technologies

MDPI

Review

Unsupervised Domain Adaptation in Semantic Segmentation: A Review

Marco Toldo , Andrea Maracani, Umberto Michieli and Pietro Zanuttigh *

Department of Information Engineering, University of Padova, 35131 Padova, Italy;
toldomarco@dei.unipd.it (M.T.); maracani@dei.unipd.it (A.M.) ; umberto.michieli@dei.unipd.it (U.M.)
* Correspondence: zanuttigh@dei.unipd.it

Received: 22 May 2020; Accepted: 17 June 2020; Published: 21 June 2020

Abstract: The aim of this paper is to give an overview of the recent advancements in the Unsupervised Domain Adaptation (UDA) of deep networks for semantic segmentation. This task is attracting a wide interest since semantic segmentation models require a huge amount of labeled data and the lack of data fitting specific requirements is the main limitation in the deployment of these techniques. This field has been recently explored and has rapidly grown with a large number of ad-hoc approaches. This motivates us to build a comprehensive overview of the proposed methodologies and to provide a clear categorization. In this paper, we start by introducing the problem, its formulation and the various scenarios that can be considered. Then, we introduce the different levels at which adaptation strategies may be applied: namely, at the input (image) level, at the internal features representation and at the output level. Furthermore, we present a detailed overview of the literature in the field, dividing previous methods based on the following (non mutually exclusive) categories: adversarial learning, generative-based, analysis of the classifier discrepancies, self-teaching, entropy minimization, curriculum learning and multi-task learning. Novel research directions are also briefly introduced to give a hint of interesting open problems in the field. Finally, a comparison of the performance of the various methods in the widely used autonomous driving scenario is presented.

Keywords: unsupervised domain adaptation; semantic segmentation; unsupervised learning; deep learning; transfer learning; survey

1. Introduction

Over the past few years, deep learning techniques have shown impressive results and have achieved great success in many visual applications. However, they typically require a huge amount of labeled data matching the considered scenario to obtain reliable performances. The collection and annotation of large datasets for every new task and domain is extremely expensive, time-consuming and error-prone. Furthermore, in many scenarios sufficient training data may not be available for various reasons, but it often happens that a large amount of data is available for other domains and tasks that are in some way related to the considered one. Hence, the ability to use a model trained on correlated samples from a different task would highly benefit real-world applications for which there is scarce data [1]. These considerations are especially true for semantic segmentation, where the learning architectures require a huge amount of manually labeled data, which is extremely expensive to obtain since a per-pixel labeling is needed.

1.1. Semantic Segmentation

Semantic segmentation is one of the most challenging tasks in automatic visual understanding, leading to a deeper understanding of the image content if compared with simpler problems like image classification or object detection. An overview of the most common visual tasks is given in Figure 1.

In *image classification*, a single label is assigned to the whole image and denotes the pre-dominant object in the scene. In *object detection*, the objects are identified by means of a bounding box and a label is assigned to each box. In *image segmentation*, the scene is clustered into regions corresponding to the various objects and structures but the regions are not labeled. *Semantic segmentation*, instead, is the task of assigning to each pixel in the image a label corresponding to its semantic content. For this reason, it is often referred to as a dense labeling task as opposed to other simpler problems where fewer labels are given as output. Semantic segmentation is a very wide research field and a huge number of approaches have been proposed to tackle it. In particular, deep learning architectures have recently allowed to obtain substantial improvements.

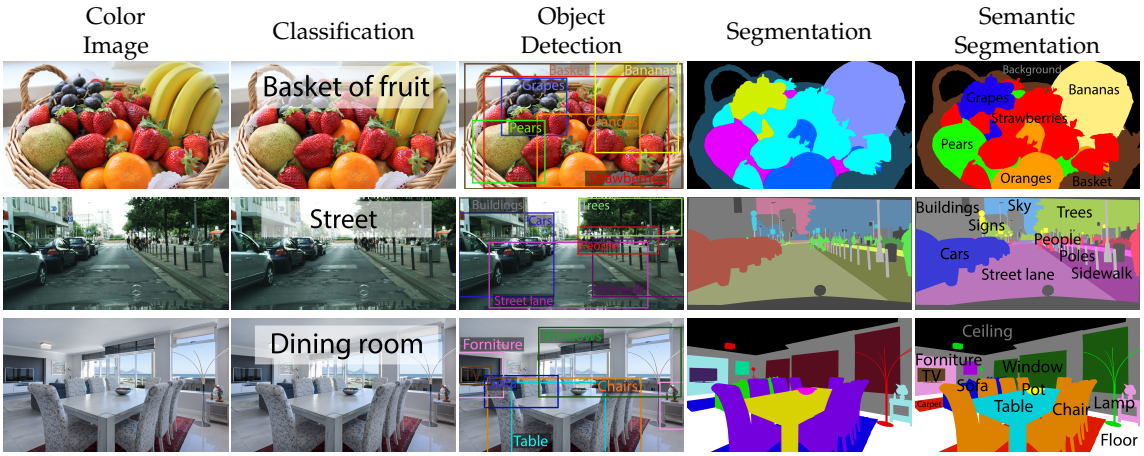

Figure 1. Overview of some possible visual tasks on a few sample images from classification (sparse task) to semantic segmentation (dense task).

Historically, semantic segmentation has moved its origins as an enriched representation and understanding of the scene with respect to the simpler task of image classification: the advent of novel problems to address requiring a higher level of interpretation of the scenes and the possibility to accomplish it, thanks to novel architectures and paradigms (e.g., deep learning), have paved the way to the wide success of semantic image segmentation.

While image classification allows to classify what is contained in an image at a global level (i.e., one label is assigned to each image), semantic image segmentation generates a pixel-wise mask of each object in the images (i.e., one label is assigned to each pixel of each image). Being the former a much simpler task, it has been tackled for a long time with both traditional techniques based on a feature extraction step (e.g., using SIFT or other feature extractors) followed by a classification stage (e.g., using SVM, LDA or Random Forests) and, more recently, with deep learning ones. For this reason, some early-stage works in semantic segmentation build up from classification works, adapting and extending them. The most recent state-of-the-art approaches rely on an autoencoder structure, composed by an encoder and a decoder in order to extract global semantic clues, while retaining input spatial dimensionality.

Starting from the well-known Fully Convolutional Networks (FCN) architecture [2], many models have been proposed, such as PSPNet [3], DRN [4] and the various versions of the DeepLab architecture [5–7]. These models can achieve impressive performance, but this is strictly related to the availability of a massive amount of labeled data required for their training. For this reason, even though the pixel-wise annotation procedure is highly expensive and time consuming, many datasets have been created: for example, Cityscapes [8] and Mapillary [9] for urban scenes; Pascal VOC [10], MS-COCO [11] and ADE20K [12] for visual objects in common contexts; NYUD-v2 [13] and SUN-RGBD [14] for indoor scenes with depth information. In light of these considerations, many recent works try to exploit knowledge extracted from other sources or domains, where labels are plentiful and easily accessible, to reduce the amount of required manually annotated data.

1.2. Domain Adaptation (DA)

Most machine learning models, including Neural Networks (NNs), typically assume that training and test samples are drawn according to the same distribution. However, there are many cases in practical settings where the training and the test data distributions differ. In this survey we focus on the case where a model is trained in one or more domains (called source domains) and then applied in another different, but related, domain (called target domain) [15]. Such learning task is known as Domain Adaptation (DA) and is a fundamental problem in machine learning. Nowadays, it has gained wide attention from the scientific community and represents a long-standing problem in many real-world applications, such as computer vision [16], natural language processing [17], sentiment analysis [18], email filtering, and several others.

Domain Adaptation can be regarded as a particular case of Transfer Learning (TL) that utilizes labeled data in one or more relevant source domains to execute new tasks in a target domain. The aim of DA methodologies is to address the distribution change or the domain shift, which typically greatly degrades the performance of the models [19]. Hence, the domain adaptation problem can be considered a natural extension of the semi-supervised learning task, as the challenge derives not only from the presence of unlabeled training samples, but also from an additional statistical discrepancy occurring between the supervised (source) and unsupervised (target) sets of data. Over the past decades, various DA methods have been proposed to address the shifts between the source and target domains for both traditional machine learning strategies and recent deep learning architectures. The intrinsic nature of source and target domains highly influence the final performance of the DA algorithms. Indeed, they are assumed to be somehow related to each other, but not identical. The more correlated they are, the easier the DA task becomes, allowing to achieve high results on the test data. Hence, a key ingredient for a well-performing strategy is the ability to discover suitable source data to extract useful clues from.

Good reviews of the domain adaptation field can be found in [1,15,16,19,20], which provide a comprehensive sight of the domain adaptation problem in its theoretical form [15,19] or its generic application to visual tasks [1,16,20]. Our work differs from those in that it specifically addresses unsupervised domain adaptation for semantic segmentation. Indeed, we are motivated by the fact that this research area has recently attracted huge interest and a remarkable effort has been undertaken for its solution.

1.3. Unsupervised Domain Adaptation (UDA)

The domain adaptation task can be performed using only data from the source domain or using also some samples from the target domain. The simplest solution that could be adopted is to train only on labeled samples from the source domain without using data from the target domain, hoping that no adaptation is needed (source only). In practice, this leads to poor performances, even when only a small visual domain shift exists. To cope with this, UDA approaches exploit labeled samples from the source domain and unlabeled samples from the target one (source to target UDA).

Especially in the semantic segmentation task where a pixel-by-pixel labeling is required, the samples annotation is the most demanding task, while data acquisition is much simpler and cheaper. For this reason, in this survey we will cover the scenario that takes the name of *Unsupervised Domain Adaptation (UDA)*. Indeed, it is the most interesting in our specific setting since there is no direct supervision on the target domain (i.e., no labels of the target domain are required).

In this scenario, the typical assumption is that the source and target domains are different but in some way related (e.g., the source could be synthetically generated data resembling the real world representations in the target one). Typically, an initial supervised training on the source domain is adapted to the target one by means of various unsupervised learning strategies aiming at achieving good performance also on the target domain (for which no labels are available). In the standard setting, the set of target classes are the same, but advanced settings where also the target labels change

can be considered (see Section 2.1). Figure 2 shows the typical flow of source and target data in the unsupervised domain adaptation process.

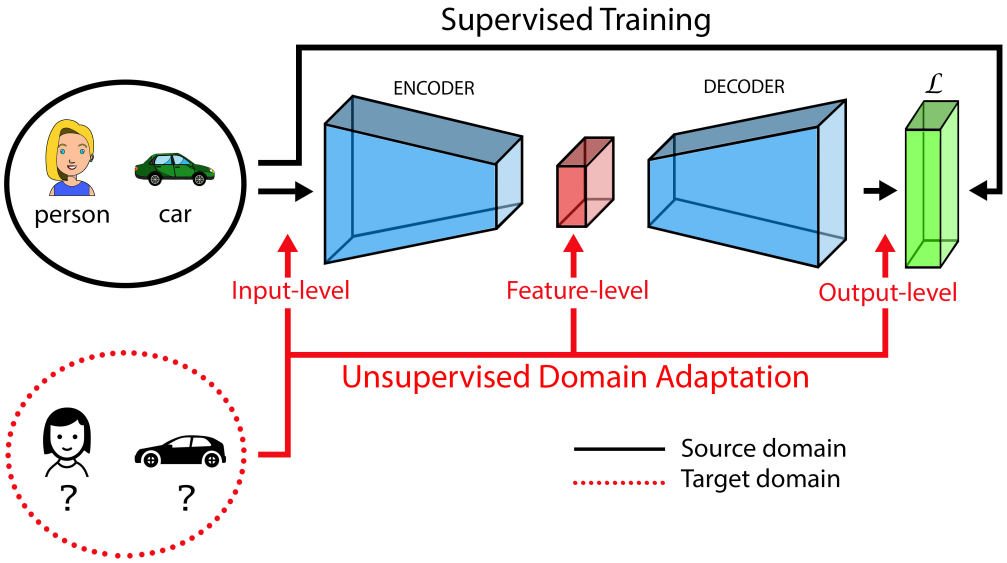

Figure 2. Graphical representation of the unsupervised domain adaptation process. A task-loss $\mathcal{L}$ (e.g., a cross-entropy loss) is used for a supervised training stage on the source domain using the semantic annotations. Unsupervised adaptation to target data without labels can be performed at different levels (e.g., input, features or output) with different strategies.

1.4. Application Motivations

There exists a large number of applications that may significantly benefit from UDA. In general, each application focuses on a very peculiar setting with images taken with a specific camera and a particular environment to solve a prefixed task. The first and easiest solution is to get as much labeled data as possible for the specific problem, but, as already mentioned, this is unfortunately very time consuming and expensive, thus making it unfeasible in many real-world contexts. On the other hand, large and publicly available labeled datasets typically contain generic data and their direct use in specific applications does not grant good performance in the relevant application-specific domain. A second solution would be to transfer source knowledge acquired on a broader scenario and adapt it to the specific setup being targeted. Such context, for example, is fairly common in industrial applications.

An example application is face recognition, which represents a challenging problem that has been actively researched for many years. Current models for face recognition perform very well when training and testing images are acquired under controlled conditions. However, their accuracy quickly degrades when the test images contain variations that are not present in the training images [20]. For instance, these variations could be changed in pose, illumination or viewpoint, and depending on the composition of training and test sets, this can be regarded as a domain adaptation problem [20,21].

Another straightforward application lies in object recognition, where one may be interested in adapting object detection capabilities from a typically larger set to a specific small-size dataset [22].

Furthermore, the recent improvements in the computer graphics field allowed the production of a large amount of synthetic data for many vision-related tasks. This allows to easily obtain large training sets but on the other side the domain shift between synthetic and real world data needs to be addressed. In this field, one of the most interesting applications is found in autonomous vehicles scenarios, where accurate understanding of the surrounding environment is crucial for a safe navigation in an urban context. At the same time, synthetic urban scenes provided with

automatically-generated annotations can be easily accessed from highly realistic computer graphic tools [23]. This allows to bypass the time-consuming and highly expensive manual labeling required for real-world training data, but in turn demands for domain adaptation techniques to safely bridge the statistical discrepancy between synthetic and real images. The synthetic to real adaptation for semantic segmentation of urban scenes will be further discussed in Section 4.

In Figure 3 we show three typical scenarios in which UDA for semantic segmentation could be highly valuable: namely, autonomous vehicles, industrial automation and domestic robots.

Figure 3. Autonomous cars, industry robots and home assistant robots are just some of the possible real world applications of Unsupervised Domain Adaptation (UDA) in semantic segmentation. (The images are modified version of pictures obtained with kind permission from Shutterstock, Inc., New York, NY, USA. The original versions have been created (from left to right) by Scharfsinn, Monopoly919 and PaO_Studio).

1.5. Outline

In this paper, we mainly focus on analyzing and discussing deep UDA methods in semantic segmentation. Recently, there has been a large number of studies related to this task. However, the motivating ideas behind these methods are different. To connect the existing work and hence to better understand the problem, we organize the current literature into some categories. We hope to provide a useful resource for the research of UDA in semantic segmentation.

The rest of the survey is organized as follows: in Section 2 a concise and precise formulation of UDA for semantic segmentation is given outlining the various stages at which the adaptation process may occur. Then, in Section 3 we give an overview of the state-of-the-art literature on the topic. We start from precursor techniques with weak supervision and then we propose a categorization based on the techniques employed to align the source and target distributions. In addition, we overview some new research directions considering more relaxed assumptions over target dataset properties, for example, dealing with the detection and classification of unseen semantic categories in target samples. In Section 4 we introduce a case study of synthetic to real unsupervised adaptation for semantic understanding of road scenes and we give an overview of the results of existing methods grouped by network architecture and evaluation scenario. In Section 5 we conclude our review with some final considerations on the different adaptation techniques and we outline some possible future directions.

2. Unsupervised Domain Adaptation for Semantic Segmentation

2.1. Problem Formulation

Image classification and image segmentation can both be reconducted to the problem of finding a function $h : \mathcal{X} \to \mathcal{Y}$ from the domain space $\mathcal{X}$ of input images to the label space $\mathcal{Y}$, that contains, respectively, the classification tags or the semantic maps. From a mathematical point of view, it is possible to suppose that all real-world labeled images $(x, y) \in \mathcal{X} \times \mathcal{Y}$ are drawn from an underlying, fixed and unknown probability distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$. The search of the function h should be limited to a predefined function space $\mathcal{H}$, called hypothesis class, chosen based on the prior knowledge on the problem. In a supervised setting, a dataset of i.i.d. samples from $\mathcal{D}$ is used by the learner to find the best mapping $h \in \mathcal{H}$ (i.e., the solution that minimizes a cost function over the training set). On the other hand, in DA, two different and related distributions over $\mathcal{X} \times \mathcal{Y}$, namely a source distribution $\mathcal{D}_S$

and a target distribution $\mathcal{D}_T$, are considered. A source domain training set $\mathcal{S}$ is sampled from $\mathcal{D}_S$ and a target domain training set $\mathcal{T}$ is sampled from $\mathcal{D}_T$ or from its marginal distribution over $\mathcal{X}$. The main purpose of DA is to use labeled i.i.d. samples from source domain $\mathcal{S}$ and labeled, or unlabeled, or a mixture of both, i.i.d. samples of the target domain $\mathcal{T}$ to find a hypothesis $h \in \mathcal{H}$ that performs well on the target domain $\mathcal{T}$. The DA task is supervised if labels in the target domain are available for all samples; it is semi-supervised if labels are available for just some samples; or it is unsupervised if the target samples are completely unlabeled (i.e., they are drawn from the marginal distribution of $\mathcal{D}_T$ over $\mathcal{X}$). Domain adaptation can be subdivided even further based on the categories (i.e., classes or labels) of the source (C_S) and target (C_T) domains, and on the categories considered in the learning process (C_L):

- **Closed Set DA**: all the possible categories appear in both the source and target domains ($C_S = C_T$);
- **Partial DA**: all the categories appear in the source domain, but just a subset appears in the target domain ($C_S \supset C_T$);
- **Open Set DA**: some categories appear in the source domain and all categories appear in the target domain ($C_S \subset C_T$);
- **Open-Partial DA**: some categories belong only to the source or to the target set and others belong to both sets ($C_S \neq C_T$ and $C_S \cap C_T \neq \emptyset$);
- **Boundless DA**: an Open Set DA where all the target domain categories are learned individually ($C_S \subset C_T$ and $C_L = C_S \cup C_T$).

It is important to remark that in Open Set DA, usually, the categories of the target set that do not belong also to the source domain are learned by the model as an *unknown* additional class, while in Boundless DA [24] they are learned individually. An overview of the aforementioned classification is given in Figure 4.

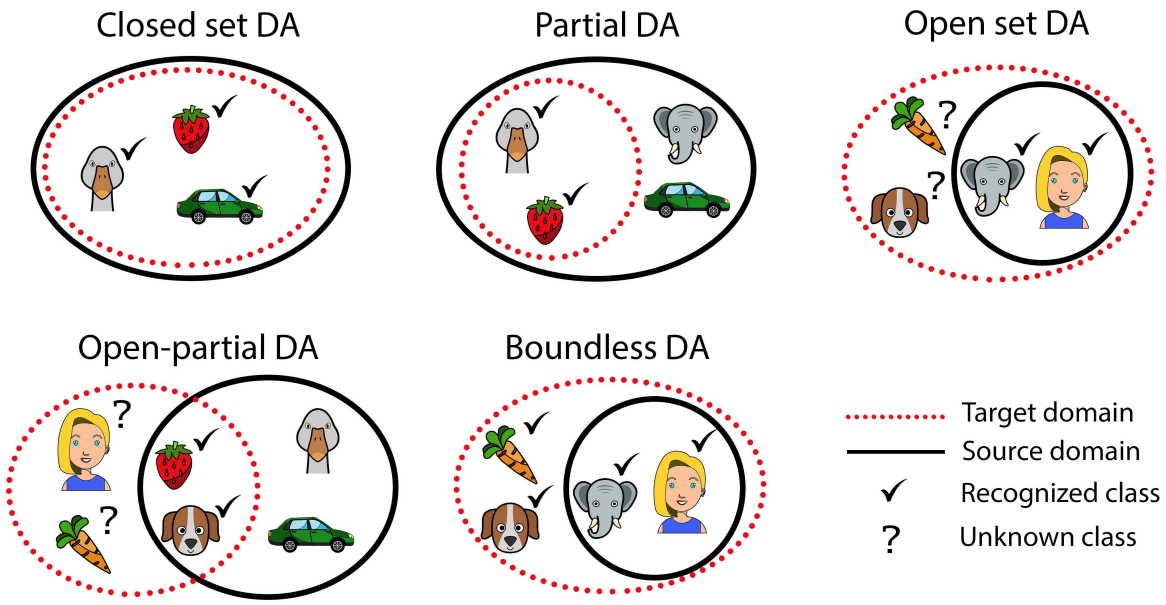

Figure 4. Possible intersections between source and target domains.

2.2. UDA in Semantic Segmentation: Adaptation Spaces

As previously discussed, there exists a domain shift phenomenon between source and target datasets, which prevents the network from yielding satisfactory results on the unsupervised target data. Therefore, the primary strategy to tackle the domain adaptation problem is to bridge the gap existing between source and target distributions. In doing so, the performance drop that affects prediction models should lessen, thus allowing effective prediction whenever the original form of

statistical discrepancy is successfully removed. In the following, a review of the different levels at which adaptation may be performed is presented, which will be also useful for the paper categorization in Section 3. A visual representation of the possible levels of adaptation is shown in Figure 5.

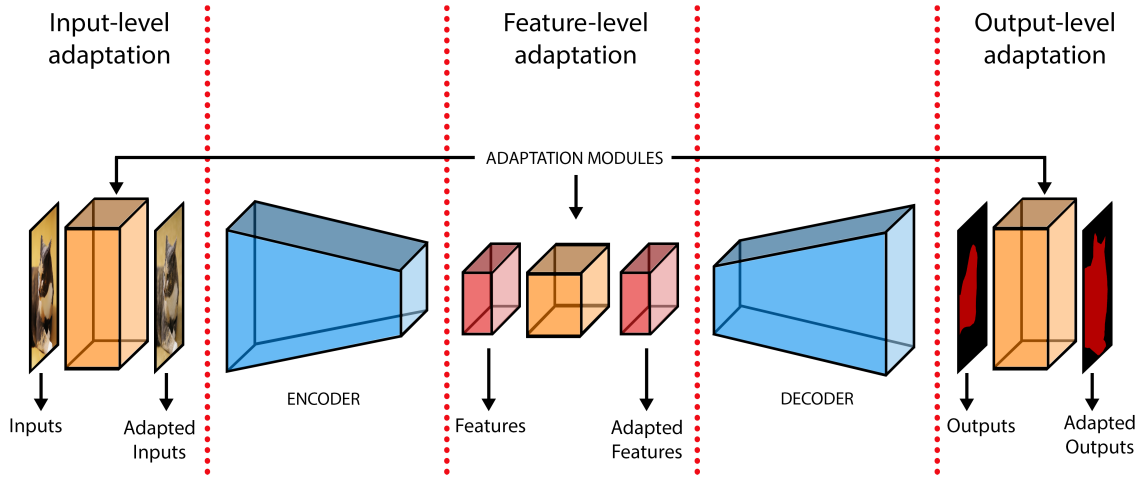

Figure 5. Domain shift adaptation could be performed at different spaces: input-level, feature-level and output-level.

Adaptation at the Input Level: one way to proceed is to address the statistical matching at the input level to achieve cross-domain uniformity of visual appearance of the input image samples. Even if source and target images carry strong high-level semantic similarity in scene content and layout, inter-domain low-level statistical discrepancy, although mostly lacking semantic significance, is likely to result in an undesirable reduction of the prediction efficacy on target samples. In light of these considerations, a rich line of works have been focusing on style transfer techniques to close the marginal distributions of source and target images from original image-level sets. The common approach is to discover a function that maps source images to a new space, where the projected samples should carry an enhanced perceptual resemblance to the target ones. Then, the image segmentation network can access samples from the domain invariant input space during training. Recently, translation on the other way around has been explored as well, by which target images are transferred to the source domain before being fed to the segmentation network.

This strategy, despite being in principle completely task independent (it is usually performed in a stage detached from the training of the task predictor), is missing sufficient discriminative power when it is employed in its simplest scheme, without any additional regularization constraints. Indeed, alignment of marginal distributions can be fully accomplished, and yet no semantic consistency may be preserved, with class-conditional distributions (not accessible at training time in the unsupervised target domain) still differing across domains. In other words, one might find many domain invariant representations, all lacking semantic discriminability to solve the segmentation task in the target domain. This for example could happen when objects of a certain class are mapped to different categories, which may be totally complying with the statistical alignment constraint, while, in fact, disregarding semantic content preservation (i.e., the preservation of semantic classes before and after the domain translation). To bypass these issues, multiple approaches have been devised to enforce semantic consistency of image translations, for example, by resorting to image reconstruction constraints, uniformity of segmentation predictions or ad-hoc engineered techniques to safely manipulate low-level statistics. Some techniques of this family are discussed in detail in Sections 3.2 and 3.3.

Adaptation at the Feature Level: a different approach is to seek for a distribution alignment of network latent embeddings. The core idea is to force the feature extractor to discover domain-invariant features, by adjusting the distribution of latent representations from source and target domains, both globally and class-wise. In this way, the network classifier should be able to learn to segment both source and target representations from the common latent space, by relying solely on the supervision from source data. Compared to the classification task, in which feature domain adaptation has been successfully applied, semantic segmentation entails a much more complex and high-dimensional feature space, which should encode both local and global visual cues. Thus, alignment at the feature level in its simplest fashion could be less effective in semantic segmentation, due to the structural and semantic complexity that feature embeddings possess, which is difficult to fully capture and handle (e.g., by an adversarial discriminator) [25]. In addition, even though adapted features should in principle retain semantic discriminability, they actually correspond to an intermediate representation in the segmentation process and there is no guarantee that the joint image-label distributions are aligned between domains, as unlabeled target images are drawn only from the marginal distribution. This can give rise to incorrect knowledge generalization to the unsupervised target representations. For such aforementioned reasons, feature adaptation has been adopted in semantic segmentation in combination with other complementary techniques or with specific arrangements to carefully overcome these major issues. Some techniques using explicit feature-level adaptation are discussed in Sections 3.2–3.4 and 3.6.

Adaptation at the Output Level: to avoid dealing with an excessively convoluted latent space, a different group of adaptation methods resort to the cross-domain distribution alignment over the segmentation output space. While retaining enough complexity and richness of semantic cues, prediction maps from the segmentation network output (or the per-class outputs of the very last layers) identify a low-dimensional space where adaptation can be performed quite effectively, for example recurring to adversarial strategies. Moreover, label statistics over segmentation maps can be easily inferred over unlabeled target data, introducing a form of self-constructed weak supervision to the segmentation task. Source priors from label distribution can be profitably imposed in the adaptation process as well, as they usually involve high-level structural properties unbounded to the specific domain. Examples of these techniques are discussed in Sections 3.5– 3.7.

Adaptation at Ad-Hoc Network Levels: in addition to the aforementioned techniques, other works resort to a distribution alignment over ad-hoc spaces upon network activations. Such methods aim at better capturing high-level patterns essential to solve the segmentation task, and ultimately achieve an improved match of source and target embeddings, thanks to gradients flowing back through the segmentation network at different levels. Hence, the adaptation is not only restricted to a particular network level, i.e., at the end of the feature extraction network, but it is achieved at intermediate levels as well.

3. Review of Unsupervised Domain Adaptation Strategies

This section reviews the most relevant approaches for Unsupervised Domain Adaptation in semantic segmentation. We start this section by presenting some weakly- and semi- supervised learning methods for semantic segmentation. Those are not purely UDA approaches since they require some minimal supervision with annotations on typically simpler tasks, but have represented the starting point in dealing with the domain adaptation problem. Then, we grouped UDA approaches into seven main categories, as shown by the visual overview in Figure 6. Domain adversarial discriminative approaches (Section 3.2) learn to produce data with a statistical distribution similar to that of training samples via adversarial learning schemes. Generative-based approaches (Section 3.3) typically use generative networks to translate data between domains in order to produce a target-like training set from source data, or alternatively to translate the source data into a representation closer to target domain characteristics that can then be fed to the network. Classifier discrepancy approaches in Section 3.4 resort to multiple dense classifiers on top of a single encoder to capture less adapted target

representations and, in turn, encourage an improved alignment of cross-domain features far from decision boundaries via an adversarial-like strategy. Self-training approaches in Section 3.5 propose to produce some form of pseudo-label (typically using some confidence estimation schemes to select the most reliable predictions) based on the current estimate to automatically guide the learning process (self-supervising it). Entropy minimization methods in Section 3.6 aim at minimizing the entropy of target output probability maps to mimic the over-confident behavior of source predictions, thus promoting well-clustered target feature representations. Curriculum learning approaches in Section 3.7 tackle one or more easy tasks first, in order to infer some necessary properties about the target domain (e.g., global label distributions) and then train the segmentation network such that the predictions in the target domain follow those inferred properties. Multi-tasking methods in Section 3.8 solve multiple tasks simultaneously to improve the extraction of invariant features representation. Finally, in Section 3.9 we conclude our digression with some considerations about recent interesting research directions to be further expanded in the future.

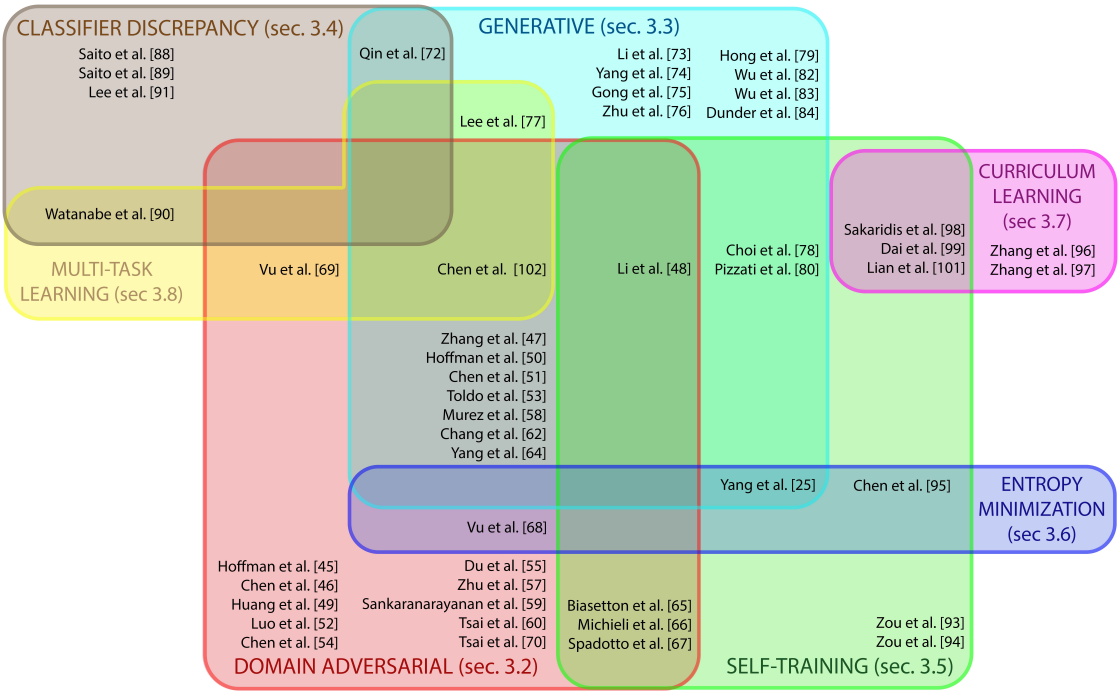

Figure 6. Venn diagram of the most popular UDA strategies for semantic segmentation. Each method falls in the set representing the adaptation techniques used. (Best viewed in color.)

3.1. Weakly- and Semi-Supervised Learning

As earlier semantic segmentation works started from image classification techniques, also the domain adaptation task was originally tackled in the classification field, in that the first DA approaches for semantic segmentation have been developed by adapting DA methods for classification. However, approaches directly targeting the semantic segmentation task started to appear soon, taking into account the specific properties of the spatial components (completely missing in the classification methods) and of the dense (pixel-wise) task. At the same time, unsupervised domain adaptation was historically preceded by techniques with weak or partial supervision, which are the focus of this section.

As we already mentioned, the training of a deep learning model for semantic segmentation requires a large amount of data with pixel-level semantic labels that are very expensive, difficult, frustrating and time-consuming to acquire. Such problem is not so relevant for other computer vision

tasks like image classification and object detection because image-level tags or bounding boxes (that in this context are called *weak* labels) are much simpler to obtain and large annotated collections are available. This is the motivation behind many works that propose to use just weakly labeled samples to train a model in the segmentation task (weakly-supervised learning) or to use a mixture of many weakly labeled samples and few samples with the more expensive pixel-level semantic map (semi-supervised learning).

A first approach to solve the problem is to cast the weakly supervised semantic segmentation as a Multiple Instance Learning problem, as shown in [26,27]. The Semantic Texton Forest (STF) traditional feature-based approach has been used as base framework and an algorithm to estimate unobserved pixel label probabilities from image label probabilities has been introduced. Then, the structure of the STF has been improved through a new algorithm that uses a geometric context estimation task as regularizer in a multi-task learning framework.

Another strategy, proposed in [28], is to implement an Expectation-Maximization (EM) method to train a deep network for the semantic segmentation task in a weakly- and semi-supervised setup. The algorithm alternates between estimating the pixel-level annotations (constrained on the weak annotations) and the optimization of the segmentation network itself. In [29], Constrained CNNs (CCNNs) have been introduced as a framework to incorporate weak supervision into the training. Linear constraints are added in the output space to describe the existence and expected distribution of labels from image level tags and a new loss function is introduced to optimize the set of constraints.

In [30], a simple to complex framework has been introduced for weakly-supervised semantic segmentation. In the paper, a distinction is made between simple and complex images: the former include a single object of just one category in the foreground and a clean background, the latter can have multiple objects of multiple categories with a cluttered background. First, salient object detection techniques are used to compute semantic maps from weak-annotated simple images and, then, starting from these, three different networks are trained, sequentially, in order to gradually enable the segmentation of complex images.

In [31], a semi-supervised approach is presented, in which the architecture is composed of three main structures: a classification network, a segmentation network and some bridging layers that interconnect the two networks. The proposed training is decoupled: first, the classification network is trained with weakly-annotated samples, then, the bridging layers and the segmentation network are jointly trained with the strong-annotated samples. The input image is first fed to the classification network, then the bridging layers extract from an intermediate layer of the classification network a class-specific activation map that is used as input for the segmentation network. In this way, it has been possible to reduce the number of parameters of the segmentation network and to make a training with just few semantic-annotated samples possible. In fact, relevant labels and spatial information are captured from the classification network and refined by the bridging layers and the task of the segmentation network is widely simplified.

In [32], an iterative procedure has been proposed to train a segmentation network just with bounding-boxes annotated samples. First, region proposal methods are used to generate many candidate segmentation masks (that are fixed throughout the training) for each image. An overlapping objective function is defined to pick the candidate mask that overlaps the ground truth bounding box as much as possible with the correct label. At every iterative step, one candidate mask is selected for each bounding box and then the resulting semantic labels are used to train the segmentation network. The outputs of the segmentation network are then used to improve the choice of the candidate labels for the next step through a feedback channel. After every iteration, the selected candidate labels and the segmentation network outputs both improve together.

Generative adversarial networks have proven to be effective in this field starting from [33], where the discriminator network is modified to accomplish the task of semantic segmentation. The discriminator assigns to every pixel of the input image either a label of one of the semantic classes or a fake label. The discriminator is trained with fake (generated) data, unlabeled data for regularization

purposes, and with labeled data with pixel-level semantic maps. Another proposed solution is to employ conditional Generative Adversarial Networks (GANs) and incorporate weak image-level annotation both at the generator and at the discriminator inputs in a weakly-supervised setup.

Many approaches of self-supervised learning have been proposed starting from [34]. The common rationale is the exploitation of inferred pixel-level activations as pseudo ground-truth in order to obtain more accurate pixel-level segmentation maps. In [35], an image classification network with classification activation maps has been used. The authors highlight how the discriminative regions, using that method, are small and sparse and they propose to use them as seed cues. Then, the regions are expanded to neighboring pixels with similar features (for example color, texture or deep features) with a classical Seeded Region Growing (SRG) algorithm to obtain accurate pixel-level labels that are used to train a segmentation network. The output of the segmentation network is used by the SRG algorithm to compute the similarity between the seed and the adjacent pixels. So, at every iteration, the segmentation network and the dynamic labels computed with SRG improve together. A similar approach that introduces a new adversarial erasing method for localizing and expanding object regions progressively is presented in [36]. Other self-learning based techniques are presented in [37–39].

A more general framework to transfer knowledge across tasks and domains is presented in [40]. Assuming to have two tasks and two domains, the proposed method works in four steps: (1) a single task network is trained on samples of both domains to solve the first task, in order to find a common feature representation for the domains; (2) a second network is trained to solve the second task just on the first domain; (3) a third network is trained on the first domain to map deep features suitable for the first task into features to be used for the second task; (4) finally, the third network is used to solve the second task on the second domain. This framework enables to adapt from a synthetic domain to a real one for the image segmentation task using depth maps of both domains. The depth maps can be considered weak annotations with respect to semantic maps because their acquisition is easier thanks to depth cameras and 3D scanners.

3.2. Domain Adversarial Discriminative

Adversarial Learning: Adversarial learning has been introduced in the form of Generative Adversarial Networks (GANs) [41] to address a generative objective (i.e., generating *fake* images similar to real world ones). Solving the generative task can be thought as seeking the evaluation of the unknown probability distribution from which the training data has been generated. In the generative context, the introduction of adversarial learning has been ground-breaking, as explicit modeling of the underlying target distribution is not required and, more importantly, no specific objective is needed to train the model. In the adversarial scheme, a generator has to learn to produce data with the same statistical distribution of training samples. To do so, it is paired with a discriminator, which has the goal of understanding whether input data comes from the original set or, instead, it has been generated. At the same time, the generator is optimized to fool the discriminator by producing samples that resemble the original ones. In the end, the statistics of generated data should match that of the training set. The GAN model is capable of learning a structured loss in the form of a learnable discriminator, which guides the generative network in its optimization procedure. For this reason, the objective function can be thought as automatically adapting to the specific context, removing, in fact, the necessity to manually design complex losses. Therefore, the adversarial learning scheme introduced in [41] can be extended under careful adjustments to address multiple tasks that would normally require different types of application-specific objectives.

Feature Adversarial Adaptation: Aiming at exploiting the statistical matching that can be achieved by the GAN model, adversarial learning has been successfully extended to the domain adaptation task [42–44]. In particular, the real-fake discriminative network in the original adversarial framework is revisited, turning into a source-target domain classifier. Thus, while the segmentation network is trained with source supervision to achieve discriminability over the semantic segmentation task, the supervisory signal provided by the domain discriminator should guide the predictor to reach

domain invariance and reduce the otherwise intrinsic bias towards the source domain. In other words, a measure of domain discrepancy is simultaneously learned and minimized within the adversarial competition.

Although the adversarial adaptation strategy has been originally introduced for the image classification task [42,43], it has been later extended to image semantic segmentation. Hoffman et al. [45] have been the first to address domain adaptation in semantic segmentation, and they resort to an adversarial approach. In particular, they devise a global domain adversarial alignment, based on a domain discriminator taking as input the feature representations from intermediate activations of the fully convolutional segmentation network. In addition, they propose a category specific distribution alignment, which is accomplished by enforcing image-level label distribution constraints on target predictions inferred from source annotations, under the assumption that high-level scene layout is in general shared among source and target images.

Following a similar approach to [45], many works have further resorted to adversarial alignment of network latent embeddings [46–55]. As previously discussed, the domain discriminator is able to infer a structured loss to capture global distribution mismatch of cross-domain image representations. However, global alignment of marginal distributions does not necessarily result in class-wise correct semantic knowledge transfer from source to target representations. Thus, adversarial learning is commonly employed in more complex frameworks working also on the internal feature representations of the network, comprising multiple complementary modules to achieve a more effective adaptation. For example, Chen et al. [46] use an additional target guided distillation loss by matching network activations from target inputs during the training phase with those from a pre-trained version on the ImageNet dataset [56]. In this way, they argue that overfitting to source data is decreased. Moreover, the feature adversarial adaptation is enforced independently over different spatial regions of the input image, thus exploiting the underlying spatial structure of input scenes. Zhang et al. [47], instead, boost the feature-level adaptation performance by providing the domain discriminator with an Atrous Spatial Pyramid Pooling (ASPP) module [5] to capture multi-scale representations. More recently, Luo et al. [52] propose a significance-aware information bottleneck (SIB) to filter out task-independent information encoded inside feature representations, so that, when enforcing adversarial adaptation, only domain invariant discriminative cues are preserved. They also introduce a significance-aware module to help the prediction of less frequent classes, which may be penalized by the information bottleneck in its original form.

Another group of researches [48,50,51,53] combines a generative approach (which will be extensively discussed in the following Section 3.3), with the adversarial feature alignment. In particular, source and target marginal distributions are matched in the input image space by a source-to-target image-to-image translation function, and then cross-domain latent representations are further brought closer by matching source original and target-like source embeddings in a domain adversarial fashion.

To accomplish category-wise adaptation, some works [54,55] revisit the original approach of Hoffman et al. [45] by assisting the global distribution alignment with class-wise adversarial learning. Chen et al. [54] propose to exploit multiple feature discriminators (one for each class), so that negative transfer among different classes in the domain bridging process should be effectively avoided. In addition, due to lack of ground-truth maps, they use grid-level soft pseudo labels from network predictions to compute the target adversarial loss. Recently, Du et al. [55] proposed a similar class-wise adversarial technique, which is improved by imposing independence during the optimization of the multiple discriminators. They argue that soft labels lead to incorrect adaptation on class boundaries, where different class discriminators may provide their guidance simultaneously. Finally, they devise an additional module to adaptively re-weight the contribution of each class component in the adversarial loss, in order to avoid the inherent dominance of classes with higher prediction probability, which turns out to be more easily well-adapted across the domains.

Different from the aforementioned techniques, other works [57–59] seek for domain alignment inside the feature space by applying a reconstruction constraint to ensure that latent embeddings

possess enough information to recover the input images from which they have been extracted. To this end, adversarial learning is applied on the reconstruction image-level space. To achieve cross-domain feature distribution alignment, the feature extractor is trained to yield latent representations that can be projected back to both source and target image spaces indistinctly. In these frameworks the backbone encoder of the segmentation network plays a min-max game against the domain discriminator. The encoder, indeed, tries to fool the discriminator on the actual originating feature's domain, by looking at the corresponding reconstructed images projected back into the image space. In other words, the objective is to learn source (target) features that can successfully generate target-like (source-like) images to promote domain invariance of those representations.

Output Adversarial Adaptation: To avoid the complexity of high-dimensional feature space adaptation, a different line of works [60–67] resort to adversarial adaptation on the low-dimensional output space spanned by the segmentation network, which is still expected to encode enough semantic information to allow effective adaptation. A domain discriminator is provided with prediction maps from source and target inputs and it is optimized to discern the domain they originate from. Conversely, the segmentation network has to fool it by aligning the distribution of predicted dense labels across domains. Tsai et al. [60] are the first to propose this type of adaptation: in order to improve the signal flow from the adversarial competition through the segmentation network, they deploy multiple dense classification modules at different depths upon which as many output-level discriminators are applied. Following the technique proposed in [60], other works adopt the output space adversarial adaptation in combination with additional modules. For example, Chen et al. [61] combine semantic segmentation and depth estimation to boost the adaptation performance. In particular, they provide the domain discriminator with segmentation and depth prediction maps jointly, in order to fully exploit the strong correlation between the two visual tasks. Moreover, Luo et al. [63] enhances the adversarial scheme by a co-training strategy that highlights regions of the input image with high prediction confidence. In this way, the adversarial loss can be effectively tuned by balancing the contribution of each spatial unit, so that more focus is directed towards less adapted areas.

Other works [65–67] revisit the adversarial output-level approach. In particular, they utilize a discriminator network that has to distinguish between source ground-truth maps and generated semantic predictions from either source and target data. In doing so, the cross-domain statistical alignment is not directly performed, but forcing the segmentation network output to be distributed as ground-truth labels for both source and target inputs leads to an indirect yet effective alignment between the two domains.

Recently, new approaches [68–70] have been proposed based on the extraction of meaningful patterns from the segmentation output space to be exploited in the adaptation process. This is done to explicitly guide the domain discriminator towards a more functional and significant insight of source and target representations, and thus to ultimately achieve a better alignment.

On this regard, Vu et al. [68,69] devise an entropy-minimization strategy (which will be described more in detail in Section 3.6) to promote more confident target predictions. They propose an indirect approach relying on the adversarial alignment of the statistics of self-information maps computed on top of source and target predictions. In particular, a domain discriminator has to detect whether a weighted self-information map comes from a source or a target prediction, whereas the segmentation network, trying to deceive the discriminator, is forced to produce low-entropy target maps as to mimic source confident ones. This process effectively pushes decision boundaries away from high-density regions in the representation space.

With a different approach, Tsai et al. [70] construct a clustered space over the output prediction space by adding a patch clustering module that discovers patch-wise modes on segmentation maps. First, the module is trained, while supervised, on source data by leveraging the available annotations, then it is exploited to achieve a patch-wise distribution alignment by enforcing adversarial cross-domain adaptation between its clustered source and target representations. The idea behind this approach is to capture high-level structured patterns, which are essential to solving the semantic

segmentation task, to be provided to the domain discriminator for an improved domain statistical alignment. Thus, the achieved domain uniformity on a patch-level should ensure, in principle, that the segmentation task can be effectively solved also in the target domain.

3.3. Generative-Based Approaches

Unsupervised image-to-image translation is a class of generative techniques where the objective is to learn a function that maps images across domains, relying solely on the supervision provided by unpaired training data sampled from the considered domains. The idea is to extract characteristics peculiar to a specific set of images and transfer those properties to a different data collection. In a more formal definition, the image-to-image translation task aims at discovering a joint distribution of images from different domains. Notice that, since the problem is, in fact, ill-posed, as an infinite set of joint distributions can be inferred from the marginal ones, appropriate constraints must be applied to obtain acceptable solutions.

Image-to-image translation can be effectively exploited in domain adaptation: discovering the conditional distribution of the target set with respect to the source one, should allow, in principle, to bridge the statistical gap between source and target pixel-level statistics, thus removing the original covariate shift responsible for the classifier performance drop. The goal, in fact, is to transfer visual attributes from the target domain to the source one, while preserving source semantic information. Following this idea, many works have proposed an input-level adaptation strategy based on a generative module that translates images between source and target domains. Despite the wide range of different approaches, all these works share the same idea of achieving a form of domain invariance in terms of visual appearance, by mitigating the cross-domain discrepancy in image layout and structure. This allows to learn a segmentation network on translated source domain data (that should have a target-like statistical distribution) allowing to make use of source annotations.

A considerable amount of research [48,50,51,53,58,71–75] has been resorting to the successful CycleGAN [76] unsupervised image-to-image translation framework to accomplish input-level domain adaptation. The framework proposed by Zhu et al. [76] is built on top of a pair of generative adversarial models, concurrently performing conditional image translation between a couple of domain sets, in both the source-to-target and target-to-source directions. The two adversarial modules are further tied by a cycle-consistency constraint, which encourages the cross-domain projections to be one the inverse of the other. This reconstruction requirement is essential to preserve structural geometrical properties of the input scene, but provides no guarantees about the semantic consistency of translations. In fact, while retaining geometrical coherence, the mapping functions could completely disrupt the semantic classification of input data.

Taking this into account, a number of works [48,50,51,53,71] address semantic consistency by taking advantage of the semantic discriminative capability of the segmentation network. In particular, cross-domain image translations are forced to preserve semantic content as perceived by the semantic predictor, which represents a measure of semantic discrepancy between an original image and its translated counterpart, which is minimized in the optimization of the translation network. Still, with the prediction maps being intrinsically flawed, especially in the target domain where annotations are missing, the inaccurate semantic information provided to the generative module could hurt the learning of the image projections. Thus, some works propose to simultaneously optimize the generative and discriminative framework components in a single stage [53], or even split the segmentation network into separate source and target predictors [51]. Li et al. [48] further extend the CycleGAN-based adaptation strategy formulating a bidirectional learning framework. The image-to-image translation and segmentation modules are alternately trained, in an optimization scheme by which each module is provided with positive feedback from the other. The segmentation network benefits from the target-like translated source images with original supervision, while the generative network is aided by the predictor in retaining semantic consistency. This closed-loop structure effectively allows for

a progressive adaptation, with both image-to-image translations quality and semantic prediction accuracy gradually enhanced.

Other works [73,74] resort to different approaches to provide semantic-awareness to the CycleGAN-based adaptation. Li et al. [73] propose to assist the cycle-consistent image-to-image translation framework by a soft gradient-sensitive loss to preserve semantic content in the cross-domain projection focusing on semantic boundaries. The idea behind this approach is that, no matter how low-level visual attributes change between domains, the edges defining semantic uniform regions should be easily detectable, regardless of the distribution the image is drawn from. Thus, a gradient-based edge detector should discover consistent edge maps between original images and their transformed versions. In addition, following the intuition that semantically different regions of an image should face a different adaptation, they devise a semantic-aware discriminator structure. In doing so, the discriminator can semantically-wise evaluate resemblance between original and translated samples.

Very recently, Yang et al. [74] introduced a phase consistency constraint to the CycleGAN pixel-level adaptation module, observing that the semantic content of an image is mostly encoded in the phase of its Fourier transform, whereas alterations of the amplitude to the representation in frequency does not change its composition.

With a different adaptation perspective, Gong et al. [75] adapted the CycleGAN model to generate a continuous flow of domains ranging from source to target ones, by conditioning the generative networks with a continuous variable representing the domain. The reason behind the retrieval of intermediate domains spanning between the two original ones is to ease the adaptation task, by progressively characterizing the domain shift affecting the input data distributions. Moreover, they suggest that resorting to target-like training data from diverse target-like domain distributions improves the generalization capability of the segmentation network.

To reduce the computational burden of the bi-directional structure of CycleGAN (which entails a total of at least four neural networks to be added to the semantic predictor) other works [61,77–79] discard the backward source-to-target projection branch, seeking for a more light-weight input-level adaptation module, still based on generative adversarial framework. The translation consistency is granted, for example, by the correlation to a related task (e.g., depth estimation) [61,77], which is jointly addressed with the semantic segmentation. Choi et al. [78], instead, improve the generator of the original GAN framework with feature normalization modules at multiple depths to provide style information to source representations, whereas source content is preserved. Furthermore, a semantic consistency loss from a pre-trained segmentation network promotes coherence of image translations, providing, in fact, a regularizing effect in absence of the cycle-consistency one. Hong et al. [79] use a conditional generative function to model the residual representation between source and target feature maps, which is optimized in an adversarial framework. In doing so, they avoid any reliance on a shared domain-invariant latent space assumption, which may not be satisfied due to the highly structured nature of semantic segmentation. The generator takes as input low-level source feature maps, together with a noise sample, and is encouraged to produce high-level feature maps with target-like distribution by a discriminator, which expresses a measure of statistical distance between original and reproduced target representations. Both source original and domain-transformed representations are provided to a dense classifier to compute the cross-entropy loss.

In order to lessen the bias towards the source domain, Yang et al. [64] resort to the target-to-source image-to-image translation, in place of the more common source-to-target one, generally employed to generate a form or target supervision from source translated data. The source-like target images are then employed in the supervised training of the predictor thanks to pseudo-labeling. In addition, training the segmentation network directly in the source domain allows to fully exploit the original source annotations, avoiding the risk of semantic alterations, which may happen in the source-to-target pixel-level adaptation scenario. Moreover, to align feature representations between domains, they introduce a label-driven reconstruction network. However, differently from the

feature-based reconstruction techniques [57–59] (Section 3.2), the generative recreation of input images is performed starting from semantic maps from the segmentation output. In doing so, they seek to guide a category-wise alignment of the segmentation network embeddings, since reconstructions that deviate from their target are penalized, thus providing semantic consistency to network predictions.

A different category of adaptation strategies explores style-transfer techniques to achieve image-level appearance invariance between source and target domains. These approaches are based on the principle that every image can be disentangled into two separate representations, namely content and style. As the style encodes low-level domain-specific texture information, the content expresses domain-invariant high-level structural properties. Thus, being able to combine style properties from target data with semantically preserving source content should effectively allow for the construction of target-distributed training data, still retaining original source annotations. Some techniques [62,80] involve content and style decomposition in the latent space. Translating a source image, then, means extracting its feature content representation and recombining it with a random target style representation. In a recent work [80], the authors perform multi-modal source-to-target image translation based on the MUNIT architecture [81]. The original datasets are augmented with additional web-crawled data, in order to reduce the gap in terms of task-unrelated data properties between sets, while at the same time highlighting the relevant task-related visual features to be matched. Furthermore, the style transfer method allows for multi-modal translation, i.e., multiple target styles can be transferred to a single source image, thus increasing training data diversity and, in turn, enforcing the adaptation robustness.

Other works [25,47,82–84] completely avoid the computational complexity of generating high resolution images with GANs by exploiting different types of style transfer techniques. Zhang et al. [47] adopt traditional techniques of neural style transfer [85,86] to separate style (low-level feature) from image content (high-level features). In particular, multi-level response maps of a pre-trained CNN are exploited for image synthesis, where image style is expressed by the correlation between feature maps in the form of Gram matrices. Alternative approaches [78,83] opt for the re-normalization of source feature maps, so that their first and second order statistics match those of the target ones, by means of the AdaIN module [87]. Differently, Dundar et al. [84] make use of a photo-realistic style transfer algorithm for an iterative optimization by which both the segmentation network and the translation algorithm performances are constantly improved. Finally, Yang et al. [25] remove domain-dependent visual attributes from source images by replacing the low-level frequency spectrum with that of target images, without affecting high-level semantic interpretability. They argue that this simple approach, despite not requiring any additional learnable module, results in a remarkably robust adaptation performance when embedded in a multi-band framework that averages predictions with different degrees of spectral alteration.

3.4. Classifier Discrepancy

As discussed in Section 3.2, feature-level adversarial domain adaptation in its original form entails the competition between the task feature extractor and a domain critic (the discriminator), whose supervisory action in principle should guide towards the cross-domain alignment of feature representations. Task-discrimination instead is granted by a source supervised task objective (i.e., the standard cross-entropy loss for semantic segmentation).

As highlighted by [88,89], the major drawback of this primary form of adversarial adaptation lies in the lack of semantic awareness from the domain discriminator network. Even when the critic manages to grasp a clear expression of marginal distributions, thus effectively leading to a global statistical alignment, category-level joint-distributions necessarily remain unknown to the domain discriminator, as it is not provided with semantic labels when discriminating feature representations. A side effect of this semantic-unaware adaptation is that features can be placed close to class boundaries, increasing the chances of incorrect classification. Furthermore, target representations may be incorrectly

transferred to a semantic category different from the actual one in the domain invariant adapted space (negative transfer), as decision boundaries are ignored in the adaptation process.

To overcome these issues, Saito et al. [88] propose an Adversarial Dropout Regularization (ADR) approach for UDA to provide cross-domain feature alignment away from decision boundaries. To do so, they completely revisit the original domain adversarial scheme, by providing the task-specific dense classifier (i.e., the encoder) with a discriminative role. In particular, by means of dropout, the classifier is perturbed in order to get two distinct predictions over the same encoder output. Since the prediction variability is subject to an inverse relationship with the proximity to decision boundaries, the feature extractor is forced to produce representations far from those boundaries by minimizing the discrepancy of the two output probability maps. At the same time, the classifier has to maximize its output variation, in order to boost its capability to detect less-adapted features. In this redesigned adversarial scheme, the dense classifier is trained to be sensitive to semantic variations of target features, as to capture all the information stored in its neurons, which in turn are encouraged to be as diverse as possible from each other by the adversarial dropout maximization. On the other hand, the encoder is focused on providing categorical certainty to extracted target features, since removing task-unrelated cues weakens the possibility to achieve dissimilar predictions from the same latent representations.

Following the same principle of Adversarial Dropout Regularization, other approaches resort to adaptation techniques based on classifier discrepancy to achieve a semantically consistent alignment [63,72,89–91]. Saito et al. [89] improve the framework in [88] by modifying the way of accessing multiple predictions over the same latent space. In place of dropout on classifier's weights, they introduce a couple of separate decoders, which are simultaneously trained with source supervision, while being forced to produce dissimilar predictions by the maximization of a discrepancy loss. The objective is to avoid the noise sensitivity acquired by the single classifier in ADR, which is essential for the individual decoder to capture the proximity to the support of target samples, but requires an additional training stage to correctly learn the segmentation model as a whole.

The co-training strategy of exploiting a couple of distinct classifiers to infer the degree of target adaptation is further merged to the more traditional generator-discriminator adversarial framework by Luo et al. [63]. They use the discrepancy map from the two classifiers' output to weight the adversarial objective. Thereby semantic inconsistent regions highlighted by strong prediction variability get a major focus in the objective, as they should suffer from a more prominent domain shift. Additionally, Lee et al. [91] re-propose a form of adversarial dropout to get divergent predictions from a single classifier. However, they drop the adversarial scheme for a non-stochastic virtual dropout mechanism, to discover minimum distance adversarial dropout masks that maximize prediction discrepancy. In the end, they resort to a single unified objective, for a combined optimization of the encoder and decoder to align features between domains, while progressively pushing dense regions and decision boundaries far away from each other.

3.5. Self-Training

The self-training strategy entails using highly confident network predictions inferred on unlabeled data to generate pseudo-labels, to be used, in turn, to reinforce the training of the predictor with the self-taught supervision. This approach has been commonly employed in semi-supervised learning (SSL) [92] to exploit additional unlabeled data in order to improve the prediction accuracy. Recently, self-training techniques have been extended to address unsupervised domain adaptation, since UDA can be considered as a variant of the SSL task, even though the additional complexity of UDA from the statistical shift of the unlabeled target data must be further taken into account. Indeed, concurrently learning from source annotations and target pseudo-labels implicitly promotes feature-level cross-domain alignment, while still retaining the task specificity. On the contrary, lacking a unified loss, other adaptation approaches, as the most successful adversarial ones, have to take care of the task-relatedness with additional training objectives. The critical point is that this strategy is

self-referential, so careful arrangements must be adopted to avoid catastrophic error propagation. Self-training, in fact, naturally promotes more confident predictions, as the network probability output is encouraged to reach a peaked distribution (at the limit a Dirac distribution) close to the one-hot pseudo-labels. Since no form of external supervision is available on unlabeled target data, the network could yield over-confident predictions by wrongly classifying uncertain pixels. In turn, the iterative self-teaching strategy enforces prediction mistakes, through a propagation mechanism that makes the output progressively deviating from the correct solution. For this reason, the majority of self-training based adaptation approaches rely on various forms of pseudo-label filtering, to allow self-learning only from top confident target predictions, which are implicitly assumed to have a higher chance of being correct.

A first class of adaptation solutions based on self-training [48,93,94] employs offline techniques for pseudo-label computation: at every update step a confidence threshold is computed by looking at the entire training set. Target segmentation maps are then directly filtered according to some confidence-based thresholding policy and used in combination with original source annotated data for the supervised learning of the segmentation network.

In this regard, Zou et al. [93] propose one of the first UDA techniques based on self-training. They devise an iterative self-training optimization scheme, which alternates steps of segmentation network training on both source original and target artificial supervision and target pseudo-label estimation. In particular, the target pseudo-labels are treated as discrete latent variables to be computed through the minimization of a unified training objective. In addition, motivated by the fact that class-unaware pseudo-labels confidence filtering is intrinsically biased towards the easy (i.e., more confident) classes, they devise a class-balancing strategy by setting category-wise confidence thresholds. This should promote inter-class balance, as the same amount of top confident pixels are considered for each class, thus resulting in class-wise uniform contributions to the learning process. Finally, since source and target domains are supposed to share high-level scene layout, they also utilize spatial priors from source label statistics, which are inferred for each semantic category and incorporated in the training objective. More recently, Zou et al. [94] revisited their previous work in [93] by extending the pseudo-label space from one-hot maps to a continuous space defined by a probability simplex. In this way, by avoiding clear-cut overconfident self-supervision in the whole input image, the effect of the inherent misleading incorrect pixel predictions should be effectively reduced. A continuous pseudo-label space further allows them to introduce a confidence regularizing term in the training objective targeting both pseudo-label (treated as latent variables) and network weights, with the purpose of achieving output smoothness in place of sparse segmentation maps.

In order to avoid slow offline dataset-wise processing, Pizzati et al. [80] introduce self-training with weighted pseudo-labels. A learnable confidence threshold is employed for both pseudo-label refinement and weighting, thus making pseudo-labels belong to a continuous space, while concurrently balancing the impact of uncertain pixels. Target weighted self-generated labels are computed over a single batch, but still retain a global view, since the confidence threshold is learned throughout the entire training phase.

A different group of researches [65–67] construct a self-training strategy on top of an adversarial discriminative adaptation module applied over the segmentation network output. In particular, on the belief that the fully convolutional discriminator can be regarded as performing a measure of reliability of network estimations, they exploit the discriminator output to identify reliable target predictions, which are then preserved in the pseudo-label filtering operation. Michieli et al. [66] further improve the pseudo-label selection mechanisms by a region growing strategy. Moreover, Spadotto et al. [67] propose to adopt a class-wise adaptive thresholding approach. They select the same fraction of highly confident target pixels for each semantic class, by looking at the batch-wise distribution of the discriminator probability output. In doing so, they provide the adaptation framework with both inter-class confidence flexibility and time adaptability over the training phase.

Another line of works [25,71,78,95] utilize various forms of prediction ensembling to yield more reliable predictions over target data, on top of which pseudo-labeling is performed. Chen et al. [95] enhance the adaptation of low-level features by introducing an additional ASPP dense classification module. Hence, self-produced guidance in the form of pseudo-labels from the combined knowledge of low and high level target predictions is exploited as an additional training objective. Yang et al. [25] train multiple instances of the segmentation network with multi-band spectrum adaptation to obtain distinct semantic predictors. Then, target pseudo-labels are generated from the mean prediction of the different segmenter instances, resulting in a more robust adaptation when dealing with multiple rounds of self-training.

Rather than operating directly on the predictor output, other self-training approaches [71,78] resort to an additional network to produce self-guidance over the unlabeled samples. Choi et al. [78] propose a self-ensembling adaptation technique, by which a teacher network derived from student network's weights average yields predictions the student network is compelled to follow. In other words, an auxiliary predictor (the teacher network) is providing a sort of pseudo-labels, which are then used to transfer reliable knowledge to the actual predictor (the student network) by supervised training on target data. With regularization purposes, Gaussian noise is additionally injected on input target images and dropout weight perturbation is applied to the segmentation network to improve adaptation robustness, as student-teacher prediction consistency is enforced even under different random disturbance. Recently, the student-teacher self-ensembling adaptation approach is extended by Zhou et al. [71], with the introduction of an uncertainty module that filters out unreliable teacher predictions by looking at self-information maps.

3.6. Entropy Minimization

As already pointed out, semi-supervised learning and unsupervised domain adaptation are closely related tasks: indeed, once source and target distributions are matched, the UDA task merely scales down to learning from an unlabeled subset of the training data. Therefore, it is natural that SSL approaches may inspire domain adaptation strategies, as discussed for self-training (Section 3.5). Among the successful techniques used to address semi-supervised learning, entropy minimization has been recently introduced to UDA [68]. The principle behind minimizing target entropy to perform domain adaptation follows the observation that source predictions are likely to show more confidence, which in turn translates into high entropy probability outputs. On the contrary, the segmentation network is likely to display a more uncertain behavior on target-distributed samples, as target prediction entropy maps happen to be overall quite unstable, typically being the noise pattern not confined just to the semantic boundaries. Thus, forcing the segmentation network to mimic the over-confident source behavior when applied to the target domain too, should effectively reduce the accuracy gap between domains. In other words, entropy minimization aims at penalizing classification boundaries in the latent space crossing high density regions, while jointly encouraging well-clustered target representations properly sorted out by decision boundaries.

In its simplest fashion [68] entropy minimization is performed at a pixel-level, so that each spatial unit of the prediction map brings an independent contribution to the final objective. However, the basic approach suffers from some intrinsic limitations, demanding further arrangements to boost the adaptation performance [25,68,95]. To leverage structural information of semantic maps, Vu et al. [68] propose a global adversarial optimization to enforce distribution alignment over source and target entropy maps. In doing so, they rely on a domain discriminator to capture global patterns differentiating samples from separate domains, thus achieving a more semantically meaningful cross-domain match of entropy behavior. Class-wise priors on label distributions inferred from source annotations are further enforced on target predictions to avoid class imbalance towards easy classes.

In a following publication, Chen et al. [95] observe that the entropy minimization objective can be seriously hindered by the gradient predominance of more confident predictions. Indeed, moving from high to low uncertainty areas, the gradient rapidly increases, and its value tends to infinity

as the output probability distribution tends to the delta function. This probability imbalance in general prevents the segmentation network from learning over areas with little accuracy, in which the gradients result much lower than those of easy-to-transfer image regions. To address this issue, they devise a maximum squares loss, which produces a gradient signal that grows linearly with the input probability. They also face class unbalance by introducing a category-wise weighting factor based on target distribution from prediction maps in place of source annotations, as they argue that source class statistics may significantly deviate from target ones.

Very recently, Yang et al. [25] added an entropy minimization technique as an additional module to their adaptation scheme. The intent is to seek a regularization effect over the training on unlabeled target data, accomplished by pushing the decision boundaries away from high-density regions in the target latent space, with basically no overhead to the actual framework. The strength of the approach is enhanced by the combined application of other adaptation modules to achieve domain alignment. This, in fact, shifts the UDA task towards SSL, thus making entropy minimization more effective. Moreover, to avoid excessive emphasis on low entropy predictions, they adopt a penalty function that increases the focus on less-adapted high entropy regions of target images.

3.7. Curriculum Learning

Another research area regards curriculum learning approaches, where some easy tasks are solved first, inferring some important and useful properties related to the target domain. Then, this information is used to support the training of a network dealing with a more challenging task, like image segmentation. This family of approaches shares many similarities in spirit with self-training. The main difference between the two approaches lies in the content of the pseudo-labels. While in the self-training approaches the pseudo-label is an estimate of the desired annotation on the target set and it is used as such during training, in curriculum approaches the pseudo-label is represented by some inferred statistical properties of the target domain (different from the labels for the task) and the network is trained to reproduce such inferred properties in the target predictions.

The first work of this family is [96] and its extension [97], where a couple of easy tasks that are less sensitive to domain discrepancy are solved: namely, the label distribution over global images and the label distribution over local landmark superpixels. The former property is evaluated in the source domain, as the number of pixels in the labels associated to each category, normalized by the total number of pixels. On the other hand, target labels are not available in unsupervised domain adaptation and consequently a machine learning model should be trained on the source domain to estimate them. In the papers, it is argued that this task can be solved more easily than image segmentation and that the results can be used to guide the adaptation of the segmentation task. To estimate the first property on the target domain, a logistic regression model is employed. While the first property is useful to guarantee that the ratio among different categories matches the ones of the target domain, samples with semantic maps not following the estimated label distribution on the target domain are still penalized. To solve this problem, a second clue is introduced. Images are divided into superpixels and an SVM classifier is used to select the most representative anchor superpixels and the label distribution is estimated over them. The final objective is a mixture of the pixel-wise cross-entropy of the source samples and the cross-entropy on the two properties on the target domain discussed before. In [98,99], a technique to adapt the domain of a segmentation model from clear weather to dense fog images is introduced. A novel method, called Curriculum Model Adaptation (CMAda), is proposed to gradually adapt the model to segment images with an incrementally growing amount of fog. A new method to add synthetic fog to images and a new fog density estimator are introduced. It is important to remark that the fog generator has a tunable parameter β that controls the density of the fog to add to the images. This made it possible to generate samples from the dataset Cityscapes with different synthetic fog density and to use them to train an AlexNet model [100] to perform a regression problem to discover β from images. The trained fog density estimator, then, can also be used to estimate the fog density of real foggy images. The algorithm presented starts from a source domain of clear weather

images and progresses through intermediate target domains of incrementally denser fog, and, finally, reaches the target domain of dense fog images. During training, the labels of source domain and of synthetic fog images are available, while the real foggy images are unlabeled. The segmentation model, initially, is pre-trained with supervision on the source domain and then, with as many adaptation training steps as the number of denser fog steps, it is gradually shifted towards the target domain. Starting from the assumption that images with lighter fog are easier to segment, the model of the current step is used to evaluate the labels for real foggy images with intensity less than the *beta* of the current step. Then these samples are used together with images with synthetic fog of density *beta* of the next step to train the model with supervision. Iterating this process for all the steps towards the target dataset, the model adapts, in an unsupervised way (labels of the real foggy images are not used), to segment real dense fog images. In [101], the connection between curriculum learning and self-training is highlighted and a method (called self-motivated pyramid curriculum domain adaptation, PyCDA) that uses and merges both techniques is presented. The authors remind that in self-training there are two main training steps that alternate: (1) the evaluation of pseudo-labels for the target domain and (2) the supervised training of the segmentation network with the labeled source domain images and with the target domain images with pseudo-labels. In curriculum learning there are also two steps that alternate: (1) the inferring of properties of the target domain (e.g., frequency label distributions over global images or image regions, like superpixels) and (2) the update of the network parameters using the labeled source domain and the target domain inferred properties. In PyCDA the two approaches are merged: the pseudo-labels used in self-training are considered as a property of the curriculum approach. The papers also substitute the superpixels used in [96] with small squared regions to improve the algorithm efficiency and also all the curriculum properties are inferred with the segmentation networks themselves and additional models (for example, SVMs or logistic regression models) are not needed.

3.8. Multi-Tasking

Some works exploit additional types of information available in the source domain dataset, for example, depth maps, to improve the performance in the target domain. In other words, the models are trained to solve additional tasks (for example depth regression) simultaneously to image segmentation in order to build an invariant and generic embedding of the images. In [77], the authors highlight that when the source domain is made of synthetic data we could include other information about the dataset samples beyond the semantic maps, for example, depth maps. This is called Privileged Information (PI) and it includes all properties that may be useful for the training. The method proposed in [77] is called Simulator Privileged Information and Generative Adversarial Networks (SPIGAN), which uses an adversarial learning scheme performing source-to-target image translation together with a network trained on source images and on adapted images that tries to predict their privileged information (e.g., the depth map). In particular, the PI is used as regularization for the domain adaptation. A different use of the extra depth information in the source domain to enhance the appearance features and improve the alignment of the source-target domains is presented in [69]. The method introduced is called Depth-Aware Domain Adaptation (DADA) and includes a specific architecture and a learning strategy. The architecture starts from an existing segmentation network and includes some extra modules to predict the monocular depth and to feed the information of this task back in the main stream. Residual auxiliary blocks are used for this purpose. To perform domain adaptation, images from source and target domains are fed to the network to compute the class-probability and depth maps. Then, the former is processed into self-information maps and merged to the latter to produce depth aware maps. Finally, these maps are used in an adversarial training to adapt the source domain. It is important to remark that the depth information is not used as regularization, but it is directly considered while deriving prediction for the main task. In the paper, it is argued that this is a more explicit and more useful way to exploit the depth information than the method presented in [77].

A third different use of the depth maps is introduced in [102] where a method called Geometrically Guided Input-Output Adaptation (GIO-Ada) is presented. The geometric information is exploited to improve the adaptation both at the input-level and at the output-level. The former adaptation tries to reduce the visual differences of the images of the source and target domains. A transform network accepts as input source images together with their semantic maps and depth maps to compute adapted images, visually similar to images of the target domain. A discriminator is used in an adversarial learning with the transform network to distinguish real target domain images from adapted ones. The main contribution of the paper in this adaptation is the use of semantic maps and depth maps as additional inputs for the transform network. The output-level adaptation is built with a task network that computes, for each input, the semantic map and the depth map. Such outputs are fed to an additional discriminator which tries to distinguish whether they were computed from a real or adapted image. In [90], a network composed of a feature generator followed by two classifiers (that computes semantic maps) has been adopted and a maximum classifier discrepancy approach is used for the unsupervised adaptation from a synthetic source domain to a real target domain. Two techniques are presented to improve the performance of the network: a data-fusion approach and a multi-task one. The former merges the RGB image information and the depth information and use the result as an input of the network. In the latter, only RGB images are used as inputs, however three tasks are solved simultaneously by different networks after the feature generator to boost the overall performance of the network in the target domain: namely, semantic segmentation, depth regression and boundary detection.

3.9. New Research Directions

Unsupervised Domain Adaptation in its original interpretation aims at addressing the domain shift by transferring representations concerning a specific and well-defined set of semantic categories shared across source and target data. This follows the assumption that the target domain contains only instances of the classes which can be found in source samples. Nevertheless, while being a reasonable hypothesis that does not hinder the generality of the adaptation task, in practice it is common that images from a novel domain may contain objects from unseen categories.

Moving in the direction of a more general definition of the adaptation objective, some works have tackled the open-set domain adaptation [103] applied to the image classification task, which entails unknown categories peculiar to the target domain not present in the source one, but they still retain a somewhat strict prior definition of the adaptation settings class-wise. Recently, a few novel approaches [104,105] have proposed to relax the common premises on the domain adaptation settings, to effectively move towards a more realistic scenario where little can be inferred a priori about target data properties, thus widening the applicability to real-world solutions.

Saito et al. [104], for instance, introduce the Universal Domain Adaptation problem, allowing for basically no beforehand characterization of target classes. In particular, they resort to a neighborhood clustering technique to assign each target sample to either a source class or to the unknown category without any supervision. Then, the matching of cross-domain representations is enforced by entropy minimization to achieve domain alignment.

A step further is attained by solving the recognition of unseen target categories, which have to be individually learned rather than simply acknowledged as unknown. Zhuo et al. [105] address what they call the Unsupervised Open Domain Recognition task, where the objective is to learn to correctly classify target samples of unknown classes. To do so, they reduce the domain shift between source and target sets by an instance matching discrepancy minimization, weighted according to feature similarity. Once the semantic predictor has achieved domain invariance, classification knowledge can be safely transferred from known to unknown categories by a graph CNN module.

Despite that the aforementioned adaptation approaches have proven to be quite effective for the image classification task, further adjustments need to be done to deal with the additional complexity of feature representations of a semantic segmentation network. In this regard, a novel

Boundless Unsupervised Domain Adaptation (BUDA) task is proposed by Bucher et al. [24] specifically for semantic segmentation. Similarly to UODR [105], the standard domain adaptation problem is *unbounded* to explicitly handle instances of new unseen target classes, while relying solely on a minimal semantic prior in the form of class names, which are supposed to be known in advance. Thus, the overall task is decoupled in the domain adaptation and zero-shot learning problems. First, the domain adaptation of categories in common between source and target domains is performed, via an entropy minimization technique carefully designed to avoid incorrect alignment over unseen target classes. Then, a zero-shot learning strategy [106] is exploited to transfer knowledge from seen to unseen classes, by a generative model able to synthesize visual features conditioned by class descriptors.

Another closely related research direction, which is currently gaining wider and wider interest among the research community, is the continual learning task. Continual learning could be regarded as a particular case of transfer learning, where the data domain distribution changes at every incremental step and the models should perform well on all the domain distributions. For instance, in class-incremental learning, the learned model is updated to perform a new task whilst preserving previous capability. Initially proposed for image classification [107,108] and object detection [109], it has been recently explored also for semantic segmentation [110–112]. Another formulation of this problem regards the coarse-to-fine refinement at the semantic level, in which previous knowledge acquired on a coarser task is exploited to perform a finer task, hence modifying the labels distribution [113].

4. A Case Study: Synthetic to Real Adaptation for Semantic Understanding of Road Scenes

The main aspect for UDA techniques is the ability to transfer knowledge acquired on one dataset to a different context. Hence, the considered data play a fundamental role in the design and evaluation of UDA algorithms. In this section, we focus on one of the most interesting application scenarios: i.e., the ability to transfer knowledge acquired on synthetic datasets (source domain), where labels are relatively inexpensive and can be easily produced with computer graphics engines, to real world ones (target domain), where annotations are highly expensive, time-consuming and error-prone. Many of the works dealing with this task focus on urban scenes mainly for four reasons:

1. Autonomous driving is nowadays one of the biggest research areas and massive fundings support this research [114];
2. Many synthetic and real world datasets are publicly available for this scenario [8,9,23,115];
3. Autonomous vehicles should fully understand the surrounding environment to plan decisions [116] and such navigation task in the environment could be encountered in many other applications, for example, in the robotics field;
4. The first works on the topic addressed this setting and it has become the de-facto standard for performance comparison with the state-of-the-art in the UDA for semantic segmentation field.

Before presenting the more commonly used synthetic and real world datasets, we stress that in the unsupervised domain adaptation scenario the expensive labels of real samples in the target domain are not needed for training. However, a limited number of real target samples must be manually labeled for testing (and sometimes validating) the performance of the algorithms. On the other hand, large synthetic datasets corresponding to the source domain are equipped with annotations that are exploited for supervised training.

4.1. Source Domain: Synthetic Datasets of Urban Scenes

One of the first large scale synthetic datasets for urban driving is the **GTA5** dataset [23]. It contains 24,966 synthetic 1914×1052 px images with pixel-level semantic annotation. The images have been rendered using the open-world video game *Grand Theft Auto V* and are all from the car perspective in the streets of American-style virtual cities (resembling the ones in California). The images have an impressive visual quality and are very realistic since the rendering engine comes from a high budget

commercial production. The data is labeled into 19 semantic classes, which are compatible with the ones of real-world datasets as Cityscapes or Mapillary (after a proper re-mapping of labels).

The **SYNTHIA-RAND-CITYSCAPES** dataset has been sampled using the same simulator as the **SYNTHIA** dataset [115] and contains 9400 synthetic 1280 × 760 px images with pixel level semantic annotation. The images have been rendered with an ad-hoc graphic engine, allowing to obtain a large variability of photo-realistic street scenes (in this case they come from virtual European-style towns in different environments under various light and weather conditions). On the other hand, the visual quality is lower than the commercial video game GTA5. The semantic labels are compatible with 16 classes of real world datasets like Cityscapes and Mapillary. For the evaluation, either 13 or 16 classes are taken into consideration.

CARLA (CAR Learning to Act) [117] is an open-source simulator for autonomous driving research built over the Unreal Engine 4 rendering software. It has been designed to grant large flexibility and both the physics and the rendering simulations are quite realistic. Two virtual towns have been designed: *Town 1* with 2.9 km of drivable roads and *Town 2* with 1.4 km of drivable roads. Three-dimensional artists first laid roads and sidewalks, then they placed houses, terrain, vegetation, and traffic infrastructure to resemble a realistic environment. Then, dynamic objects, like cars and pedestrians, spawn from specific coordinates. The APIs of CARLA give access to a semantic segmentation camera that can distinguish 13 different classes. This feature makes it possible to sample urban datasets very quickly, easily and with a large control on the variability of samples. Another relevant aspect is that anyone can create their own dataset based on the specific needs customizing the open-source simulator.

4.2. Target Domain: Real World Datasets of Urban Scenes

The **Cityscapes** dataset [8] contains 2975 color images of resolution 2048 × 1024 px captured on the streets of 50 European cities. The images have pixel-level semantic annotation with a total of 34 semantic classes. For the evaluation of UDA approaches, typically the original training set (without the labels) is used for unsupervised adaptation, while the 500 images in the original validation set are used as a test set (since the test set labels have not been made available).

The **Mapillary** dataset [9] contains 25,000 variable (typically very high) resolution color images taken from different devices in many different locations around the world. The variability in classes, appearance, acquisition settings and geo-localization makes the dataset the most complete and the one of the highest quality in the field. The 152 object-level semantic annotations are often re-conducted to the classes present in the Cityscapes dataset, for example, following the mapping in [118]. The training images (without the labels) are used for unsupervised adaptation and the images in the original validation set are exploited as a test set.

The **Oxford RobotCar Dataset** [119] contains about 1000 km of images recorded driving in the central part of Oxford (UK). The same route (approximately 10 km long) has been repeatedly traversed for almost a year and 20 million images under different weather and light conditions have been collected by the six cameras the car was equipped with. All data are associated also to LiDAR, GPS and INS ground truth.

In [54], **Google Street View** is used to collect a large number of not annotated street images from Rome, Rio, Tokyo and Taipei. These cities have been chosen to ensure enough visual variations and the locations in the cities have been randomly selected. Using the time-machine feature of Google Street View it has been possible to capture images of the same street scene at different times in order to extract static objects priors. They collected 1600 (647 × 1280 px) image pairs at 1600 different locations per city. A set of 100 random images for each city have been annotated with pixel-level semantic labels for testing purposes.

4.3. Methods Comparison

In this section, the main results of the approaches described in the previous chapters are summarized and briefly discussed. For the sake of brevity, only the most widely used datasets are considered in this section: namely, GTA5 and SYNTHIA as source datasets, Cityscapes and Mapillary as target datasets. The others previously introduced datasets are less commonly used even if they could in principle be used being equipped with semantic segmentation annotations.

Before digging into the description of the results of existing methods, we warn the reader to be aware that different evaluation protocols and experimental setups exist making the direct comparison of the final accuracy results not always faithful. For instance, differences in input image resolution, batch size, backbone network architecture and other training parameters may alter the comparison. As for the metrics, the per-class Intersection over the Union (IoU_i for a generic class i) and the mean IoU (mIoU) are the most common. They are defined as:

$$\text{IoU}_i = \frac{TP_i}{FP_i + FN_i + TP_i} \tag{1}$$

$$\text{mIoU} = \sum_{i=1}^{N} \frac{\text{IoU}_i}{N} \tag{2}$$

where TP_i, FP_i and FN_i represent respectively the number of true positive, false positive and false negative pixels for a generic class i, and N is the number of classes.

Table 1 shows the mIoU results for different methods grouped by the employed backbone network when adapting source knowledge from GTA5 to Cityscapes. The results are grouped by backbone and we can appreciate that the results could greatly vary depending on the method and on the evaluation protocol used.

The entries in this table are scattered in Figure 7 grouped by backbone architecture to show the mIoU values and the corresponding mean (only the backbones with at least three entries considered). In general, we can see that ResNet-based approaches outperform the competitors. Moreover, the most widely diffused architectures are ResNet-101 and VGG-16. Employing the ResNet-101 architecture looks to be the best option for full comparison with the existing literature.

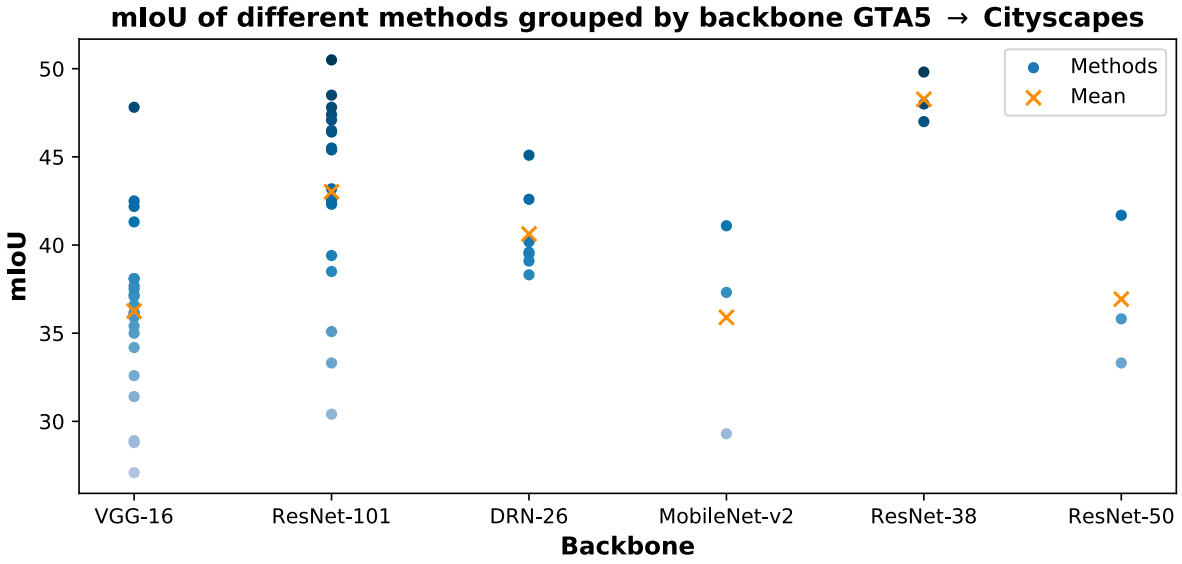

Figure 7. Mean IoU (mIoU) of different methods grouped by backbone in the scenario adapting source knowledge from GTA5 to Cityscapes (see Table 1). Backbones are sorted by decreasing the number of entries. Orange crosses represent the per-backbone mean mIoU. Only the backbones with 3 or more entries are displayed.

Similarly, Table 2 reports the results grouped by backbone when adapting source knowledge from SYNTHIA to Cityscapes. The table reports two setups, i.e., considering either 13 or 16 classes, since both are often considered in this case. The respective entries of $mIoU_{16}$ are scattered in Figure 8 to give an overview of the most widely used techniques and of the results achieved. Also in this scenario, VGG-16 is the most popular architecture followed by ResNet-101, which generally shows higher results.

Table 1. Mean IoU (mIoU) for different methods grouped by backbone in the scenario adapting source knowledge from GTA5 to Cityscapes.

Method	Backbone	mIoU	Method	Backbone	mIoU
Biasetton et al. [65]	ResNet-101	30.4	Chen et al. [46]	VGG-16	35.9
Chang et al. [62]	ResNet-101	45.4	Chen et al. [51]	VGG-16	38.1
Chen et al. [46]	ResNet-101	39.4	Choi et al. [78]	VGG-16	42.5
Chen et al. [95]	ResNet-101	46.4	Du et al. [55]	VGG-16	37.7
Du et al. [55]	ResNet-101	45.4	Hoffman et al. [45]	VGG-16	27.1
Gong et al. [75]	ResNet-101	42.3	Hoffman et al. [50]	VGG-16	35.4
Hoffman et al. [50]	ResNet-101	42.7 *	Huang et al. [49]	VGG-16	32.6
Li et al. [48]	ResNet-101	48.5	Li et al. [48]	VGG-16	41.3
Lian et al. [101]	ResNet-101	47.4	Lian et al. [101]	VGG-16	37.2
Luo et al. [52]	ResNet-101	42.6	Luo et al. [52]	VGG-16	34.2
Luo et al. [63]	ResNet-101	43.2	Luo et al. [63]	VGG-16	36.6
Michieli et al. [66]	ResNet-101	33.3	Saito et al. [89]	VGG-16	28.8
Spadotto et al. [67]	ResNet-101	35.1	Sankaranarayanan et al. [59]	VGG-16	37.1
Tsai et al. [60]	ResNet-101	42.4	Tsai et al. [60]	VGG-16	35.0
Tsai et al. [70]	ResNet-101	46.5	Tsai et al. [70]	VGG-16	37.5
Vu et al. [68]	ResNet-101	45.5	Vu et al. [68]	VGG-16	36.1
Wu et al. [82]	ResNet-101	38.5	Wu et al. [82]	VGG-16	36.2
Yang et al. [25]	ResNet-101	50.5	Yang et al. [25]	VGG-16	42.2
Zhang et al. [47]	ResNet-101	47.8	Zhang et al. [96]	VGG-16	28.9
Zou et al. [94]	ResNet-101	47.1	Zhang et al. [97]	VGG-16	31.4
Murez et al. [58]	ResNet-34	31.8	Zhou et al. [71]	VGG-16	47.8
Lian et al. [101]	ResNet-38	48.0	Zhu et al. [57]	VGG-16	38.1 *
Zou et al. [93]	ResNet-38	47.0	Zou et al. [93]	VGG-16	36.1
Zou et al. [94]	ResNet-38	49.8	Hong et al. [79]	VGG-19	44.5
Lee et al. [91]	ResNet-50	35.8	Chen et al. [51]	DRN-26	45.1
Saito et al. [88]	ResNet-50	33.3	Dundar et al. [84]	DRN-26	38.3
Wu et al. [82]	ResNet-50	41.7	Hoffman et al. [50]	DRN-26	39.5
Hoffman et al. [50]	MobileNet-v2	37.3 *	Huang et al. [49]	DRN-26	40.2
Toldo et al. [53]	MobileNet-v2	41.1	Liu et al. [120]	DRN-26	39.1 *
Zhu et al. [76]	MobileNet-v2	29.3 *	Yang et al. [74]	DRN-26	42.6
Murez et al. [58]	DenseNet	35.7	Zhu et al. [76]	DRN-26	39.6 *
Huang et al. [49]	ERFNet	31.3	Saito et al. [89]	DRN-105	39.7

*: values from results of competing works.

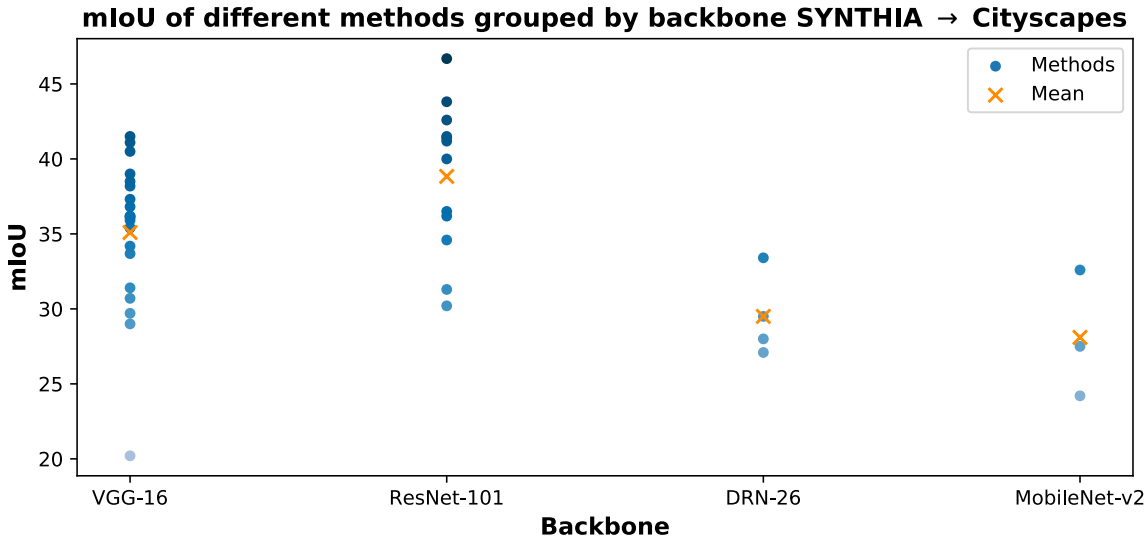

Figure 8. Mean IoU on 16 classes ($mIoU_{16}$) of different methods grouped by backbone in the scenario adapting source knowledge from SYNTHIA to Cityscapes (see Table 2). Backbones are sorted by decreasing number of entries. Orange crosses represent the per-backbone mean mIoU. Only the backbones with 3 or more entries are displayed.

Table 2. Mean IoU (mIoU) for different methods grouped by backbone in the scenario adapting source knowledge from SYNTHIA to Cityscapes. The table reports the mIoU computed over 13 or 16 semantic classes depending on the label set used in the corresponding paper.

Method	Backbone	$mIoU_{13}$	$mIoU_{16}$	Method	Backbone	$mIoU_{13}$	$mIoU_{16}$
Biasetton et al. [65]	ResNet-101	-	30.2	Chen et al. [54]	VGG-16	35.7	-
Bucher et al. [24]	ResNet-101	-	36.2	Chen et al. [46]	VGG-16	-	36.2
Chang et al. [62]	ResNet-101	-	41.5	Chen et al. [46]	VGG-16	41.8 *	36.2 *
Chen et al. [95]	ResNet-101	48.2	41.4	Chen et al. [51]	VGG-16	-	38.2
Du et al. [55]	ResNet-101	50.0	-	Chen et al. [102]	VGG-16	43.0	37.3
Li et al. [48]	ResNet-101	51.4	-	Choi et al. [78]	VGG-16	46.6	38.5
Lian et al. [101]	ResNet-101	53.3	46.7	Du et al. [55]	VGG-16	43.4	-
Luo et al. [52]	ResNet-101	46.3	-	Hoffman et al. [45]	VGG-16	17.0	20.2 *
Luo et al. [63]	ResNet-101	47.8	-	Huang et al. [49]	VGG-16	-	30.7 *
Michieli et al. [66]	ResNet-101	-	31.3	Lee et al. [77]	VGG-16	42.4 *	36.8
Spadotto et al. [67]	ResNet-101	-	34.6	Li et al. [48]	VGG-16	-	39.0
Tsai et al [70]	ResNet-101	46.5	40.0	Lian et al. [101]	VGG-16	42.6	35.9
Tsai et al. [60]	ResNet-101	46.7	-	Luo et al. [63]	VGG-16	39.3	-
Vu et al. [68]	ResNet-101	48.0	41.2	Luo et al. [52]	VGG-16	37.2	-
Vu et al. [69]	ResNet-101	49.8	42.6	Sankaran. et al. [59]	VGG-16	42.1 *	36.1
Wu et al. [82]	ResNet-101	-	36.5	Tsai et al [70]	VGG-16	39.6	33.7
Yang et al. [25]	ResNet-101	52.5	-	Tsai et al. [60]	VGG-16	37.6	-
Zou et al. [94]	ResNet-101	50.1	43.8	Vu et al. [68]	VGG-16	36.6	31.4
Zou et al. [93]	ResNet-38	-	38.4	Wu et al. [82]	VGG-16	-	35.4
Wu et al. [82]	ResNet-50	48.4	42.5	Yang et al. [25]	VGG-16	-	40.5
Hoffman et al [50]	MobileNet-v2	-	27.5 *	Yang et al. [74]	VGG-16	48.7	41.1
Toldo et al. [53]	MobileNet-v2	-	32.6	Zhang et al. [96]	VGG-16	34.8 *	29.0
Zhu et al. [76]	MobileNet-v2	-	24.2 *	Zhang et al. [97]	VGG-16	-	29.7
Chen et al. [51]	DRN-26	-	33.4	Zhou et al. [71]	VGG-16	48.6	41.5
Dundar et al. [84]	DRN-26	-	29.5	Zhu et al. [57]	VGG-16	40.3 *	34.2 *
Liu et al. [120]	DRN-26	-	28.0*	Zou et al. [93]	VGG-16	36.1	35.4
Zhu et al. [76]	DRN-26	-	27.1 *	Hong et al. [79]	VGG-19	-	41.2
Saito et al. [89]	DRN-105	43.5 *	37.3 *				

*: values from results of competing works.

Some recent works also consider the Mapillary dataset adapting from either GTA5 or SYNTHIA as source datasets, as before. While, in this case, the comparison is quite limited [65–67] (currently, to the best of our knowledge, the highest performing approach is [67], which achieves a mIoU of 41.9 when adapting from GTA5).

5. Conclusions and Future Directions

In this paper, we presented a comprehensive overview of the recent advancements in Unsupervised Domain Adaptation for semantic segmentation. This is a very relevant task since deep learning architectures for semantic segmentation require a huge amount of labeled training samples, which in many practical settings are not available due to the complex labeling procedure. For this reason, a wide range of different UDA approaches for this task have been proposed in the recent years.

In order to organize the wide range of existing approaches we started from grouping them at a high-level, based on where the domain adaptation is performed: namely, at input-level (i.e., on the images provided to the network), at feature-level, at output-level or at some ad-hoc network levels.

After this macroscopic subdivision, we moved to the actual review of the literature in the field, dividing the existing works into seven (non mutually exclusive) categories: i.e., based on adversarial learning, on generative approaches, on the analysis of the classifier discrepancies, on self-training, on entropy minimization, on curriculum learning and, finally, on multi-task learning. For each category, we presented the most successful approaches and we summarized the main ideas of each contribution.

Then, we considered a case study: the synthetic to real adaptation for semantic understanding of road scenes. Besides being a very relevant task since it is one of the key enabling technologies for autonomous driving, it has also been used for the evaluation of many papers in the field and we concluded the survey comparing the accuracy of many different works grouped by the backbone architecture on this task.

We believe that UDA for semantic segmentation is an open research field with large room for improvements, as proved by the fact that even the best approaches have performance still far from the ones of supervised training on the target dataset. More refined and performing approaches based on the various schemes presented in this review are continuously appearing.

More complete and variable datasets are also needed to deal with a wider range of real world scenarios and, to this extent, we believe that the Mapillary dataset should be taken into consideration for future works on synthetic to real adaptation in the autonomous driving scenario.

Further novel research directions will regard open problems in the field, as for example, open-set and boundless-set UDA in the semantic segmentation task. Additionally, knowledge acquired for UDA may be beneficial for other closely related tasks such as continual learning, where the data distribution changes multiple times.

Author Contributions: Conceptualization, M.T., U.M., A.M. and P.Z.; Supervision, M.T., U.M. and P.Z.; Writing—original draft, M.T., U.M., A.M.; and Writing—review and editing, M.T., U.M., A.M. and P.Z. All authors have read and agreed to the published version of the manuscript.

Funding: Our work was in part supported by the Italian Minister for Education (MIUR) under the “Departments of Excellence" initiative (Law 232/2016).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wang, M.; Deng, W. Deep visual domain adaptation: A survey. *Neurocomputing* **2018**, *312*, 135–153. [CrossRef]
2. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.

3. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
4. Yu, F.; Koltun, V.; Funkhouser, T.A. Dilated Residual Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 636–644.
5. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)* **2018**, *40*, 834–848. [CrossRef] [PubMed]
6. Chen, L.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 833–851.
7. Chen, L.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv* **2017**, arXiv:1706.05587.
8. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The Cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), LasVegas, NV, USA, 27–30 June 2016; pp. 3213–3223.
9. Neuhold, G.; Ollmann, T.; Rota Bulo, S.; Kontschieder, P. The Mapillary vistas dataset for semantic understanding of street scenes. In Proceedings of the International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 4990–4999.
10. Everingham, M.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The Pascal Visual Object Classes (VOC) challenge. *Int. J. Comput. Vis. (IJCV)* **2010**, *88*, 303–338. [CrossRef]
11. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*; Springer: Berlin, Germany, 2014; pp. 740–755.
12. Zhou, B.; Zhao, H.; Puig, X.; Fidler, S.; Barriuso, A.; Torralba, A. Scene parsing through ade20k dataset. In Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 633–641.
13. Silberman, N.; Derek Hoiem, P.K.; Fergus, R. Indoor Segmentation and Support Inference from RGBD Images. In Proceedings of the European Conference on Computer Vision (ECCV), Florence, Italy, 7–13 October 2012.
14. Song, S.; Lichtenberg, S.P.; Xiao, J. Sun rgb-d: A rgb-d scene understanding benchmark suite. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 567–576.
15. Sun, S.; Shi, H.; Wu, Y. A survey of multi-source domain adaptation. *Inf. Fusion* **2015**, *24*, 84–92. [CrossRef]
16. Csurka, G. Domain adaptation for visual applications: A comprehensive survey. *arXiv* **2017**, arXiv:1702.05374.
17. Jiang, J.; Zhai, C. Instance weighting for domain adaptation in NLP. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, 23–30 June 2007; pp. 264–271.
18. Fang, F.; Dutta, K.; Datta, A. Domain adaptation for sentiment classification in light of multiple sources. *INFORMS J. Comput.* **2014**, *26*, 586–598. [CrossRef]
19. Jiang, J. A Literature Survey on Domain Adaptation of Statistical Classifiers. 2008. Available online: http://www.mysmu.edu/faculty/jingjiang/papers/da_survey.pdf (accessed on 19 June 2020)
20. Patel, V.M.; Gopalan, R.; Li, R.; Chellappa, R. Visual Domain Adaptation: A survey of recent advances. *IEEE Signal Process. Mag.* **2015**, *32*, 53–69. [CrossRef]
21. Ho, H.T.; Gopalan, R. Model-driven domain adaptation on product manifolds for unconstrained face recognition. *Int. J. Comput. Vis. (IJCV)* **2014**, *109*, 110–125. [CrossRef]
22. Saenko, K.; Kulis, B.; Fritz, M.; Darrell, T. Adapting visual category models to new domains. In Proceedings of the European Conference on Computer Vision (ECCV), Crete, Greece, 5–10 September 2010; Springer: Berlin/Heidelberg, Germany, 2010; pp. 213–226.

23. Richter, S.R.; Vineet, V.; Roth, S.; Koltun, V. Playing for Data: Ground Truth from Computer Games. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; Leibe, B.; Matas, J., Sebe, N., Welling, M., Eds. Springer International Publishing: Berlin/Heidelberg, Germany, 2016; Volume 9906, pp. 102–118.
24. Bucher, M.; Vu, T.H.; Cord, M.; Pérez, P. BUDA: Boundless Unsupervised Domain Adaptation in Semantic Segmentation. *arXiv* **2020**, arXiv:2004.01130.
25. Yang, Y.; Soatto, S. FDA: Fourier Domain Adaptation for Semantic Segmentation. *arXiv* **2020**, arXiv:2004.05498.
26. Vezhnevets, A.; Buhmann, J.M. Towards weakly supervised semantic segmentation by means of multiple instance and multitask learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA, 13–18 June 2010; pp. 3249–3256.
27. Pathak, D.; Shelhamer, E.; Long, J.; Darrell, T. Fully convolutional multi-class multiple instance learning. *arXiv* **2014**, arXiv:1412.7144.
28. Papandreou, G.; Chen, L.C.; Murphy, K.P.; Yuille, A.L. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In Proceedings of the International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1742–1750.
29. Pathak, D.; Krahenbuhl, P.; Darrell, T. Constrained convolutional neural networks for weakly supervised segmentation. In Proceedings of the International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1796–1804.
30. Wei, Y.; Liang, X.; Chen, Y.; Shen, X.; Cheng, M.M.; Feng, J.; Zhao, Y.; Yan, S. STC: A simple to complex framework for weakly-supervised semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)* **2017**, *39*, 2314–2320. [CrossRef] [PubMed]
31. Hong, S.; Noh, H.; Han, B. Decoupled deep neural network for semi-supervised semantic segmentation. In Proceedings of the Neural Information Processing Systems (NeurIPS), Montreal, QC, Canada, 7–12 December 2015; pp. 1495–1503.
32. Dai, J.; He, K.; Sun, J. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In Proceedings of the International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1635–1643.
33. Souly, N.; Spampinato, C.; Shah, M. Semi and weakly supervised semantic segmentation using generative adversarial network. *arXiv* **2017**, arXiv:1703.09695.
34. Kolesnikov, A.; Lampert, C.H. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin, Germany, 2016; pp. 695–711.
35. Huang, Z.; Wang, X.; Wang, J.; Liu, W.; Wang, J. Weakly-supervised semantic segmentation network with deep seeded region growing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 7014–7023.
36. Wei, Y.; Feng, J.; Liang, X.; Cheng, M.M.; Zhao, Y.; Yan, S. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1568–1576.
37. Ahn, J.; Kwak, S. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4981–4990.
38. Lee, J.; Kim, E.; Lee, S.; Lee, J.; Yoon, S. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 5267–5276.
39. Ahn, J.; Cho, S.; Kwak, S. Weakly supervised learning of instance segmentation with inter-pixel relations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 2209–2218.
40. Ramirez, P.Z.; Tonioni, A.; Salti, S.; Stefano, L.D. Learning Across Tasks and Domains. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019.
41. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the Neural Information Processing Systems (NeurIPS), Montreal, QC, USA, 8–13 December 2014; pp. 2672–2680.

42. Ganin, Y.; Lempitsky, V. Unsupervised Domain Adaptation by Backpropagation. In Proceedings of the International Conference on Machine Learning (ICML), Lille, France, 7–9 July 2015; pp. 1180–1189.
43. Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; Lempitsky, V. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.* **2016**, *17*, 1–35.
44. Tzeng, E.; Hoffman, J.; Saenko, K.; Darrell, T. Adversarial discriminative domain adaptation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 7167–7176.
45. Hoffman, J.; Wang, D.; Yu, F.; Darrell, T. FCNs in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv* **2016**, arXiv:1612.02649.
46. Chen, Y.; Li, W.; Van Gool, L. Road: Reality oriented adaptation for semantic segmentation of urban scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 7892–7901.
47. Zhang, Y.; Qiu, Z.; Yao, T.; Liu, D.; Mei, T. Fully convolutional adaptation networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 6810–6818.
48. Li, Y.; Yuan, L.; Vasconcelos, N. Bidirectional Learning for Domain Adaptation of Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019.
49. Huang, H.; Huang, Q.; Krähenbühl, P. Domain Transfer Through Deep Activation Matching. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
50. Hoffman, J.; Tzeng, E.; Park, T.; Zhu, J.Y.; Isola, P.; Saenko, K.; Efros, A.; Darrell, T. CyCADA: Cycle-Consistent Adversarial Domain Adaptation. In Proceedings of the International Conference on Machine Learning (ICML), Stockholm, Sweden, 10–15 July 2018.
51. Chen, Y.C.; Lin, Y.Y.; Yang, M.H.; Huang, J.B. CrDoCo: Pixel-Level Domain Transfer With Cross-Domain Consistency. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019.
52. Luo, Y.; Liu, P.; Guan, T.; Yu, J.; Yang, Y. Significance-Aware Information Bottleneck for Domain Adaptive Semantic Segmentation. In Proceedings of the International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019.
53. Toldo, M.; Michieli, U.; Agresti, G.; Zanuttigh, P. Unsupervised Domain Adaptation for Mobile Semantic Segmentation based on Cycle Consistency and Feature Alignment. *arXiv* **2020**, arXiv:2001.04692.
54. Chen, Y.H.; Chen, W.Y.; Chen, Y.T.; Tsai, B.C.; Frank Wang, Y.C.; Sun, M. No more discrimination: Cross city adaptation of road scene segmenters. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1992–2001.
55. Du, L.; Tan, J.; Yang, H.; Feng, J.; Xue, X.; Zheng, Q.; Ye, X.; Zhang, X. SSF-DAN: Separated Semantic Feature Based Domain Adaptation Network for Semantic Segmentation. In Proceedings of the International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019.
56. Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; Li, F. ImageNet: A large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, 20–25 June 2009, pp. 248–255.
57. Zhu, X.; Zhou, H.; Yang, C.; Shi, J.; Lin, D. Penalizing top performers: Conservative loss for semantic segmentation adaptation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 568–583.
58. Murez, Z.; Kolouri, S.; Kriegman, D.J.; Ramamoorthi, R.; Kim, K. Image to Image Translation for Domain Adaptation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Munich, Germany, 8–14 September 2018.
59. Sankaranarayanan, S.; Balaji, Y.; Jain, A.; Nam Lim, S.; Chellappa, R. Learning from synthetic data: Addressing domain shift for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Munich, Germany, 8–14 September 2018; pp. 3752–3761.
60. Tsai, Y.H.; Hung, W.C.; Schulter, S.; Sohn, K.; Yang, M.H.; Chandraker, M. Learning to adapt structured output space for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Munich, Germany, 8–14 September 2018; pp. 7472–7481.

61. Chen, Y.; Li, W.; Chen, X.; Van Gool, L. Learning Semantic Segmentation from Synthetic Data: A Geometrically Guided Input-Output Adaptation Approach. *arXiv* **2018**, arXiv:1812.05040.
62. Chang, W.; Wang, H.; Peng, W.; Chiu, W. All About Structure: Adapting Structural Information Across Domains for Boosting Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 1900–1909.
63. Luo, Y.; Zheng, L.; Guan, T.; Yu, J.; Yang, Y. Taking A Closer Look at Domain Shift: Category-level Adversaries for Semantics Consistent Domain Adaptation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019.
64. Yang, J.; An, W.; Wang, S.; Zhu, X.; Yan, C.; Huang, J. Label-Driven Reconstruction for Domain Adaptation in Semantic Segmentation. *arXiv* **2020**, arXiv:2003.04614.
65. Biasetton, M.; Michieli, U.; Agresti, G.; Zanuttigh, P. Unsupervised Domain Adaptation for Semantic Segmentation of Urban Scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Long Beach, CA, USA, 16–20 June 2019.
66. Michieli, U.; Biasetton, M.; Agresti, G.; Zanuttigh, P. Adversarial Learning and Self-Teaching Techniques for Domain Adaptation in Semantic Segmentation. *IEEE Trans. Intell. Veh.* **2020**. [CrossRef]
67. Spadotto, T.; Toldo, M.; Michieli, U.; Zanuttigh, P. Unsupervised Domain Adaptation with Multiple Domain Discriminators and Adaptive Self-Training. *arXiv* **2020**, arXiv:2004.12724.
68. Vu, T.H.; Jain, H.; Bucher, M.; Cord, M.; Pérez, P. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 2517–2526.
69. Vu, T.; Jain, H.; Bucher, M.; Cord, M.; Pérez, P. DADA: Depth-Aware Domain Adaptation in Semantic Segmentation. In Proceedings of the International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 7363–7372.
70. Tsai, Y.H.; Sohn, K.; Schulter, S.; Chandraker, M. Domain Adaptation for Structured Output via Discriminative Patch Representations. In Proceedings of the International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 1456–1465.
71. Zhou, Q.; Feng, Z.; Cheng, G.; Tan, X.; Shi, J.; Ma, L. Uncertainty-Aware Consistency Regularization for Cross-Domain Semantic Segmentation. *arXiv* **2020**, arXiv:2004.08878.
72. Qin, C.; Wang, L.; Zhang, Y.; Fu, Y. Generatively Inferential Co-Training for Unsupervised Domain Adaptation. In Proceedings of the International Conference on Computer Vision Workshops (ICCVW), Seoul, Korea, 27 October–2 November 2019; pp. 1055–1064.
73. Li, P.; Liang, X.; Jia, D.; Xing, E.P. Semantic-aware Grad-GAN for Virtual-to-Real Urban Scene Adaption. In Proceedings of the British Machine Vision Conference (BMVC), Newcastle, UK, 3–6 September 2018.
74. Yang, Y.; Lao, D.; Sundaramoorthi, G.; Soatto, S. Phase Consistent Ecological Domain Adaptation. *arXiv* **2020**, arXiv:2004.04923.
75. Gong, R.; Li, W.; Chen, Y.; Gool, L.V. DLOW: Domain Flow for Adaptation and Generalization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 2477–2486.
76. Zhu, J.; Park, T.; Isola, P.; Efros, A.A. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In Proceedings of the International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
77. Lee, K.; Ros, G.; Li, J.; Gaidon, A. SPIGAN: Privileged Adversarial Learning from Simulation. In Proceedings of the International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019.
78. Choi, J.; Kim, T.; Kim, C. Self-Ensembling With GAN-Based Data Augmentation for Domain Adaptation in Semantic Segmentation. In Proceedings of the International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 6830–6840.
79. Hong, W.; Wang, Z.; Yang, M.; Yuan, J. Conditional Generative Adversarial Network for Structured Domain Adaptation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 1335–1344.
80. Pizzati, F.; Charette, R.d.; Zaccaria, M.; Cerri, P. Domain bridge for unpaired image-to-image translation and unsupervised domain adaptation. In Proceedings of the British Winter Conference on Applications of Computer Vision (WACV), Snowmass Village, CO, USA, 1–5 March 2020; pp. 2990–2998.

81. Huang, X.; Liu, M.; Belongie, S.J.; Kautz, J. Multimodal Unsupervised Image-to-Image Translation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 179–196.
82. Wu, Z.; Han, X.; Lin, Y.; Uzunbas, M.G.; Goldstein, T.; Lim, S.; Davis, L.S. DCAN: Dual Channel-Wise Alignment Networks for Unsupervised Scene Adaptation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 535–552.
83. Wu, Z.; Wang, X.; Gonzalez, J.; Goldstein, T.; Davis, L. ACE: Adapting to Changing Environments for Semantic Segmentation. In Proceedings of the International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 2121–2130.
84. Dundar, A.; Liu, M.; Wang, T.; Zedlewski, J.; Kautz, J. Domain Stylization: A Strong, Simple Baseline for Synthetic to Real Image Domain Adaptation. *arXiv* **2018**, arXiv:1807.09384.
85. Gatys, L.A.; Ecker, A.S.; Bethge, M. Texture Synthesis Using Convolutional Neural Networks. In Proceedings of the Neural Information Processing Systems (NeurIPS), Montreal, QC, Canada, 7–12 December 2015; pp. 262–270.
86. Gatys, L.A.; Ecker, A.S.; Bethge, M. Image Style Transfer Using Convolutional Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), LasVegas, NV, USA, 27–30 June 2016; pp. 2414–2423
87. Huang, X.; Belongie, S.J. Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization. In Proceedings of the International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 1510–1519.
88. Saito, K.; Ushiku, Y.; Harada, T.; Saenko, K. Adversarial Dropout Regularization. In Proceedings of the International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018.
89. Saito, K.; Watanabe, K.; Ushiku, Y.; Harada, T. Maximum Classifier Discrepancy for Unsupervised Domain Adaptation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 3723–3732.
90. Watanabe, K.; Saito, K.; Ushiku, Y.; Harada, T. Multichannel Semantic Segmentation with Unsupervised Domain Adaptation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
91. Lee, S.; Kim, D.; Kim, N.; Jeong, S.G. Drop to Adapt: Learning Discriminative Features for Unsupervised Domain Adaptation. In Proceedings of the International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 91–100.
92. Grandvalet, Y.; Bengio, Y. Semi-supervised Learning by Entropy Minimization. In Proceedings of the Actes de CAP 05, Conférence Francophone sur L'apprentissage Automatique, Nice, France, 31 May–3 June 2005; pp. 281–296.
93. Zou, Y.; Yu, Z.; Vijaya Kumar, B.; Wang, J. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 289–305.
94. Zou, Y.; Yu, Z.; Liu, X.; Kumar, B.V.; Wang, J. Confidence Regularized Self-Training. In Proceedings of the International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 5982–5991.
95. Chen, M.; Xue, H.; Cai, D. Domain Adaptation for Semantic Segmentation With Maximum Squares Loss. In Proceedings of the International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019.
96. Zhang, Y.; David, P.; Gong, B. Curriculum domain adaptation for semantic segmentation of urban scenes. In Proceedings of the International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2020–2030.
97. Zhang, Y.; David, P.; Foroosh, H.; Gong, B. A curriculum domain adaptation approach to the semantic segmentation of urban scenes. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)* **2019**, in press. [CrossRef] [PubMed]
98. Sakaridis, C.; Dai, D.; Hecker, S.; Van Gool, L. Model adaptation with synthetic and real data for semantic dense foggy scene understanding. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 687–704.

99. Dai, D.; Sakaridis, C.; Hecker, S.; Van Gool, L. Curriculum model adaptation with synthetic and real data for semantic foggy scene understanding. *Int. J. Comput. Vis. (IJCV)* **2019**, *128*, 1182–1204. [CrossRef]
100. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In Proceedings of the Neural Information Processing Systems (NeurIPS), Lake Tahoe, NV, USA, 3–6 December 2012, pp. 1106–1114.
101. Lian, Q.; Lv, F.; Duan, L.; Gong, B. Constructing Self-motivated Pyramid Curriculums for Cross-Domain Semantic Segmentation: A Non-Adversarial Approach. In Proceedings of the International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 6758–6767.
102. Chen, Y.; Li, W.; Chen, X.; Gool, L.V. Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seoul, Korea, 27 October–2 November 2019; pp. 1841–1850.
103. Busto, P.P.; Gall, J. Open Set Domain Adaptation. In Proceedings of the International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 754–763.
104. Saito, K.; Kim, D.; Sclaroff, S.; Saenko, K. Universal Domain Adaptation through Self Supervision. *arXiv* **2020**, arXiv:2002.07953.
105. Zhuo, J.; Wang, S.; Cui, S.; Huang, Q. Unsupervised Open Domain Recognition by Semantic Discrepancy Minimization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seoul, Korea, 27 October–2 November 2019; pp. 750–759.
106. Bucher, M.; Vu, T.; Cord, M.; Pérez, P. Zero-Shot Semantic Segmentation. In Proceedings of the Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 8–14 December 2019; pp. 466–477.
107. Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A.A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; et al. Overcoming catastrophic forgetting in neural networks. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 3521–3526. [CrossRef]
108. Li, Z.; Hoiem, D. Learning without forgetting. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)* **2017**, *40*, 2935–2947. [CrossRef]
109. Shmelkov, K.; Schmid, C.; Alahari, K. Incremental learning of object detectors without catastrophic forgetting. In Proceedings of the International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 3400–3409.
110. Michieli, U.; Zanuttigh, P. Incremental learning techniques for semantic segmentation. In Proceedings of the International Conference on Computer Vision Workshops (ICCVW), Seoul, Korea, 27 October–2 November 2019.
111. Michieli, U.; Zanuttigh, P. Knowledge Distillation for Incremental Learning in Semantic Segmentation. *arXiv* **2019**, arXiv:1911.03462.
112. Cermelli, F.; Mancini, M.; Bulò, S.R.; Ricci, E.; Caputo, B. Modeling the Background for Incremental Learning in Semantic Segmentation. *arXiv* **2020**, arXiv:2002.00718.
113. Mel, M.; Michieli, U.; Zanuttigh, P. Incremental and Multi-Task Learning Strategies for Coarse-To-Fine Semantic Segmentation. *Technologies* **2020**, *8*, 1. [CrossRef]
114. Smith, D.; Burke, B. Gartner's 2019 Hype Cycle for Emerging Technologies. Technical Report; Gartner: Stamford, CT, USA, August 2019.
115. Ros, G.; Sellart, L.; Materzynska, J.; Vazquez, D.; Lopez, A.M. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Vegas, NV, USA, 27–30 June 2016; pp. 3234–3243.
116. Michieli, U.; Badia, L. Game Theoretic Analysis of Road User Safety Scenarios Involving Autonomous Vehicles. In Proceedings of the IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications, Bologna, Italy, 9–12 September 2018; pp. 1377–1381.
117. Dosovitskiy, A.; Ros, G.; Codevilla, F.; Lopez, A.; Koltun, V. CARLA: An open urban driving simulator. *arXiv* **2017**, arXiv:1711.03938.
118. Kim, J.; Park, C. Attribute Dissection of Urban Road Scenes for Efficient Dataset Integration. In Proceedings of the International Joint Conference on Artificial Intelligence Workshops, Stockholm, Sweden, 13–15 July 2018; pp. 8–15.

119. Maddern, W.; Pascoe, G.; Linegar, C.; Newman, P. 1 year, 1000 km: The Oxford RobotCar dataset. *Int. J. Robot. Res.* **2017**, *36*, 3–15. [CrossRef]
120. Liu, M.Y.; Breuel, T.; Kautz, J. Unsupervised image-to-image translation networks. In Proceedings of the Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 700–708.