Caputo. (2019). "Towards Multi-source Adaptive Semantic Segmentation". In: *ICIAP*.

Saito, K., D. Kim, S. Sclaroff, T. Darrell, and K. Saenko. (2021). "Tune It the Right Way: Unsupervised Validation of Domain Adaptation via Soft Neighborhood Density". In: *ICCV*.

Saito, K. and K. Saenko. (2021). "OVANet: One-vs-All Network for Universal Domain Adaptation". In: *ICCV*.

Saito, K., Y. Ushiku, T. Harada, and K. Saenko. (2018a). "Adversarial Dropout Regularization". In: *ICLR*.

Saito, K., K. Watanabe, Y. Ushiku, and T. Harada. (2018b). "Maximum Classifier Discrepancy for Unsupervised Domain Adaptation". In: *CVPR*.

Saito, K., S. Yamamoto, Y. Ushiku, and T. Harada. (2018c). "Open Set Domain Adaptation by Backpropagation". In: *ECCV*.

Sakaridis, C., D. Dai, S. Hecker, and L. Van Gool. (2018). "Model Adaptation with Synthetic and Real Data for Semantic Dense Foggy Scene Understanding". In: *ECCV*.

Sakaridis, C., D. Dai, and L. Van Gool. (2019). "Guided Curriculum Model Adaptation and Uncertainty-Aware Evaluation for Semantic Nighttime Image Segmentation". In: *ICCV*.

Sakaridis, C., D. Dai, and L. Van Gool. (2021). "ACDC: The Adverse Conditions Dataset with Correspondences for Semantic Driving Scene Understanding". In: *ICCV*.

Salamati, N., D. Larlus, G. Csurka, and S. Süsstrunk. (2014). "Incorporating Near-Infrared Information into Semantic Image Segmentation". arXiv:1406.6147.

Sankaranarayanan, S., Y. Balaji, A. Jain, S. Nam Lim, and R. Chellappa. (2018). "Learning from Synthetic Data: Addressing Domain Shift for Semantic Segmentation". In: *CVPR*.

Saporta, A., T.-H. Vu, M. Cord, and P. Pérez. (2020). "ESL: Entropy-guided Self-supervised Learning for Domain Adaptation in Semantic Segmentation". In: *CVPR Workshops*.

Saporta, A., T.-H. Vu, M. Cord, and P. Pérez. (2021). "Multi-Target Adversarial Frameworks for Domain Adaptation in Semantic Segmentation". In: *ICCV*.

Schneider, L., M. Jasch, B. Fröhlich, T. Weber, U. Franke, M. Pollefeys, and M. Rätsch. (2017). "Multimodal Neural Networks: RGB-D for Semantic Segmentation and Object Detection". In: *SCIA*.

Schneider, S., E. Rusak, L. Eck, O. Bringmann, W. Brendel, and M. Bethge. (2020). "Improving robustness against common corruptions by covariate shift adaptation". In: *NeurIPS*.

Schroff, F., A. Criminisi, and A. Zisserman. (2006). "Single-Histogram Class Models for Image Segmentation". In: *ICCVGIP*.

Schwing, A. G. and R. Urtasun. (2015). "Fully Connected Deep Structured Networks". arXiv:1506.04579.

Selvaraju, R. R., M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. (2017). "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization". In: *ICCV*.

Seo, S., Y. Suh, D. Kim, and B. Han. (2020). "Learning to Optimize Domain Specific Normalization for Domain Generalization". In: *ECCV*.

Shen, T., D. Gong, W. Zhang, C. Shen, and T. Mei. (2019). "Regularizing Proxies with Multi-Adversarial Training for Unsupervised Domain-Adaptive Semantic Segmentation". arXiv:1907.12282.

Shen, T., G. Lin, L. Liu, C. Shen, and I. Reid. (2017). "Weakly Supervised Semantic Segmentation Based on Web Image Co-segmentation". In: *BMVC*.

Shen, T., G. Lin, C. Shen, and I. Reid. (2018). "Bootstrapping the Performance of Webly Supervised Semantic Segmentation". In: *CVPR*.

Shin, I., D.-J. Kim, J. Cho, S. Woo, K. Park, and I. S. Kweon. (2021). "LabOR: Labeling Only if Required for Domain Adaptive Semantic Segmentation". In: *ICCV*.

Shin, I., S. Woo, F. Pan, and I. S. Kweon. (2020). "Two-Phase Pseudo Label Densification for Self-training Based Domain Adaptation". In: *ECCV*.

Shotton, J., J. Winn, C. Rother, and A. Criminisi. (2009). "TextonBoost for Image Understanding: Multi-Class Object Recognition and Segmentation by Jointly Modeling Texture, Layout, and Context". *International Journal of Computer Vision (IJCV)*. 81: 2–23.

Silberman, N., D. Hoiem, P. Kohli, and R. Fergus. (2012). "Indoor Segmentation and Support Inference from RGBD Images". In: *ECCV*.

Singh, R., M. Vatsa, V. M. Patel, and N. Ratha, eds. (2020). *Domain Adaptation for Visual Understanding. Image Processing, Computer Vision, Pattern Recognition & Graphics*. Springer.

Sivaprasad, P. T. and F. Fleuret. (2021). "Uncertainty Reduction for Model Adaptation in Semantic Segmentation". In: *CVPR*.

Sofiiuk, K., O. Barinova, and A. Konushin. (2019). "AdaptIS: Adaptive Instance Selection Network". In: *ICCV*.

Song, C., Y. Huang, W. Ouyang, and L. Wang. (2019). "Box-driven Class-wise Region Masking and Filling Rate Guided Loss for Weakly Supervised Semantic Segmentation". In: *CVPR*.

Souly, N., C. Spampinato, and M. Shah. (2017). "Semi-supervised Semantic Segmentation using Generative Adversarial Network". In: *ICCV*.

Soviany, P., R. Ionescu, P. Rota, and N. Sebe. (2021). "Curriculum Learning: A Survey". arXiv:2101.10382.

Strudel, R., R. Garcia, I. Laptev, and C. Schmid. (2021). "Segmenter: Transformer for Semantic Segmentation". In: *ICCV*.

Sun, G., W. Wang, J. Dai, and L. Van Gool. (2020a). "Mining Cross-Image Semantics for Weakly Supervised Semantic Segmentation". In: *ECCV*.

Sun, K., B. Xiao, D. Liu, and J. Wang. (2019a). "Deep High-Resolution Representation Learning for Human Pose Estimation". In: *CVPR*.

Sun, K., Y. Zhao, B. Jiang, T. Cheng, B. Xiao, D. Liu, Y. Mu, X. Wang, W. Liu, and J. Wang. (2019b). "High-Resolution Representations for Labeling Pixels and Regions". arXiv:1904.04514.

Sun, Y., E. Tzeng, T. Darrell, and A. A. Efros. (2019c). "Unsupervised Domain Adaptation through Self-Supervision". arXiv:1909.11825.

Sun, Y., X. Wang, Z. Liu, J. Miller, A. Efros, and M. Hardt. (2020b). "Test-Time Training with Self-Supervision for Generalization under Distribution Shifts". In: *ICML*.

Sun, Y., W. Zuo, and M. Liu. (2019d). "RTFNet: RGB-Thermal Fusion Network for Semantic Segmentation of Urban Scenes". *IEEE Robotics and Automation Letters*. 4(3): 2576–2583.

Szegedy, C., W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. (2014). "Intriguing Properties of Neural Networks". In: *ICLR*.

Taigman, Y., A. Polyak, and L. Wolf. (2017). "Unsupervised Cross-domain Image Generation". In: *ICLR*.

Tang, M., F. Perazzi, A. Djelouah, I. Ben Ayed, C. Schroers, and Y. Boykov. (2018). "On Regularized Losses for Weakly-supervised CNN Segmentation". In: *ECCV*.

Tarvainen, A. and H. Valpola. (2017). "Mean Teachers are Better Role Models: Weight-averaged Consistency Targets Improve Semi-supervised Deep Learning Results". In: *NeurIPS*.

Tasar, O., Y. Tarabalka, A. Giros, P. Alliez, and S. Clerc. (2020). "StandardGAN: Multi-Source Domain Adaptation for Semantic Segmentation of Very High Resolution Satellite Images by Data Standardization". In: *CVPR Workshops*.

Teichmann, M. and R. Cipolla. (2019). "Convolutional CRFs for Semantic Segmentation". In: *BMVC*.

Teichmann, M., M. Weber, M. Zöllner, R. Cipolla, and R. Urtasun. (2018). "MultiNet: Real-time Joint Semantic Reasoning for Autonomous Driving". In: *IEEE Intelligent Vehicles Symposium (IVS)*.

Thoma, M. (2016). "A Survey of Semantic Segmentation". arXiv:1602.06541.

Thomas, C. and A. Kovashka. (2019). "Artistic Object Recognition by Unsupervised Style Adaptation". In: *ACCV*.

Tian, Z., C. Shen, and H. Chen. (2020). "Conditional Convolutions for Instance Segmentation". In: *ECCV*.

Tian, Z., C. Shen, X. Wang, and H. Chen. (2019). "Weakly Supervised Instance Segmentation using the Bounding Box Tightness Prior". In: *NeurIPS*.

Tian, Z., C. Shen, X. Wang, and H. Chen. (2021). "BoxInst: High-Performance Instance Segmentation with Box Annotations". In: *CVPR*.

Tighe, J., M. Niethammer, and S. Lazebnik. (2014). "Scene Parsing with Object Instances and Occlusion Ordering". In: *CVPR*.

Tobin, J., R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel. (2017). "Domain Randomization for Transferring Deep Neural Networks from Simulation to the Real World". In: *IROS*.

Toldo, M., A. Maracani, U. Michieli, and P. Zanuttigh. (2020a). "Unsupervised Domain Adaptation in Semantic Segmentation: a Review". arXiv:2005.10876.

Toldo, M., U. Michieli, G. Agresti, and P. Zanuttigh. (2020b). "Unsupervised Domain Adaptation for Mobile Semantic Segmentation based on Cycle Consistency and Feature Alignment". *Image and Vision Computing*. 95(103889).

Toldo, M., U. Michieli, and P. Zanuttigh. (2021). "Unsupervised Domain Adaptation in Semantic Segmentation via Orthogonal and Clustered Embeddings". In: *WACV*.

Tran, T., O.-H. Kwon, K.-R. Kwon, S.-H. Lee, and K.-W. Kang. (2018). "Blood Cell Images Segmentation using Deep Learning Semantic Segmentation". In: *IEEE International Conference on Electronics and Communication Engineering*.

Tranheden, W., V. Olsson, J. Pinto, and L. Svensson. (2021). "DACS: Domain Adaptation via Cross-domain Mixed Sampling". In: *WACV*.

Truong, T.-D., C. N. Duong, N. Le, S. L. Phung, C. Rainwater, and K. Luu. (2021). "BiMaL: Bijective Maximum Likelihood Approach to Domain Adaptation in Semantic Scene Segmentation". In: *ICCV*.

Tsai, Y.-H., W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker. (2018). "Learning to Adapt Structured Output Space for Semantic Segmentation". In: *CVPR*.

Tsai, Y.-H., K. Sohn, S. Schulter, and M. Chandrakero. (2019). "Domain Adaptation for Structured Output via Discriminative Patch Representations". In: *ICCV*.

Turkmen, S. (2019). "Scene Understanding Through Semantic Image Segmentation in Augmented Reality". *Tech. rep.* University of Oulu.

Tzeng, E., J. Hoffman, T. Darrell, and K. Saenko. (2015). "Simultaneous Deep Transfer Across Domains and Tasks". In: *ICCV*.

Tzeng, E., J. Hoffman, K. Saenko, and T. Darrell. (2017). "Adversarial Discriminative Domain Adaptation". In: *CVPR*.

Ulyanov, D., A. Vedaldi, and V. S. Lempitsky. (2016). "Instance Normalization: The Missing Ingredient for Fast Stylization". arXiv:1607.08022.

Valada, A., J. Vertens, A. Dhall, and W. Burgard. (2017). "AdapNet: Adaptive Semantic Segmentation in Adverse Environmental Conditions". In: *ICRA*.

Valindria, V. V., I. Lavdas, W. Bai, K. Kamnitsas, E. O. Aboagye, A. G. Rockall, D. Rueckert, and B. Glocker. (2018). "Domain Adaptation for MRI Organ Segmentation using Reverse Classification Accuracy". In: *Medical Imaging with Deep Learning*.

Varma, G., A. Subramanian, A. Namboodiri, M. Chandraker, and C. Jawahar. (2019). "IDD: A Dataset for Exploring Problems of Autonomous Navigation in Unconstrained Environments". In: *WACV*.

Vaswani, A., P. Ramachandran, A. Srinivas, N. Parmar, B. Hechtman, and J. Shlens. (2021). "Scaling Local Self-Attention for Parameter Efficient Visual Backbones". In: *CVPR*.

Venkataramani, R., H. Ravishankar, and S. Anamandra. (2019). "Towards Continuous Domain Adaptation For Medical Imaging". In: *International Symposium on Biomedical Imaging (ISBI)*.

Venkateswara, H. and S. Panchanathan, eds. (2020). *Domain Adaptation in Computer Vision with Deep Learning*. Springer.

Verbeek, J. and B. Triggs. (2007a). "Region Classification with Markov Field Aspect Models". In: *CVPR*.

Verbeek, J. and B. Triggs. (2007b). "Scene Segmentation with CRFs Learned from Partially Labeled Images". In: *NeurIPS*.

Vernaza, P. and M. Chandraker. (2017). "Learning Random-walk Label Propagation for Weakly-supervised Semantic Segmentation". In: *CVPR*.

Vezhnevets, A. and J. M. Buhmann. (2010). "Towards Weakly Supervised Semantic Segmentation by Means of Multiple Instance and Multitask Learning". In: *CVPR*.

Visin, F., M. Ciccone, A. Romero, K. Kastner, K. Cho, Y. Bengio, M. Matteucci, and A. Courville. (2016). "ReSeg: A Recurrent Neural Network-based Model for Semantic Segmentation". In: *CVPR Workshops*.

Volpi, R., P. De Jorge, D. Larlus, and G. Csurka. (2022). "On the Road to Online Adaptation for Semantic Image Segmentation". In: *CVPR*.

Volpi, R., D. Larlus, and G. Rogez. (2021). "Continual Adaptation of Visual Representations via Domain Randomization and Meta-Learning". In: *CVPR*.

Volpi, R. and V. Murino. (2019). "Model Vulnerability to Distributional Shifts over Image Transformation Sets". In: *ICCV*.

Volpi, R., H. Namkoong, O. Sener, J. C. Duchi, V. Murino, and S. Savarese. (2019). "Generalizing to Unseen Domains via Adversarial Data Augmentation". In: *NeurIPS*.

Vu, T.-H., H. Jain, M. Bucher, M. Cord, and P. Pérez. (2019a). "ADVENT: Adversarial Entropy Minimization for Domain Adaptation in Semantic Segmentation". In: *CVPR*.

Vu, T.-H., H. Jain, M. Bucher, M. Cord, and P. Pérez. (2019b). "DADA: Depth-Aware Domain Adaptation in Semantic Segmentation". In: *ICCV*.

Wang, D., E. Shelhamer, S. Liu, B. Olshausen, and T. Darrell. (2021a). "TENT: Fully Test-Time Adaptation by Entropy Minimization". In: *ICLR*.

Wang, D., C. Gu, K. Wu, and X. Guan. (2017). "Adversarial Neural Networks for Basal Membrane Segmentation of Microinvasive Cervix Carcinoma in Histopathology Images". In: *ICMLC*.

Wang, H., Y. Zhu, B. Green, H. Adam, A. Yuille, and L.-C. Chen. (2020a). "Axial-DeepLab: Stand-Alone Axial-Attention for Panoptic Segmentation". In: *ECCV*.

Wang, J., C. Lan, C. Liu, Y. Ouyang, and T. Qin. (2020b). "Generalizing to Unseen Domains: A Survey on Domain Generalization". arXiv:2103.03097.

Wang, K., C. Yang, and M. Betke. (2021b). "Consistency Regularization with High-dimensional Non-adversarial Source-guided Perturbation for Unsupervised Domain Adaptation in Segmentation". In: *AAAI*.

Wang, M. and W. Deng. (2018). "Deep Visual Domain Adaptation: A Survey". *Neurocomputing*. 312: 135–153.

Wang, P., P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell. (2018a). "Understanding Convolution for Semantic Segmentation". In: *WACV*.

Wang, P., X. Shen, Z. Lin, S. Cohen, B. Price, and A. Yuille. (2015). "Towards Unified Depth and Semantic Prediction from a Single Image". In: *CVPR*.

Wang, Q., O. Fink, L. Van Gool, and D. Dai. (2022a). "Continual Test-Time Domain Adaptation". In: *CVPR*.

Wang, S., L. Yu, C. Li, C.-W. Fu, and P.-A. Heng. (2020c). "Learning from Extrinsic and Intrinsic Supervisions for Domain Generalization". In: *ECCV*.

Wang, W., E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao. (2021c). "Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions". In: *ICCV*.

Wang, X., S. You, X. Li, and H. Ma. (2018b). "Weakly-Supervised Semantic Segmentation by Iteratively Mining Common Object Features". In: *CVPR*.

Wang, X., T. Kong, C. Shen, Y. Jiang, and L. Li. (2020d). "SOLO: Segmenting Objects by Locations". In: *ECCV*.

Wang, Y., J. Peng, and Z. Zhang. (2021d). "Uncertainty-Aware Pseudo Label Refinery for Domain Adaptive Semantic Segmentation". In: *ICCV*.

Wang, Z., Y. Lu, Q. Li, X. Tao, Y. Guo, M. Gong, and T. Liu. (2022b). "CRIS: CLIP-Driven Referring Image Segmentation". In: *CVPR*.

Wang, Z., Y. Wei, R. Feris, J. Xiong, W.-M. Hwu, T. S. Huang, and H. Shi. (2020e). "Alleviating Semantic-level Shift: A Semi-supervised Domain Adaptation Method for Semantic Segmentation". In: *CVPR Workshops*.

Wang, Z., M. You, Y. Wei, R. Feris, J. Xiong, W.-m. Hwu, T. S. Huang, and H. Shi. (2020f). "Differential Treatment for Stuff and Things: A Simple Unsupervised Domain Adaptation Method for Semantic Segmentation". In: *CVPR*.

Wei, C., H. Fan, S. Xie, C. Wu, A. L. Yuille, and C. Feichtenhofer. (2021). "Masked Feature Prediction for Self-Supervised Visual Pre-Training". arXiv:2112.09133.

Wilson, G. and D. J. Cook. (2020). "A Survey of Unsupervised Deep Domain Adaptation". *IEEE Transactions on Intelligent Systems and Technology*. 11(5).

Winn, J. and J. Shotton. (2006). "The Layout Consistent Random Field for Recognizing and Segmenting Partially Occluded Objects". In: *CVPR*.

Wrenninge, M. and J. Unger. (2018). "Synscapes: A Photorealistic Synthetic Dataset for Street Scene Parsing". arXiv:1810.08705.

Wu, F.-Y. (1982). "The Potts Model". *Reviews of Modern Physics*. 54(1): 235–268.

Wu, Z., C. Shen, and A. van den Hengel. (2017). "Real-time Semantic Image Segmentation via Spatial Sparsity". arXiv:1712.00213.

Wu, Z., X. Han, Y.-L. Lin, M. Gokhan Uzunbas, T. Goldstein, S. Nam Lim, and L. S. Davis. (2018). "DCAN: Dual Channel-wise Alignment Networks for Unsupervised Scene Adaptation". In: *ECCV*.

Wu, Z., X. Wang, J. E. Gonzalez, T. Goldstein, and L. S. Davis. (2019). "ACE: Adapting to Changing Environments for Semantic Segmentation". In: *ICCV*.

Wulfmeier, M., A. Bewley, and I. Posner. (2017). "Addressing Appearance Change in Outdoor Robotics with Adversarial Domain Adaptations". In: *IROS*.

Xia, H., H. Zhao, and Z. Ding. (2021). "Adaptive Adversarial Network for Source-Free Domain Adaptation". In: *ICCV*.

Xia, W., C. Domokos, J. Dong, L.-F. Cheong, and S. Yan. (2013). "Semantic Segmentation without Annotating Segments". In: *CVPR*.

Xiao, T., Y. Liu, B. Zhou, Y. Jiang, and J. Sun. (2018). "Unified Perceptual Parsing for Scene Understanding". In: *ECCV*.

Xie, B., L. Yuan, S. Li, C. H. Liu, and X. Cheng. (2022a). "Towards Fewer Annotations: Active Learning via Region Impurity and Prediction Uncertainty for Domain Adaptive Semantic Segmentation". In: *CVPR*.

Xie, C., J. Wang, Z. Zhang, Y. Zhou, L. Xie, and A. L. Yuille. (2017). "Adversarial Examples for Semantic Segmentation and Object Detection". In: *ICCV*.

Xie, E., W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo. (2021). "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers". In: *NeurIPS*.

Xie, Z., Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu. (2022b). "SimMIM: A Simple Framework for Masked Image Modeling". In: *CVPR*.

Xiong, Y., R. Liao, H. Zhao, R. Hu, M. Bai, E. Yumer, and R. Urtasun. (2019). "UPSNet: A Unified Panoptic Segmentation Network". In: *CVPR*.

Xu, D., W. Ouyang, X. Wang, and N. Sebe. (2018). "PAD-Net: Multi-Tasks Guided Prediction-and-Distillation Network for Simultaneous Depth Estimation and Scene Parsing". In: *CVPR*.

Xu, J., A. G. Schwing, and R. Urtasun. (2015). "Learning to Segment Under Various Forms of Weak Supervision". In: *CVPR*.

Xu, J., L. Xiao, and A. M. López. (2019a). "Self-Supervised Domain Adaptation for Computer Vision Tasks". *IEEE Access*. 7: 156694–156706.

Xu, L., W. Ouyang, M. Bennamoun, F. Boussaid, and D. Xu. (2022). "Multi-class Token Transformer for Weakly Supervised Semantic Segmentation". In: *CVPR*.

Xu, W., Y. Xu, T. Chang, and Z. Tu. (2021). "Co-Scale Conv-Attentional Image Transformers". In: *ICCV*.

Xu, Y., B. Du, L. Zhang, Q. Zhang, G. Wang, and L. Zhang. (2019b). "Self-Ensembling Attention Networks: Addressing Domain Shift for Semantic Segmentation". In: *AAAI*.

Xu, Z., W. Li, L. Niu, and D. Xu. (2014). "Exploiting Low-rank Structure from Latent Domains for Domain Generalization". In: *ECCV*.

Yang, J., R. Xu, R. Li, X. Qi, X. Shen, G. Li, and L. Lin. (2020a). "An Adversarial Perturbation Oriented Domain Adaptation Approach for Semantic Segmentation". In: *AAAI*.

Yang, J., W. An, S. Wang, X. Zhu, C. Yan, and J. Huang. (2020b). "Label-driven Reconstruction for Domain Adaptation in Semantic Segmentation". In: *ECCV*.

Yang, J., C. Li, W. An, H. Ma, Y. Guo, Y. Rong, P. Zhao, and J. Huang. (2021). "Exploring Robustness of Unsupervised Domain Adaptation in Semantic Segmentation". In: *ICCV*.

Yang, L., W. Zhuo, ei Qi, Y. Shi, and Y. Gao. (2022). "ST++: Make Self-training Work Better for Semi-supervised Semantic Segmentation". In: *CVPR*.

Yang, L., P. Meer, and D. J. Foran. (2007). "Multiple Class Segmentation Using A Unified Framework over Mean-Shift Patches". In: *CVPR*.

Yang, L., Y. Balaji, S.-N. Lim, and A. Shrivastava. (2020c). "Curriculum Manager for Source Selection in Multi-Source Domain Adaptation". In: *ECCV*.

Yang, M., K. Yu, C. Zhang, Z. Li, and K. Yang. (2018). "DenseASPP for Semantic Segmentation in Street Scenes". In: *CVPR*.

Yang, T.-J., M. D. Collins, Y. Zhu, J.-J. Hwang, T. Liu, X. Zhang, V. Sze, G. Papandreou, and L.-C. Chen. (2019). "DeeperLab: Single-Shot Image Parser". arXiv:1902.05093.

Yang, Y., D. Lao, G. Sundaramoorthi, and S. Soatto. (2020d). "Phase Consistent Ecological Domain Adaptation". In: *CVPR*.

Yang, Y. and S. Soatto. (2020). "FDA: Fourier Domain Adaptation for Semantic Segmentation". In: *CVPR*.

Yang, Y., S. Hallman, D. Ramanan, and C. C. Fowlkes. (2012). "Layered Object Models for Image Segmentation". *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*. 34(9): 1731–1743.

Ye, C., W. Wang, S. Zhang, and K. Wang. (2019a). "Multi-Depth Fusion Network for Whole-Heart CT Image Segmentation". *IEEE Access*. 7: 23421–23429.

Ye, L., M. Rochan, Z. Liu, and Y. Wang. (2019b). "Cross-Modal Self-Attention Network for Referring Image Segmentation". In: *CVPR*.

You, K., X. Wang, M. Long, and M. I. Jordan. (2019). "Towards Accurate Model Selection in Deep Unsupervised Domain Adaptation". In: *ICML*.

Yu, C., J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang. (2018a). "BiSeNet: Bilateral Segmentation Network for Real-time Semantic Segmentation". In: *ECCV*.

Yu, F., H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell. (2020). "BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning". In: *CVPR*.

Yu, F. and V. Koltun. (2015). "Multi-Scale Context Aggregation by Dilated Convolutions". arXiv:1511.07122.

Yu, L., Z. Lin, X. Shen, J. Yang, X. Lu, M. Bansal, and T. L. Berg. (2018b). "MAttNet: Modular Attention Network for Referring Expression Comprehension". In: *CVPR*.

Yuan, J., Y. Liu, C. Shen, Z. Wang, and H. Li. (2021a). "A Simple Baseline for Semi-Supervised Semantic Segmentation With Strong Data Augmentation". In: *ICCV*.

Yuan, L., D. Chen, Y.-L. Chen, N. Codella, X. Dai, J. Gao, H. Hu, X. Huang, B. Li, C. Li, C. Liu, M. Liu, Z. Liu, Y. Lu, Y. Shi, L. Wang, J. Wang, B. Xiao, Z. Xiao, J. Yang, M. Zeng, L. Zhou, and P. Zhang. (2021b). "Florence: A New Foundation Model for Computer Vision". arXiv:2111.11432.

Yuan, Y., X. Chen, and J. Wang. (2020). "Object-Contextual Representations for Semantic Segmentation". In: *ECCV*.

Yue, X., Y. Zhang, S. Zhao, A. Sangiovanni-Vincentelli, K. Keutzer, and B. Gong. (2019). "Domain Randomization and Pyramid Consistency: Simulation-to-Real Generalization without Accessing Target Domain Data". In: *ICCV*.

Yun, S., D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo. (2019). "CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Feature". In: *ICCV*.

Yurtsever, E., J. Lambert, A. Carballo, and K. Takeda. (2020). "A Survey of Autonomous Driving: Common Practices and Emerging Technologies". *IEEE Access*. 8: 58443–58469.

Zang, D., L. Yang, D. Meng, D. Xu, and J. Han. (2017). "SPFTN: A Self-Paced Fine-Tuning Network for Segmenting Objects in Weakly Labelled Videos". In: *CVPR*.

Zhan, X., Z. Liu, P. Luo, X. Tang, and C. C. Loy. (2017). "Mix-and-Match Tuning for Self-Supervised Semantic Segmentation". In: *AAAI*.

Zhang, J., Z. Ding, W. Li, and P. Ogunbona. (2018a). "Importance Weighted Adversarial Nets for Partial Domain Adaptation". In: *CVPR*.

Zhang, L. and X. Gao. (2019). "Transfer Adaptation Learning: A Decade Survey". arXiv:1903.04687.

Zhang, M., Y. Zhou, J. Zhao, Y. Man, B. Liu, and R. Yao. (2020a). "A Survey of Semi- and Weakly Supervised Semantic Segmentation of Images". *Artificial Intelligence Review*. 53: 2402–2417.

Zhang, Q., J. Zhang, W. Liu, and D. Tao. (2019). "Category Anchor-Guided Unsupervised Domain Adaptation for Semantic Segmentation". In: *NeurIPS*.

Zhang, R., P. Isola, and A. A. Efros. (2016a). "Colorful Image Colorization". In: *ECCV*.

Zhang, W., Z. Huang, G. Luo, T. Chen, X. Wang, W. Liu, G. Yu, and C. Shen. (2022). "TopFormer: Token Pyramid Transformer for Mobile Semantic Segmentation". In: *CVPR*.

Zhang, Y., P. David, H. Foroosh, and B. Gong. (2020b). "A Curriculum Domain Adaptation Approach to the Semantic Segmentation of Urban Scenes". *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*. 42(8): 1823–1841.

Zhang, Y., D. Sidibé, O. Morel, and F. Mériaudeau. (2021). "Deep Multimodal Fusion for Semantic Image Segmentation: A Survey". *Image and Vision Computing*. 105(104042).

Zhang, Y., Z. Qiu, T. Yao, C.-W. Ngo, D. Liu, and T. Mei. (2020c). "Transferring and Regularizing Prediction for Semantic Segmentation". In: *CVPR*.

Zhang, Y. and A. C. S. Chung. (2018). "Deep Supervision with Additional Labels for Retinal Vessel Segmentation Task". In: *MICCAI*.

Zhang, Z., Z. Cui, C. Xu, Z. Jie, X. Li, and J. Yang. (2018b). "Joint Task-Recursive Learning for Semantic Segmentation and Depth Estimation". In: *ECCV*.

Zhang, Z., S. Fidler, and R. Urtasun. (2016b). "Instance-Level Segmentation for Autonomous Driving with Deep Densely Connected MRFs". In: *CVPR*.

Zhao, H., R. T. des Combes, K. Zhang, and G. Gordon. (2019a). "On Learning Invariant Representation for Domain Adaptation". In: *ICML*.

Zhao, H., S. Zhang, G. Wu, J. M. F. Moura, J. P. Costeira, and G. J. Gordon. (2018a). "Adversarial Multiple Source Domain Adaptation". In: *NeurIPS*.

Zhao, H., X. Qi, X. Shen, J. Shi, and J. Jia. (2018b). "ICNet for Real-Time Semantic Segmentation on High-Resolution Images". In: *ECCV*.

Zhao, H., J. Shi, X. Qi, X. Wang, and J. Jia. (2017). "Pyramid Scene Parsing Network". In: *CVPR*.

Zhao, H., Y. Zhang, S. Liu, J. Shi, C. Loy, D. Lin, and J. Jia. (2018c). "PSANet: Point-wise Spatial Attention Network for Scene Parsing". In: *ECCV*.

Zhao, S., B. Li, X. Yue, Y. Gu, P. Xu, R. Hu, H. Chai, and K. Keutzer. (2019b). "Multi-source Domain Adaptation for Semantic Segmentation". In: *NeurIPS*.

Zhao, S., G. Wang, S. Zhang, Y. Gu, Y. Li, Z. Song, P. Xu, R. Hu, H. Chai, and K. Keutzer. (2020). "Multi-source Distilling Domain Adaptation". In: *AAAI*.

Zheng, S., S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr. (2015). "Conditional Random Fields as Recurrent Neural Networks". In: *ICCV*.

Zheng, S., J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr, and L. Zhang. (2021). "Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers". In: *CVPR*.

Zheng, S., Y. Song, T. Leung, and I. Goodfellow. (2016). "Improving the Robustness of Deep Neural Networks via Stability Training". In: *CVPR*.

Zheng, Z. and Y. Yang. (2020). "Unsupervised Scene Adaptation with Memory Regularization in vivo". In: *IJCAI*.

Zheng, Z. and Y. Yang. (2021). "Rectifying Pseudo Label Learning via Uncertainty Estimation for Domain Adaptive Semantic Segmentation". *International Journal of Computer Vision (IJCV)*. 129: 1106–1120.

Zhong, E., W. Fan, Q. Yang, O. Verscheure, and J. Ren. (2010). "Cross Validation Framework to Choose Amongst Models and Datasets for Transfer Learning". In: *PKDD*.

Zhong, Y., B. Yuan, H. Wu, Z. Yuan, J. Peng, and Y.-X. Wang. (2021). "Pixel Contrastive-Consistent Semi-Supervised Semantic Segmentation". In: *ICCV*.

Zhou, B., A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. (2016). "Learning Deep Features for Discriminative Localization". In: *CVPR*.

Zhou, B., H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba. (2018). "A Review of Semantic Segmentation using Deep Neural Networks". *International Journal of Multimedia Information Retrieval (IJMIR)*. 7: 87–93.

Zhou, B., H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba. (2019a). "Semantic Understanding of Scenes through the ADE20k Dataset". *International Journal of Computer Vision (IJCV)*. 127: 302–321.

Zhou, J., C. Wei, H. Wang, W. Shen, C. Xie, A. Yuille, and T. Kong. (2022). "iBOT:Image BERT Pre-training with Online Tokenizer". In: *ICLR*.

Zhou, K., Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy. (2020a). "Domain Generalization: A Survey". arXiv:2103.02503.

Zhou, K., Y. Yang, Y. Qiao, and T. Xiang. (2020b). "Domain Adaptive Ensemble Learning". arXiv:2003.07325.

Zhou, T., S. Ruan, and S. Canu. (2019b). "A Review: Deep Learning for Medical Image Segmentation using Multi-modality Fusion". *Array*. 3-4(100004).

Zhou, Y., H. Xu, W. Zhang, B. Gao, and P.-A. Heng. (2021). "C^3-SemiSeg: Contrastive Semi-supervised Segmentation via Cross-set Learning and Dynamic Class-balancing". In: *ICCV*.

Zhou, Z.-H. and M. Li. (2005). "Tri-training: Exploiting Unlabeled Data using Three Classifiers". *IEEE Transactions on Knowledge and Data Engineering*. 11(17): 1529–1541.

Zhu, J.-Y., T. Park, P. Isola, and A. A. Efros. (2017). "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks". In: *ICCV*.

Zhu, X., H. Zhou, C. Yang, J. Shi, and D. Lin. (2018). "Penalizing Top Performers: Conservative Loss for Semantic Segmentation Adaptation". In: *ECCV*.

Zhu, Y., Z. Zhang, C. Wu, Z. Zhang, T. He, H. Zhang, R. Manmatha, M. Li, and A. Smola. (2021). "Improving Semantic Segmentation via Efficient Self-Training". *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*.

Zhu, Y., F. Zhuang, and D. Wang. (2019). "Aligning Domain-Specific Distribution and Classifier for Cross-Domain Classification from Multiple Sources". In: *AAAI*.

Zou, Y., Z. Yu, B. V. Kumar, and J. Wang. (2018). "Unsupervised Domain Adaptation for Semantic Segmentation via Class-Balanced Self-Training". In: *ECCV*.

Zou, Y., Z. Yu, X. Liu, B. V. K. V. Kumar, and J. Wang. (2019). "Confidence Regularized Self-Training". In: *ICCV*.

Zurbrügg, R., H. Blum, C. Cadena, R. Siegwart, and L. Schmid. (2022). "Embodied Active Domain Adaptation for Semantic Segmentation via Informative Path Planning". arXiv:2203.00549.