A Survey on Deep Learning-Based Semi-Supervised Semantic Segmentation

ADRIÁN PELÁEZ-VARGAS, Computer Science and Artificial Intelligence, University of Granada, Granada, Spain and Andalusian Institute of Data Science and Computational Intelligence (DaSCI), Granada, Spain
IGNACIO AGUILERA-MARTOS, Computer Science and Artificial Intelligence, University of Granada, Granada, Spain and Andalusian Institute of Data Science and Computational Intelligence (DaSCI), Granada, Spain
PABLO MESEJO, Computer Science and Artificial Intelligence, University of Granada, Granada, Spain and Andalusian Institute of Data Science and Computational Intelligence (DaSCI), Granada, Spain
JULIÁN LUENGO, Computer Science and Artificial Intelligence, University of Granada, Granada, Spain and Andalusian Institute of Data Science and Computational Intelligence (DaSCI), Granada, Spain

Semantic segmentation is one of the most challenging tasks in computer vision. However, in many applications, a frequent obstacle is the lack of labeled images, due to the high cost of pixel-level labeling. In this scenario, it makes sense to approach the problem from a semi-supervised point of view, where both labeled and unlabeled images are exploited. In recent years this line of research has gained much interest and many approaches have been published in this direction. Therefore, the main objective of this study is to provide an overview of the current state of the art in semi-supervised semantic segmentation, offering an updated taxonomy of all existing methods to date. This is complemented by an experimentation with a variety of models representing all the categories of the taxonomy on the most widely used benchmark datasets in the literature, and a final discussion on the results obtained, the challenges and the most promising lines of future research.

CCS Concepts: • **Computing methodologies** → **Image segmentation**.

Additional Key Words and Phrases: Image segmentation, semi-supervised semantic segmentation, semi-supervised learning, deep learning, convolutional neural networks, adversarial methods, pseudo-labeling, consistency regularization, contrastive learning

1 Introduction

Segmentation is one of the oldest and most widely studied computer vision (CV) problems [86, 125]. It consists of dividing an image into different non-overlapping regions and assigning the corresponding label to each pixel in the image. This task can be considered as a pixel-level classification problem, which leads to a significant increase in complexity compared to other CV problems, such as image-level classification or object detection [125]. We can differentiate between two different types of image segmentation problems. On one hand, semantic segmentation classifies each pixel with the corresponding semantic class, thus giving the same class label to

Authors' Contact Information: Adrián Peláez-Vargas, Computer Science and Artificial Intelligence, University of Granada, Granada, Granada, Spain and Andalusian Institute of Data Science and Computational Intelligence (DaSCI), Granada, Granada, Spain; e-mail: adrianpelaez@ugr.es; Ignacio Aguilera-Martos, Computer Science and Artificial Intelligence, University of Granada, Granada, Granada, Spain and Andalusian Institute of Data Science and Computational Intelligence (DaSCI), Granada, Granada, Spain; e-mail: nacheteam@ugr.es; Pablo Mesejo, Computer Science and Artificial Intelligence, University of Granada, Granada, Granada, Spain and Andalusian Institute of Data Science and Computational Intelligence (DaSCI), Granada, Granada, Spain; e-mail: pmesejo@decsai.ugr.es; Julián Luengo, Computer Science and Artificial Intelligence, University of Granada, Granada, Granada, Spain and Andalusian Institute of Data Science and Computational Intelligence (DaSCI), Granada, Granada, Spain; e-mail: julianlm@decsai.ugr.es.

ACM 1557-7341/2026/4-ART
https://doi.org/10.1145/3806766

all objects or regions of the image that belong to this class. On the other hand, instance segmentation attempts to go one step further and tries to distinguish between different occurrences of the same class (Figure 2). This paper focuses on semantic segmentation (SS), which has gained much interest in recent years with important applications in different areas such as medical imaging [83], autonomous driving [95], aerial scene analysis [91] or metallographic images [80], among others [60, 106].

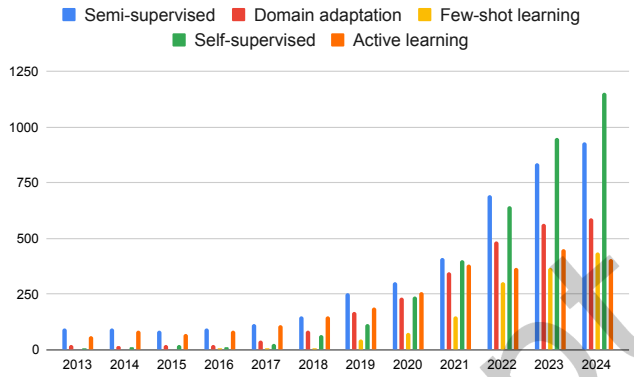


Fig. 1. Histogram of publications related to different learning approaches with sparse annotations in SS. The queries, run in Elsevier Scopus on the 12th of December, 2024, showing a clear growing tendency.

Methods based on deep learning (DL) have recently shown great potential [41, 67, 107], becoming the state-of-the-art methods in many CV problems [52, 75, 86, 138, 141]. The SS problem has traditionally been addressed by classical image processing and CV techniques, such as thresholding techniques or clustering algorithms [11, 110]. However, with the emergence of DL methods it has been possible to give a leap of quality in the segmentation results, as in many other CV problems. That is why all the semi-supervised methods that form the state of the art in semi-supervised SS, and included in this study, are based on DL.

Depending on the degree of detail of the ground truth labeling (i.e. the output considered correct) and the portion of labeled examples in relation to the total number of available images we can face different scenarios: fully supervised scenario, weakly supervised scenario, semi-supervised scenario and unsupervised scenario. Due to the difficulty and effort involved in labeling images at the pixel level, approaches based on semi-supervised learning (SSL) [9], in which we have a reduced amount of labeled images, and a larger amount of unlabeled images, are becoming more and more relevant. These semi-supervised methods extract knowledge from the labeled data in a supervised way, and in an unsupervised way from the unlabeled data, thus reducing the labeling effort required in a fully supervised scenario, and obtaining notably better results than in an unsupervised scenario.

Fig. 2. Visual representation of the different variants of the image segmentation problem. From left to right, original image, semantic segmentation and instance segmentation examples.

To the best of our knowledge, there is only one survey that tries to address semi-supervised SS methods [157]. However, the aforementioned review does not cover certain aspects that we consider important:

- First of all, this previous study published in 2019, does not include the methods proposed in recent years, when the problem has gained more interest. Figure 1 shows a histogram where it is clear that the majority of publications and citations in this field are concentrated in the last two years, which are outside the previous survey.
- Secondly, it does not focus exclusively on the semi-supervised scenario, but also includes the methods proposed for a weakly supervised scenario. It does not go into sufficient depth for the semi-supervised field, which is currently a sufficiently broad scenario to be addressed exclusively.
- Finally, it does not include an experimental study that allows a fair comparison between methods and allows the reader to have a clear idea about the performance of each one of them.

Other surveys related to our target field, but not totally focused on it, have been proposed recently. Several of these works focus on the SSL paradigm. Some of them review methods based on DL [96, 149], and others carry out the review of methods from a general point of view [132, 136]. These studies take the image classification problem as the basic problem, so they do not cover the wide field of semi-supervised SS. On the other hand, there are other surveys that focus on the SS problem, but do not address the semi-supervised scenario or treat it very superficially [36, 47, 86].

The main contributions of this work are summarized as follows:

- We provide an updated taxonomy of semi-supervised SS methods as well as a description of them.
- We carry out an experimentation with a wide range of state-of-the-art semi-supervised segmentation methods on the most widely used datasets in the literature.
- A discussion is proposed on the results obtained, advantages and shortcomings of the current methods, challenges and future lines of work in this field.

The content of this article is organized as follows. The key concepts and background about the problem under discussion, as well as the existing datasets, are presented in Section 2. Then, in Section 3, the proposed taxonomy is described, followed by a subsection for each of the categories that form our taxonomy, where we go into detail about the different method proposals belonging to each category. A detailed description of the proposed experimentation as well as the results obtained are shown and discussed in Section 5, while a reflection on the main difficulties, challenges and future directions is presented in Section 6. Finally, Section 7 ends with the conclusions obtained.

2 Background

2.1 Problem formulation

The SSL paradigm is halfway between fully supervised learning and unsupervised learning, and it deals with data sets that are just partly annotated. Moreover, the ratio between the amount of labeled and unlabeled data is usually, in real-world problems, very unfavorable for the labeled part. Specifically, in semantic image segmentation, the imbalance between labeled and unlabeled data is often even more frequent and pronounced due to the difficulty of annotating an image at the pixel level [9].

In this context, we have a dataset $X = \{X_L, X_U\}$ where $X_L = \{(x_i, y_i)\}_{i=1}^{l}$ is the subset of labeled data and $X_U = \{x_i\}_{i=1}^{u}$ is the subset of unlabeled data, x_i is an input image and y_i its corresponding label map, l and u are the number of labeled and unlabeled data, respectively, and commonly $l << u$.

The purpose of semi-supervised semantic segmentation is defined as the development of a function $\mathcal{F} : \mathcal{M}_{n\times m\times 3} \rightarrow \mathcal{M}_{n\times m}(\mathbb{N} \cup \{0\})$. This is achieved through the use of a combination of loss functions applied both labeled $\mathcal{L}_L(\mathcal{F}(x_i), y_i)$ where $(x_i, y_i) \in X_L$ and unlabeled data $\mathcal{L}_U(\mathcal{F}(x_i), \mathcal{G}(x_i))$ where $x_i \in X_U$ and $\mathcal{G}$ being a

modification function, the identity or an adversarial or cooperative model. By leveraging both loss functions, information is extracted from both labeled and unlabeled data.

2.2 Classical approaches to SS

In this section we introduce those methods for SS that were proposed before the DL era (i.e., before 2012). Although our study focuses on deep segmentation models, it is relevant to analyze prior art. We consider three levels of supervision: fully-supervised, unsupervised, and semi-supervised.

It is important to note that the first methods proposed for image segmentation are essentially unsupervised. Among these methods we can find basic image processing and CV techniques that stand out for their simplicity and efficiency when applied. Among them we find methods such as image thresholding [110], region growing [131] and deformable models [129]. On the other hand, the application of unsupervised machine learning algorithms, such as clustering algorithms (e.g. k-means) [11] or graph-based models [32], has also been proposed.

Subsequently, proposals based on supervised machine learning algorithms started to emerge. For instance, Random Forest [108], SVMs [31] and conditional or Markov random fields [39, 89] were proposed and adapted to the image segmentation problem.

The semi-supervised setting was the last to be addressed. Some extensions of fully supervised methods were proposed to equip them with the ability to handle unlabeled data. To the best of our knowledge, the first method proposed specifically for semi-supervised segmentation was a mixed model based on a tree-structured patch-based approach and the random forest algorithm [4]. A model based on weighted graphs was also proposed for the semi-supervised segmentation of 3D surfaces [6]. Due to its fast semi-supervised classification and its interpretability, the random forest algorithm is used in other works to address the semi-supervised segmentation problem, as is the case of [81] where the authors propose its use on abdominal magnetic resonance. Finally, a last proposal prior to the emergence of DL proposes a method that incorporates Gaussian mixture models, random walk models and SVMs [99].

It is noteworthy that many of these methods require the manual setting of specific hyperparameters, such as threshold values, number of clusters, etc., which stands as a disadvantage compared to more modern models that will be discussed in the following sections.

2.3 Deep learning for SS

The performance of the semi-supervised methods largely depend on the good choice, fit and training of the supervised model on which it is based. In this section we present the background of these supervised segmentation models.

Initially, DL techniques, generally convolutional neural networks (CNN) [103], were proposed and applied to the problem of image classification, obtaining a leap in quality with respect to the traditional techniques that had been used until then. Due to the good results obtained, these techniques were extended to other areas of CV, trying to solve increasingly complex and fine-grained problems, such as object detection [104] and segmentation [14, 105].

The key idea underlying most DL models for SS is the fully convolutional neural networks (FCNN), proposed in [111]. In this work, the approach proposed by the authors consists in reusing well-known CNN (such as VGG [114], ResNet [49] or EfficientNet [126]) originally proposed and applied in image classification problems, adapting them to address the SS problem. This adaptation consists in replacing the final fully connected layers of these models by convolutional layers, thus obtaining as output feature maps instead of a vector of classification scores. Finally, the resulting feature maps are upsampled by using deconvolution operations [154] to obtain the final segmentation map. FCNN achieves a performance gains (20% improvement) in Pascal VOC [29], one of the main SS benchmarks.

The FCNN approach demonstrated that the problem of SS could be addressed through DL techniques, thus opening a new line of research that today is in a really advanced state, with many new methods that improve the original proposal of FCNN. The main difference between these methods lies in the way they upsample the output of the convolutional network to obtain the final segmentation map. For instance, encoder-decoder architectures (e.g. U-Net model [105]) chain a decoder to the CNN. Another well-known example is the DeepLab model [13–15] that uses atrous convolution to increase its range of vision and increase its capacity to capture contextual information.

In recent years, Transformer models [137], originally proposed in the field of language processing, have been adapted and applied to computer vision problems [28]. In particular, for the SS problem, several Transformer-based proposals [122, 146] have shown state-of-the-art results. One of the main characteristics that these novel architectures present is the possibility of capturing global information of the whole image when classifying each pixel. In contrast, CNNs can only pay attention to a local context, which sometimes makes it difficult to learn semantic relationships between objects located in different areas of the image.

2.4 Datasets

The availability of pixel-level annotated datasets to address the SS problem is not as high as those in other VC problems, such as image classification, mainly due to the difficulty of performing pixel-level annotations manually.

For instance, to annotate a single image from the Cityscapes dataset requires three hours of manual work [23]. However, due to the interest that this problem has raised in recent years, efforts have been made to annotate datasets to train segmentation models, and we currently have a wide range of these datasets of different types and domains. Table 1 shows a summary of some of the most commonly used datasets and their domain of application.

Table 1. Summary table of the most widely used datasets for SS classified according to the nature of the images.

Images content	Datasets
General images	PASCAL VOC 2012 [29], CIFAR-10/100 [64], ADE20K [160], PASCAL Context [90],
Street Views	CamVid [7], Cityscapes [23], Mapillary Vistas [92], KITTI [38],
Indoor environments	SUN RGB-D [118], ScanNet [24], Standford 2D [3], NYUD v2 [113],
Outdoor environments	INRIA-Graz-02 [82], Freiburg Forest [134], PASCAL SBD [48], Sift-Flow [74]
Human Pictures	Adobe's portrait [112], Helen [66], LIP [40], DeepFashion 2 [37],
Material images	MINC [5], UHCS [25], MetalDAM [80]
Satellite and aerial images	EuroSAT [51], FloodNet [101], xBD [45], AIRS [16], GID [130], iSAID [153]
Medical images	Medical Segmentation Decathlon [115] Drive [120], GlaS [116], CoNSeP [43],

Below we provide a description of the datasets used in our experimentation. On the one hand we include PASCAL VOC 2012 and Cityscapes, the two most widely used benchmark datasets in the literature, and on the other hand, MetalDAM, a real industrial use case.

- **PASCAL VOC 2012** [29]: The PASCAL VOC 2012 dataset[1] is the most widely used as benchmark in SS studies. It is composed of general situation and object-centered images with variable size. This dataset has 20 object classes and an additional background class. The partitions for training, validation and test consist of 1464, 1449 and 1456 images, respectively. An augmented version with 9118 extra images from the Segmentation Boundary Dataset (SBD) [48] is often used, bringing the training set to 10582 images with associated pixel-wise labeling.
- **Cityscapes** [23]: The Cityscapes dataset[2] is another of the most widely used datasets and, specifically, one of the most important ones for autonomous driving applications. This dataset is composed of a series of sequential street view images taken from a vehicle in different European cities, with size 2048 × 1024 and 19 classes. The official partitions for training, validation and test are composed of 2975, 500 and 1525 images respectively.
- **MetalDAM** [80]: The MetalDAM dataset[3] is introduced as a public benchmark for semantic segmentation of metallographic images, particularly those generated using additive manufacturing techniques. It comprises 42 steel images with resolutions of 1280 × 895 and 1024 × 703, segmented into 5 classes.

3 Semi-Supervised Semantic Segmentation Methods

In this section, a review and explanation of the techniques proposed for the semi-supervised segmentation problem and a taxonomy is carried out. First, the proposed taxonomy is presented, followed by a detailed section for each of its categories, with detailed explanations of the representative methods.

3.1 Related Works

Existing surveys of semi-supervised learning, tend to focus on classification or broad settings, rather than pixel-wise segmentation. Likewise, many semantic segmentation surveys address only fully supervised methods or mix different supervision regimes. Zhang et al. [157] review both semi and weakly supervised segmentation techniques, which dilutes the focus on pure semi-supervised approaches and is already out of date. Other reviews target narrow domains: for instance, Jiao et al. [57] survey semi-supervised methods for medical image segmentation, which do not necessarily generalize to generic vision tasks. In the same context, Han et al. [46] presents a survey on semi-supervised methods for medical image segmentation, arguing the low availability of labels in medical problems. Even more recent compendia like Ran et al. [102] restrict attention to a subclass of techniques, in this case pseudo-label methods, leaving out consistency-based, adversarial or contrastive approaches. We base our motivation in the following key points:

- Broad semi-supervised learning surveys as Van Engelen and Hoos[136] present a general survey on semi-supervised learning, but devoted to classification and not considering the semantic segmentation task.
- Zhang et al. [157] review semi and weakly supervised semantic segmentation methods, however this mix of settings implies newer pure semi-supervised methods are not treated, as the review dates from 2020.
- Surveys like Jiao et al. [57] focus on medical image segmentation, which limits their applicability. They do not address challenges in general RGB image segmentation under the semi-supervised lens.
- Newer papers as Ran et al. [102] offer a survey only of pseudo-label strategies, omitting entire classes of methods.
- General segmentation surveys do not cover the use of unlabeled data at all, and thus ignore the semi-supervised regime entirely.

[1] http://host.robots.ox.ac.uk/pascal/VOC/
[2] https://www.cityscapes-dataset.com/
[3] https://dasci.es/transferencia/open-data/metal-dam/

Crucially, none of the above fully incorporates new semi-supervised learning segmentation techniques introduced between 2021 and 2025. For instance, Dynamic Mutual Training (DMT), proposed in 2020, demonstrated state-of-the-art performance in segmentation tasks [33]. However, its adaptive re-weighting mechanism (along with subsequent extensions) has not been addressed in earlier reviews. Likewise, hybrid frameworks that combine multiple semi-supervised learning strategies, or consistency-based methods with network perturbations (injecting noise or dropout into the model) have emerged recently with impressive results [78]. These include approaches that enforce prediction consistency under diverse input, feature, or model perturbations. Such developments, along with other modern ideas like multi-scale contrastive learning, generative modeling for segmentation, and large pre-training, fall outside the scope of prior surveys. This gap motivates the present survey: we synthesize all recent semi-supervised segmentation methods (including DMT, hybrid models, perturbation-driven consistency, etc) in one place, highlighting new trends that no previous review has fully addressed to the best of our knowledge.

3.2 Methodology

To construct a comprehensive and representative taxonomy of recent advances in semi-supervised semantic segmentation, we conducted a systematic literature review across three major repositories: Scopus, IEEE Xplore, and arXiv. The search was guided by a set of carefully selected keywords, including "semi-supervised semantic segmentation," "SSL segmentation," and related terms, aiming to retrieve a broad yet focused collection of relevant works. We restricted our inclusion to methods based on deep learning architectures, explicitly discarding traditional machine learning approaches or heuristic-based pipelines. In order to ensure the reproducibility and practical relevance of the selected methods, we only considered publications that either provided open-source implementations or reported quantitative results on standard benchmarks such as PASCAL VOC or Cityscapes.

Furthermore, we excluded methods falling under weak supervision, domain adaptation, or few-shot learning paradigms, unless these works addressed exclusively a semi-supervised setting and performed pixel-level segmentation. This ensured a clear focus on methods designed specifically to exploit unlabeled data in a semi-supervised context. After filtering an initial set of several hundred publications, we manually curated a final selection of 50 representative and up-to-date models, which span a wide range of strategies and have had a significant impact on the field. This curated corpus served as the basis for the taxonomic structure presented in this work.

We established a set of consistent criteria to guide the classification process. Each model was examined according to the underlying mechanism used to leverage unlabeled data, which typically fell into one or more of the following categories: adversarial learning, consistency regularization, pseudo-labeling, contrastive learning, or hybrid strategies combining these approaches. These distinctions reflect the conceptual foundations that define how each method exploits unlabeled data to enhance segmentation performance under limited supervision.

Beyond the methodological paradigm, we further analyzed the architectural components of each model to capture structural and operational similarities. Key factors included the presence of generator-discriminator architectures (typical in adversarial settings), teacher-student frameworks (often used in consistency-based methods), specialized data augmentation techniques such as ClassMix or CutMix, and modules enabling contrastive memory mechanisms. Additionally, we categorized each method based on the nature of the perturbation or supervisory signal it applied—whether at the input level, within the feature space, across network components, or through generated pseudo-labels. This multifaceted analysis enabled the construction of a hierarchical categorization, where models were grouped according to shared mechanisms and architectural patterns, forming the foundation of the taxonomy proposed in this survey.

3.3 Taxonomy

According to the nature and main characteristics of the existing methods in the semi-supervised SS literature, we propose a taxonomy that classifies these methods into five categories. This taxonomy is represented graphically by the dendogram displayed in Figure 3, and a list of all existing methods in each category is provided in Table 2.

- The first category includes those methods that adopt a GAN-like structure and adversarial training between two networks, one acting as a generator and the other as a discriminator (Section 3.4).
- The next category corresponds to consistency regularization methods. These methods include a regularization term in the loss function to minimize the differences between different predictions of the same image, which are obtained by applying perturbations to the images or to the models involved (Section 3.5).
- Another category comprises methods that are based on pseudo-labeling of unlabeled data. In general terms, these methods rely on predictions previously made on the unlabeled data with a model trained on the labeled data to obtain pseudo-labels. By so they are able to include the unlabeled data in the training process (Section 3.6).
- The fourth category includes methods based on contrastive learning. This learning paradigm groups similar elements and separates them from dissimilar elements in a certain representation space, often different from the output space of the models (Section 3.7).
- Finally, we group in a fifth category those methods that present characteristic elements of several of the previously exposed categories. We may find hybrid methods between consistency regularization, pseudo-labeling and contrastive learning (Section 3.8).

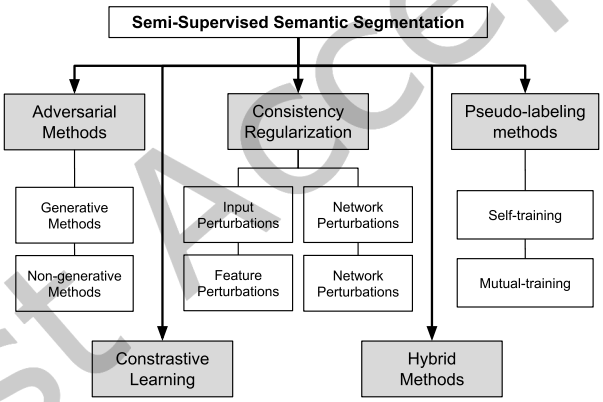


Fig. 3. Dendrogram showing the taxonomy proposed in this paper. References to the methods belonging to each category are located next to each leaf node.

The proposed taxonomy shares certain similarities with other taxonomies of semi-supervised learning introduced in different contexts (e.g., [149], which focuses on the classification problem), although it diverges in several aspects. We introduce a category dedicated to constructive methods, given their frequent use in the field of segmentation.

Within our taxonomy, there are methods that have been adapted from the classification problem. Examples of these methods are the Mean Teacher [127] method and the basic Self-Training method [148], which will be discussed in depth in the category description section. These methods outline a semi-supervised learning strategy applicable to both segmentation and classification tasks, with the only requirement being the use of task-specific base models.

However, most of the methods included in our study are explicitly designed and proposed to address segmentation tasks, adapted to the particularities presented by this task. Some of the methods that incorporate data augmentation techniques are explicitly designed with the pixel-level information available in this task in mind. For example, the ClassMix [94] method takes advantage of this information to mix images. In the case of generative models, dealing with these labels in the form of segmentation maps poses an added challenge that often makes the direct application of generative approaches difficult, as documented in [69]. However, in many cases, having pixel-level annotations is an advantage that some methods exploit. For example, numerous methods (e.g., pseudo-labeling methods) employ the pixel-level confidence map generated by base segmentation models to guide semi-supervised retraining, focusing on areas where the model exhibits higher confidence and discarding regions where model certainty is less evident.

In the following sections, we will delve into the specific characteristics of the methods that comprise the state of the art of semi-supervised segmentation.

3.4 Adversarial methods

Generative adversarial networks (GANs) [42] have become a very popular framework due to the good performance they have demonstrated in a multitude of problems such as image generation [100], object detection [140] or SS [79], among many others. A typical GAN framework consists of two networks, generator and discriminator. The purpose of the generator is to learn the distribution of the target data, thus allowing the generation of synthetic images from random noise. The purpose of the discriminator is to distinguish between real images (belonging to the real distribution) and fake images (generated by the generator). The training process of these networks is carried out in an adversarial way. The generator tries to confuse the discriminator, generating images increasingly similar to the target distribution, and the discriminator attempts to increase its ability to distinguish between real and fake images. This adversary training process is formally defined below:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim X}[log(D(x))] + \mathbb{E}_{z\sim p_z(z)}[log(1-D(G(z)))] \tag{1}$$

The Equation 1 represents the min max game played by the discriminator D and the generator G. The purpose of the first term is to maximize the accuracy obtained by D, while the second term attempts to increase the quality of the images generated by G, from random noise z.

Methods based on adversarial training for semi-supervised SS are divided into two subcategories in the proposed taxonomy. The key aspect that differentiates these methods is the inclusion or not of a generative model in the training process. In this sense, one of the subcategories groups those models that employ a generative model [69, 119], thus generating new synthetic images that can be used as additional input examples for the segmentation model (Figure 4). On the other hand, the other subcategory groups those methods that do not include a generative model in their GAN-like structure [27, 56, 58, 62, 73, 84, 87, 147, 156]. In these cases, a segmentation network assumes the role of generator, and the objective of the discriminator is differentiating those segmentation maps generated from the segmentation network from the real segmentation maps (Figure 5). Generative methods, however, face certain limitations[121] including vulnerabilities to attacks targeting the generator, the discriminator, or data poisoning, which can undermine the learning paradigm. Below we present and explain the different methods proposed in each of the two subcategories.

3.4.1 Generative methods. The methods based on GANs, in general, were the first approaches of DL techniques to the problem of semi-supervised SS. Previously, only weakly supervised approaches had been proposed, which do not take advantage of unlabeled data.

Table 2. List of existing semi-supervised SS methods in the literature, classified according to the defined taxonomy.

Method	Category	Subcategory	Year
[69]	Adversarial methods	Generative	2021
[119]			2017
SemiRoadExNet[12]			2023
[58]		Non-generative	2021
[147]			2021
[27]			2021
[84]			2020
(GCT) [62]			2020
[155]			2020
(S4GAN) [87]			2019
[73]			2019
[56]			2018
[93]			2023
(ComplexMix) [21]	Consistency regularization	Input perturbations	2021
[44]			2021
(ClassMix) [94]			2021
(CutMix) [35]			2020
[71]			2020
[63]			2020
[123]			2023
(CCT) [97]		Feature perturbations	2020
[2]		Network perturbations	2022
(CPS) [20]			2020
[98]			2020
[142]		Combined perturbations	2022
[78]			2021
[85]			2022
(ST++) [148]	Pseudo-labeling	Self-training	2021
(GIST & RIST) [128]			2021
[70]			2021
[151]			2021
[50]			2021
[164]			2020
[22]			2020
[10]			2024
(DMT) [33]		Mutual-training	2022
[161]			2022
(ReCo) [76]	Contrastive learning	-	2021
[1]			2021
[72]			2023
(CTT) [144]	Hybrid	-	2022
[8]			2022
(AEL) [55]			2021
(GuidedMix-Net) [133]			2021
(CAC) [65]			2021
[159]			2021
[162]			2021
(PseudoSeg) [165]			2020
[61]			2020
SSDA(DS+US)[53]			2023
SemiVL[54]			2024
segWCD[143]			2024

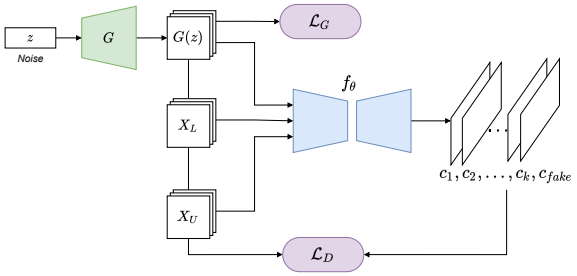


Fig. 4. Generative adversarial method structure for semi-supervised segmentation. The generator G receives random noise as input and generates new synthetic images. Then, the segmentation network f_θ receives both synthetic ($G(z)$) and real (X_L, X_U) images as input and classifies each pixel into its corresponding class $c_1, c_2, ..., c_k$ or into an additional fake class c_{fake} which indicates that it is a synthetic pixel. $\mathcal{L}_D$ and $\mathcal{L}_G$ are the discriminator and generator loss functions, respectively.

In particular, the first method [119] proposed to address the segmentation problem in a semi-supervised way, without requiring weak labels, consists of a GAN framework adapted for the segmentation problem. This framework aims, on the one hand, to handle and extract knowledge from a large amount of unlabeled data, and on the other hand, to increase the number of training examples available through the synthetic generation of images. Specifically, this method includes a generative network that approximates the distribution of the target images, thus achieving the ability to generate new training examples. A segmentation network assumes the role of discriminator and segments the images received as input, both real and synthetic. This network classifies each pixel with its corresponding class, or with an extra fake class, which indicates that this pixel or region of the image has been generated by the generator. This type of architecture can be seen represented in Figure 4. This approach adapts the loss function proposed for the original GAN to the SS problem. Both the loss function used to optimize the generator ($\mathcal{L}_G$) and the segmentation model that acts as a discriminator ($\mathcal{L}_D$) are shown below:

$$\mathcal{L}_D = -\mathbb{E}_{x\sim X}[log(f_\theta(x))] - \mathbb{E}_{z\sim p_z(z)}[log(1 - f_\theta(G(z)))] + \gamma\mathbb{E}_{x,y\sim X_L}[CE(y, f_\theta(x))] \quad (2)$$

$$\mathcal{L}_G = \mathbb{E}_{z\sim p_z(z)}[log(1 - f_\theta(G(z)))] \quad (3)$$

The discriminator loss function $\mathcal{L}_D$ (Equation 2) is composed of three terms. The first term penalizes the model when it labels a real sample as fake. The second term penalizes the model when it labels a fake sample as real. The last term is the supervised component, which forces the correct classification of each pixel of the labeled set in its corresponding class. γ is the weight of the supervised component in the training process. The generator loss function $\mathcal{L}_G$ (Equation 3), seeks to increase the quality of the generated images by penalizing G when f_θ detects synthetic images.

Another generative method [69] has been proposed recently for semi-supervised SS, due to the recent success of StyleGAN [59]. The proposed model extends the StyleGAN model, adding a label synthesis branch, and attempts to capture the joint distribution of images and labels, gaining the ability to generate new image-label pairs. However, due to the high complexity of this generative problem, the authors themselves state that this approach is still far from being able to deal with the segmentation of natural and generic images, and limit its success cases to very specific domains such as skin lesions and facial parts segmentation.

3.4.2 Non-generative methods. On the other hand, we grouped those methods that use adversarial training and have a similar structure to GAN, but do not include a generative model. All the methods that we group under this subcategory share the characteristic of replacing the typical generative network of the classical GAN

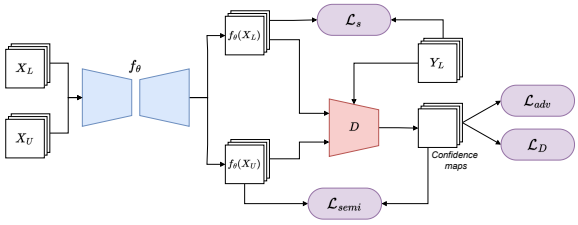


Fig. 5. Non-generative adversarial method structure for semi-supervised segmentation. Segmentation network f_θ acts as generator. Supervised cross-entropy loss function ($\mathcal{L}_s$) is used to train f_θ in a supervised way. Discriminator D is trained to distinguish between real and predicted (by f_θ) segmentation maps. The output of D (*confidence maps*) is used to perform the semi-supervised learning ($\mathcal{L}_{semi}$) with unlabeled data, and also is used for discriminator and adversarial loss functions ($\mathcal{L}_D$ and $\mathcal{L}_{adv}$).

by a segmentation network. Its output is directed towards a discriminator that distinguishes between the real segmentation maps, and those generated by the segmentation network.

This GAN-like architecture for SS was originally proposed in [79], and adapted for a semi-supervised scenario in [56]. The authors present a fully convolutional discriminator that receives both segmentation maps (the one coming from the ground truth and the one predicted by the segmentation model, in this case DeepLabV2 [13]). The discriminative network is adversarially trained, together with the segmentation model, to distinguish real label maps from predicted ones. In this sense, it produces a probability map as output, of the same dimension as the input image, where it represents, for each pixel, the confidence of being a real example or a prediction made by the segmentation network. In this way, this confidence map indicates the quality of the segmentation in a certain area, so that the confidence map of the unlabeled images can be used to detect those areas where the predicted labels have enough quality to be used in the training process of the segmentation model. This structure is represented in Figure 5. The formulation of the loss functions involved in these methods is presented below:

$$\mathcal{L}_D = -\mathbb{E}_{y\sim X_L}[log(D(y))] - \mathbb{E}_{x\sim X}[log(1 - D(f_\theta(x)))] \tag{4}$$

$$\mathcal{L}_{seg} = \mathcal{L}_{sup} + \lambda_{adv}\mathcal{L}_{adv} + \lambda_{semi}\mathcal{L}_{semi} \tag{5}$$

$$\mathcal{L}_{sup} = \mathbb{E}_{x,y\sim X_L}[CE(y, f_\theta(x))] \tag{6}$$

$$\mathcal{L}_{adv} = -\mathbb{E}_{x\sim X}[log(D(f_\theta(x)))] \tag{7}$$

$$\mathcal{L}_{semi} = -\mathbb{E}_{x\sim X_U}[I(D(f_\theta(x)) > \mathcal{T}) \cdot \hat{y} \cdot log(f_\theta(x))] \tag{8}$$

The discriminator loss function $\mathcal{L}_D$ (Equation 4) is composed of two terms, each of which forces the discriminator D to detect the segmentation maps coming from the ground truth and those generated by the segmentation network f_θ. The segmentation network loss function $\mathcal{L}_{seg}$ (Equation 5) is composed of three terms. The first is the supervised component $\mathcal{L}_{sup}$ (Equation 6), formed by the cross-entropy loss function. The second is the adversarial component $\mathcal{L}_{adv}$ (Equation 7) that penalizes the cases in which D detects segmentation maps generated by the segmentation network. The third term $\mathcal{L}_{semi}$ (Equation 8) allows to take into account the unlabeled images whose segmentation exceeds a confidence threshold $\mathcal{T}$ by D. λ_{adv} and λ_{semi} are parameters that weight the use of their respective terms.

Based on the previous approach, other alternatives have been proposed to improve the structure of the original method in different ways. S4GAN [87] proposes the use of a simpler discriminator that generates an output for the entire segmentation map rather than for each pixel. It also includes an additional processing branch where a classifier is trained. It is used to filter the segmentation maps obtained, removing those labels that are false positives in view of the classifier. Confrontation Network [27] method also incorporates image-level discriminator

and improves the generator loss function by adding a variance regularization term. Other approaches [73, 147] propose the use of two discriminators, one at the image level and the other at the pixel level. Both are used together in order to increase the accuracy in the definition of confidence areas in the images.

Error-Correcting Supervision (ECS) [84] and Guided Collaborative Training (GCT) [62] are based on a collaborative strategy, very close to the original adversary strategy. These approaches introduce a new network which assumes the role of discriminator, called correction network in the case of ECS and flaw detector in GCT. These approaches provide, in addition to a confidence map at the pixel level, a correction for those areas where confidence is low.

Other adversarial approaches incorporate attention modules with the objective of modeling long-range semantic dependencies. This is the case in [156] which also incorporates spectral normalization to reduce the instability in the training process. Another approach [58] proposes the use of attention modules in combination with sparse representation module that helps the segmentation model to emphasize the edges and locations of objects.

3.5 Consistency regularization

SSL makes some assumptions without which successful knowledge extraction from unlabeled data would not be possible. Specifically, consistency regularization methods are based on the assumption of smoothness [9]. This assumption says that, for two nearby points in the input space, their labels must be the same. In other words, a robust model should obtain similar predictions for both a point and a locally modified version of it.

In this sense, SSL methods based on consistency regularization take advantage of unlabeled data by applying perturbations on them, and training a model that is not affected by these perturbations. This is achieved by adding a regularization term to the loss function that measures the distance between the original and perturbed predictions. The following is the formal definition of the described loss function:

$$\mathcal{L} = \mathcal{L}_{sup} + \lambda \mathcal{L}_{cons} \tag{9}$$

$$\mathcal{L}_{sup} = \mathbb{E}_{x,y \sim X_L}[CE(y, S(x))] \tag{10}$$

$$\mathcal{L}_{cons} = \mathbb{E}_{x \sim X_U}[R(f_\theta(x), f_{\theta'}(x))] \tag{11}$$

where $\mathcal{L}_{sup}$ is the supervised cross-entropy (CE) loss function and $\mathcal{L}_{cons}$ is the unsupervised regularization term. R is a function that measure the distance between two predictions obtained from the student network f_θ and teacher network $f_{\theta'}$. λ is used to weight the relevance of $\mathcal{L}_{cons}$. This learning paradigm is notably affected by a decline in performance when regions of low data density occur near class boundaries in the input image. This issue arises from smooth transitions between classes, which hinder the accurate delineation of class boundaries [34].

The basic method on which all other approaches are based is Mean Teacher [127]. It forces consistency between the predictions of a student network and a teacher network. The weights of the teacher network are calculated by an exponential moving average (EMA) of the weights of the student network. Figure 6 shows a graphical representation of the structure of this method.

The main difference between methods based on consistency regularization for semi-supervised SS lies in the way they incorporate perturbations to the data. Based on this, we can group these methods into four subcategories. On the one hand, the methods based on input perturbations [21, 35, 44, 63, 71, 94]. These methods apply perturbations directly to the input images using data augmentation techniques. They force the model to predict the same label for both the original image and the augmented image (Figure 7). Second, the methods based on feature perturbations, which incorporate perturbations internally in the segmentation network, thus obtaining modified features [97] (Figure 8). In third place, the methods based on network perturbations, which obtain perturbed predictions by using different networks, for instance, networks with different starting weights

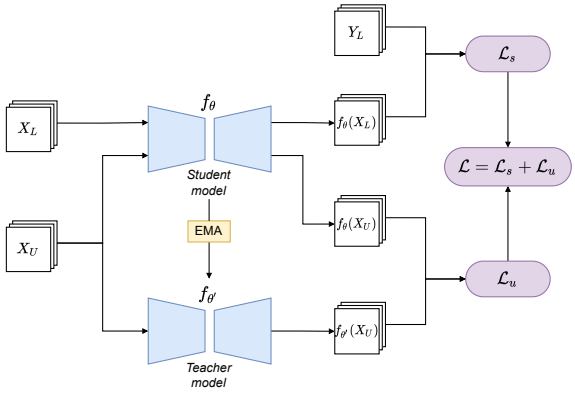


Fig. 6. Mean Teacher [127] method structure. $\mathcal{L}_s$ is used to train the student model f_θ in a supervised way. $\mathcal{L}_u$ is a regularization term that forces consistency between f_θ and teacher model $f_{\theta'}$ predictions.

[2, 20, 98] (Figure 9). Finally, we can identify a last subcategory that combines some of the three previous types of perturbations [78, 142].

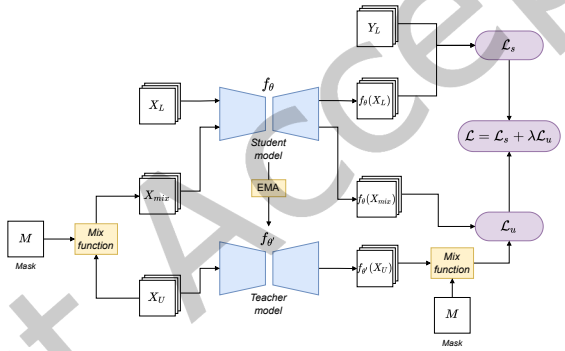


Fig. 7. Input perturbations based consistency regularization method structure for semi-supervised segmentation. It presents a Mean Teacher base structure (Figure 6), and incorporates input perturbations in unlabeled data by means of the *mix function* and M mask.

3.5.1 Input perturbations. In a first subcategory we group those consistency regularization methods that apply perturbations directly to the unlabeled input images using data augmentation techniques. Then, these methods train a segmentation model that is not sensitive to these input perturbations, and predicts segmentation maps that are as similar as possible for both the original images and their augmented versions. The key aspect that differentiates these methods is the way they perform modifications to the data. We can find in the literature different proposals for data augmentation techniques that have been applied to the semi-supervised SS problem. The consistency term incorporated in these data augmentation-based methods is defined as follows:

$$\mathcal{L}_{cons} = \mathbb{E}_{x_a,x_b\sim X_U}[R(mix(f_{\theta'}(x_a), f_{\theta'}(x_b), M), f_\theta(mix(x_a, x_b, M)))] \tag{12}$$

where *mix* is a mixing function that receives as input two images x_a, x_b (or segmentation maps $f_{\theta'}(x_a), f_{\theta'}(x_b)$) and returns a combination of them. This combination is done by means of a predefined mask M. Below we detail the different data augmentation techniques for semi-supervised SS proposed in the literature.

CutOut and CutMix techniques are applied to SS in [35]. Previously, these techniques have been applied in image classification [26, 152]. These techniques use a rectangular mask over the images. CutOut discards the rectangular section marked by the mask in the training process. Then, the consistency between the predictions of the original image and the modified image is forced by the regularization term. On the other hand, CutMix combines two images using a rectangular mask, obtaining a new image where the sections marked by the mask belong to one of the original images, and the rest of the sections belong to the other image (the inverse image is also obtained). Another approach [63] extends the previous method by adding a new term to the loss function called consistency structured loss that incorporates the concept of pair-wise knowledge distillation [77].

ClassMix [94] is proposed and designed specifically for the SS problem. This technique differs from the previous CutMix technique in the form of the mask that is applied to mix images. In this case, the sections marked by the mask coincide with areas belonging to the same class in the image, so that sections completely belonging to one class are copied into another image, thus generating the new augmented images. The difference between original and augmented predictions is calculated in the same way as the previous technique using the regularization term. ComplexMix [21] proposes the combined use of the previous data augmentation techniques, CutMix and ClassMix.

Besides these types of methods that propose a specific data augmentation technique for segmentation, other approaches [71] use classical data augmentation techniques (e. g. cropping, color jittering or flipping) to obtain the perturbed versions of the original images. Focused on efficiency, a method [44] is proposed that performs photometric and geometric perturbations only in the teacher model.

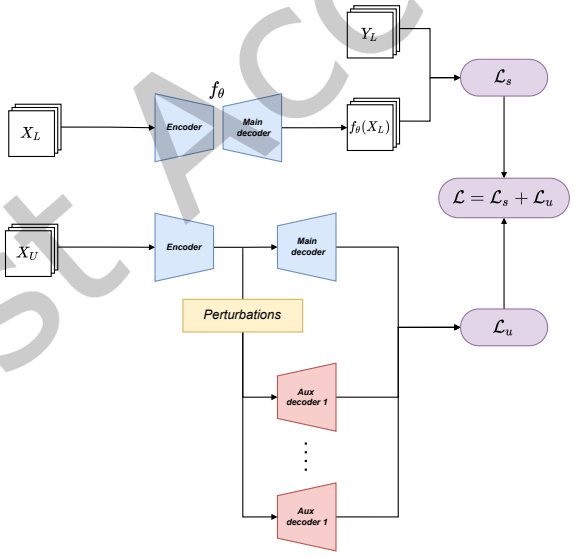


Fig. 8. Feature perturbations based consistency regularization method structure for semi-supervised segmentation. It presents a Mean Teacher base structure (Figure 6), and incorporates perturbations in an internal representation of the segmentation network, obtaining different outputs from auxiliary decoders. These outputs are forced to be consistent through the regularization term $\mathcal{L}_u$.

3.5.2 Feature perturbations. The second way to introduce perturbations in the training process consists in perturbing the internal features of the segmentation network. Cross-Consistency Training (CCT) [97] is proposed to address the semi-supervised SS problem following this idea. The architecture presented extends a supervised segmentation model with an encoder-decoder structure (e.g. DeepLabV3+ [15]) with some auxiliary decoders. First, a supervised training is carried out with the available labeled data, using the main decoder. Next, to take advantage of the unlabeled data, the encoder output is perturbed in different ways, resulting in different versions of the same features, which are directed to different auxiliary decoders. Finally, consistency between the outputs of the auxiliary decoders is enforced, favoring similar predictions for different perturbed versions of the encoder output features. The consistency term incorporated in these feature perturbation-based methods is defined as follows:

$$\mathcal{L}_{cons} = \mathbb{E}_{x \sim X_U} [\frac{1}{k} \sum_{k=1}^{K} R(h(x), h^k(x))] \tag{13}$$

where h is the main decoder, h^k is the k-th auxiliary decoder, and K is the number of auxiliary decoders.

3.5.3 Network perturbations. Another way of introducing perturbations in the training process is to use different segmentation networks. The differences between the networks constitute the perturbations in the resulting predictions. This is the case of the Cross Pseudo Supervision (CPS) method [20], which follows a training process similar to Mean Teacher. In this case the training of the two networks involved is carried out in a parallel and independent way, instead of updating one according to the EMA of the other. In addition, although both networks share the same architecture, they are initialized with different random weights, thus increasing the difference between them. An extension of the above method by including three networks in the training process can be seen in [2]. Another approach [98] emphasizes the importance of enforcing diversity across networks and proposes the use of adversarial samples and re-sampling strategy to train the models on different sets.

As in the other consistency regularization methods, the consistency between the predictions of the networks is enforced by a regularization term included in the loss function. This regularization term is defined as follows (for the case where two networks are used):

$$\mathcal{L}_{cons} = \mathbb{E}_{x \sim X_U} [R(f_\theta(x), g_\phi(x))] \tag{14}$$

where f_θ and g_ϕ are different networks trained independently.

3.5.4 Combined perturbations. Finally, a last subcategory includes those methods that jointly apply several of the different types of perturbations described above.

A method that proposes the combination of input, feature, and network perturbations is presented in [78]. This method emphasizes the fact that a greater variety and strength of perturbations may cause more problems if the predictions are not sufficiently accurate. In this sense, to ensure accurate predictions for unlabeled images, this method extends the Mean Teacher method by adding a confidence-weighted cross-entropy loss function, instead of the mean square error (MSE) used by the classic Mean Teacher method. In addition, it also proposes a new way of performing feature perturbations by means of virtual adversarial training [88].

The combination of input perturbations, specifically the CutMix technique, and feature perturbations is proposed in [142]. Instead of adding different auxiliary decoders, as in CCT [97] , this method proposes the application of perturbations directly on the features, while the decoders share the weights.

3.6 Pseudo-labeling methods

Pseudo-labeling methods, also known as bootstrapping [96], wrapper [136] or self-labeled [132] methods, are among the most widely known and the first semi-supervised methods to appear [145]. This type of method consists of an intuitive approach to extend existing supervised models to a semi-supervised scenario, allowing

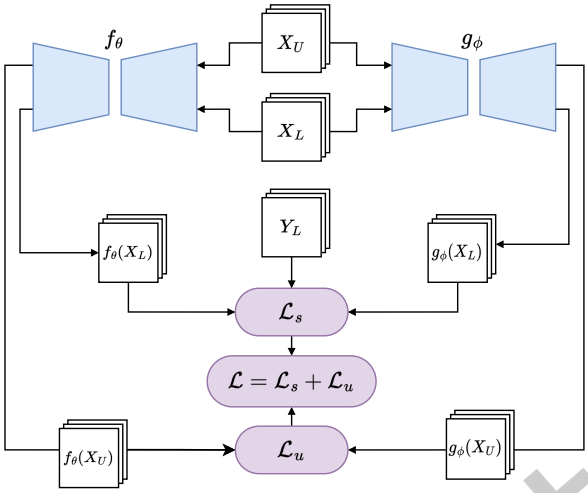


Fig. 9. Network perturbations based consistency regularization method structure for semi-supervised segmentation. It presents a Mean Teacher base structure (see Figure 6) and changes the teacher model to a second segmentation network g_ϕ that is trained independently. The outputs of both networks are forced to be consistent by the regularization term $\mathcal{L}_u$.

them to handle unlabeled data. The idea behind pseudo-labeling methods is simple: generate pseudo-labels of the unlabeled images from the predictions made by a model previously trained on the labeled data. Then, extend the labeled dataset with these new pairs of images and pseudo-labels, and train a new model on this new dataset. This idea is formalized with the loss function:

$$\mathcal{L} = \mathbb{E}_{x,y\sim X_L}[CE(y, f_\theta(x))] + \lambda \mathbb{E}_{x\sim X_U}[CE(\hat{y}, f_\theta(x))] \tag{15}$$

where $\hat{y}$ is the pseudo-label for image x, generated from the predicted probabilities with the segmentation model f_θ, in many cases by one-hot encoding, and λ is a parameter that weights the unsupervised part of the loss function.

Based on the differences between models in the training process and the way pseudo-labels are generated, in our taxonomy we differentiate between two types of pseudo-labeling methods. The first are self-training methods [22, 50, 70, 128, 148, 151, 164], based on one supervised base model and representing the simplest form of pseudo-labeling, where pseudo-labels are generated from their own high-confidence predictions (Figure 10). Secondly, mutual-training methods [33, 161], which involve multiple models with explicit differences such as different initialization weights or training on different views of the data. Each of the models are retrained with the unlabeled images and the corresponding pseudo-labels generated by other models involved in the process (Figure 11). This model operates under the assumption of a well-distributed labeled dataset across all classes. In cases of significant class imbalance, performance would degrade because the tag generator would lack sufficient information about underrepresented classes [50].

3.6.1 Self-training. Self-training methods are the simplest pseudo-labeling and semi-supervised methods, first proposed in [150], thoroughly reviewed in [132] and applied for the first time with deep neural networks in [68]. These methods consist in retraining a base supervised model by feeding back the training set with its own predictions. The typical self-training process consists of the following steps:

(1) The supervised model is trained on the available labeled data.

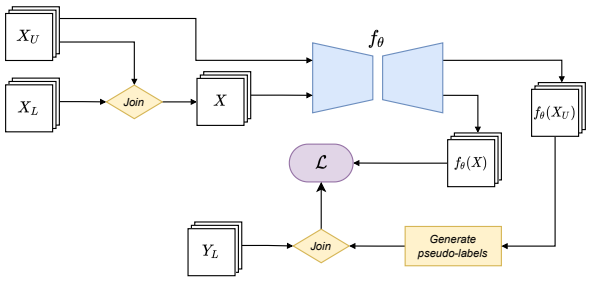


Fig. 10. Self-training method structure for semi-supervised segmentation. Firstly, pseudo-labels are generated for the unlabeled images using the segmentation network f_θ (usually pre-trained with labeled images). Then, pseudo-labels are joined to the ground truth and the loss function $\mathcal{L}$ is computed in a supervised way for all images.

(2) Predictions are obtained from the unlabeled data using the previously trained model. Those predictions with a confidence level higher than a predefined threshold become pseudo-labels for unlabeled data and are included in the labeled data set.
(3) The supervised model is retrained with this new data set composed of the labeled and the pseudo-labeled data.

This process can be repeated in an iterative way, obtaining new pseudo-labels with the model resulting from step 3, refining the quality of the pseudo-labels at each iteration, until no prediction exceeds the confidence threshold necessary to be treated as a pseudo-label.

The methods grouped in this subsection are based on this training process applied to the SS problem, each of them contributing some variant to the original algorithm that improves the learning capacity. For instance, the method proposed in [164] extends the original self-training process with a centroid sampling technique. The purpose is to solve the problem of class imbalance in the pseudo-labels.

Other proposals consist of adding some auxiliary network to the self-training process. For example, in [70] the authors extend the self-training process by adding a residual network. This network is trained with the labeled images, and is subsequently used to refine the pseudo-labels obtained by the segmentation model. The pseudo-labels predicted by a model may have a substantially different label space than the ground truth. This can be a problem when training a model with both label inputs, since it can lead to different gradient directions, resulting in a chaotic back-propagation process. A possible solution proposed in [22] consists in the use of a segmentation model that shares the encoder (i.e. ResNet101) and incorporates two different decoders, one for each label space.

The integration of data augmentation techniques within the self-training process has also been proposed in different approaches. The ST++ [148] method applies data augmentation techniques on the unlabeled images during the self-training process. This is combined with a selective stage in which, on each iteration of the self-training process, those images with reliable pseudo-labels are prioritized, and those images that present a higher probability of suffering from errors in the pseudo-labels are discarded.

Nevertheless, the application of data augmentation may alter the distribution of the mean and variance in the batch normalization. To solve this problem, the use of distribution-specific batch normalization is proposed in [151]. Additionally, this method also integrates a self-correction loss function which performs a dynamic re-weighting based on confidence, in order to avoid over-fitting noisy labels and under-learning of the most difficult classes.

A common issue faced by this type of methods is the distribution mismatch between ground truth and pseudo-labels, where the latter are often biased towards the majority classes. In order to obtain unbiased pseudo-labels, a

strategy of distribution alignment and random sampling with class-wise thresholding is proposed in [50], also in combination with data augmentation techniques.

Another proposal focuses on the difficulty of defining an optimal ratio between the actual labeled data and the pseudo-labeled data to be used in the self-training process. In this sense, two strategies are proposed to approach this optimal value during the iterative retraining process, one of them is based on a randomized search (RIST) and the other one employs a greedy algorithm (GIST) [128].

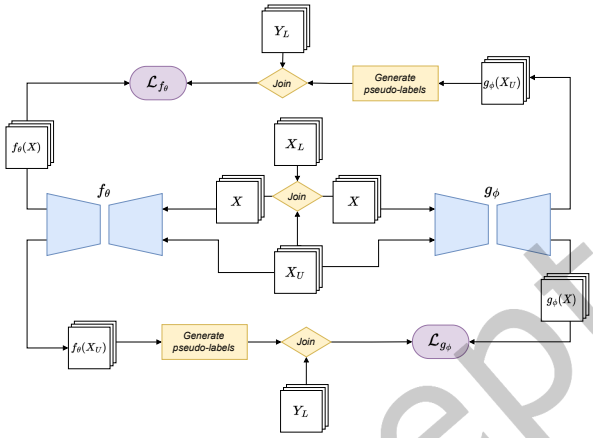


Fig. 11. Mutual-training method structure for semi-supervised segmentation. This approach extends the classical self-training (see Figure 10) with an additional segmentation network g_ϕ. The pseudo-labels used to retrain each of the networks are computed with the other network.

3.6.2 Mutual-training. One of the main disadvantages of previously described self-training methods is the absence of a mechanism for detecting their own errors. Instead of learning from their own predictions, mutual learning [158] methods extend self-training methods and involve multiple learning models, each of which train with the pseudo-labels generated by other models. The diversity present among the participating models is one of the key aspects for the proper performance of this type of methods [139]. That is why the different existing proposals try to explicitly induce differences between the base supervised models that compose the co-training method, for instance, by initializing such models with different pre-trained weights or by training each of the models with different views or subsets of the training set. In other studies, similar methods have been categorized as disagreement-based [149, 163], since they rely on exploiting the predictive differences between the models involved, multi-view training [96] or co-training [136].

DMT is a mutual learning approach adapted to semi-supervised scenario and SS problem proposed to take advantage of the disagreement between models as a way to detect errors in the generated pseudo-labels. This method takes these differences into account by means of a loss function that is dynamically re-weighted during training based on the discrepancies between two different models, which are trained independently, using the pseudo-labels generated by the other model. In this sense, a greater disagreement in a specific pixel indicates a greater probability of error, so it is weighted with a low value in the loss function, and has less influence on the training than other pixels or areas of the image where the discrepancy between models is smaller [33].

Another approach consists in extending the previous method (DMT) with a pseudo-label enhancement strategy [161]. This publication focuses on the problem of catastrophic forgetting. This problem points out the difficulty that models have to maintain the acquired knowledge when they receive inputs with some variants. This could

be the case of pseudo-labels. In order to maintain the acquired knowledge during the whole training process, and to avoid that the models suffer a bias towards the last classes learned, the authors propose a strategy that takes into account the pseudo-labels generated in previous stages to refine the current ones.

3.7 Contrastive learning

Contrastive learning focuses on high-level features to differentiate between classes in the absence of ground truth. In other words, these types of methods group similar samples and move them away from different samples in feature space. In many contrastive learning methods, the target sample to be compared is called the *query*, while the similar and dissimilar samples are called the *positive* and *negative* keys, respectively. Due to the lack of annotations in the data, samples considered similar in the training process are augmented versions of the same sample, while the rest of the data are considered different samples. Specifically, in the most relevant contrastive methods, pairs of augmented images are commonly obtained in different ways. Some of them apply data augmentation techniques (e. g. cropping, color jittering or flipping) as in the SimCLR method [17, 18]. Other methods divide the image into different overlaying sub-patches and considering these patches independent images as in the CPC method [135].

Due to the success of this type of methods, even outperforming its supervised counterpart in some specific problems, such as Pascal VOC object detection [30], in recent years a series of contrastive learning methods specifically designed for SS have been proposed. The ReCo method [76] is one of the first contrastive learning proposals for SS. This method consists in chaining on top of the segmentation model encoder an auxiliary decoder that maps the input feature to a higher dimensional representation space, in which the sampling of queries and keys is carried out. By means of the proposed contrastive loss function the query is enforced to be close to the positive key in the representation space, and away from the negative key. Because using all pixels of a high-dimensional image to compute the contrastive loss function is impractical, ReCo method incorporates an active sampling strategy that samples less than 5% of the total pixels in the image. On the one hand, this method gives a higher probability of being selected as key negative those pixels belonging to classes that are usually confused with the query class. On the other hand, it relies on prediction confidence to select those pixels that are more difficult to classify for the segmentation model as query pixels. However, these methods encounter challenges when densely packed decision boundaries are present [124]. In such cases, effective clustering based on similarity becomes infeasible, often leading to label noise.

Another contrastive learning method proposed for semi-supervised SS is based on positive-only contrastive learning [19], in which only positive keys are sampled. The key element of this method is the creation and dynamic updating of a memory bank containing a subset of samples from the labeled set. The samples with a higher prediction confidence are selected to be stored. Subsequently, a contrastive loss function ensures that the features of a sample are close to the features of the samples of the same class stored in the memory bank [1].

3.8 Hybrid methods

The last category includes those methods that share characteristics of several of the previously introduced categories. Hybrid methods that attempt to take advantage of the benefits of pseudo-labeling and consistency regularization methods are some of the most common in this category. For instance, a three-stage self-training framework whit an intermediate stage of consistency regularization [61] is proposed. Specifically, a multi-task model is integrated in the self-training process. It is trained on the segmentation problem using consistency regularization (task 1), and statistical information is introduced into the optimization process from the pseudo-labels (task 2).

In the same way, Adaptive Equalization Learning (AEL) [55] also incorporates characteristics of consistency regularization and pseudo-labeling methods. AEL method is based on FixMatch [117], a widely used hybrid method

originally proposed for image classification. It is common in segmentation problems that models underperform in some classes, mainly due to their difficulty or negative imbalance with respect to the rest of the classes. AEL focuses on these challenging classes. This method proposes a confidence bank that dynamically stores the performance of each category during training. Data augmentation techniques and adaptive equalization sampling are used to favor the training towards those disadvantaged classes.

Pseudo-Seg [165] also integrates characteristics of consistency regularization and pseudo-labeling methods. The authors emphasize the fact that the usual ways of obtaining pseudo-labels (from the outputs of a trained segmentation model and applying a confidence threshold) can fail and result in low-quality pseudo-labels. To address this problem, an approach focused on performing a structured and quality design of pseudo-labels is proposed. This method generates the pseudo-labels from two different sources: on the one hand, the output of the segmentation model and, on the other hand, the output of a class activation map algorithm [109]. Unlike the segmentation task that seeks to obtain a dense and accurate prediction, the class activation algorithms perform a simpler task in which they only need to predict coarser-grained outputs.

A key bottleneck in semi-supervised segmentation methods can be to treat labeled and unlabeled data separately during training. This is the issue that the hybrid GuidedMix-Net method focuses on [133], allowing a transfer of knowledge from labeled to unlabeled images. This is achieved through an interpolation between pairs of labeled and unlabeled images, thus capturing interactions between them.

Interest in methods that combine consistency regularization with contrastive learning has also increased recently. In this line, methods such as directional context-aware (DCA) [65] have been proposed. The authors point out the difficulty of generalizing in a semi-supervised environment, where the contexts of a given object are limited in the reduced set of labeled images. This may cause a segmentation model to give too much importance to these specific contexts, not focusing on some important characteristics of the object to be segmented. To address this issue, The DCA method incorporates a new data augmentation technique that makes two cuts of the same image with an overlapping region. In this way it simulates two different contexts for that region, and enforces consistency between the two slices by means of a contrastive loss function.

The approach proposes in [159] tries to achieve two properties: consistency in the prediction space and contrastiveness in the feature space. On one hand, they enforce consistency between the predictions of two augmented versions of an unlabeled image using the l_2 loss. On the other hand, they integrate contrastive learning by means of a contrastive loss function that brings positive (similar) pairs closer and negative (dissimilar) pairs away in the feature space.

Another method that combines consistency regularization and contrastive learning is C3-SemiSeg, presented in [162]. In this method, consistency regularization is focused on exploiting feature alignment under perturbations, introducing a novel cross-set region-level data augmentation strategy. In addition, cross-set contrastive learning is integrated to improve the feature representation capability.

A method presented in [144] combines a consistency regularization framework based on cross-teacher training (CCT) with two complementary contrastive learning modules. CCT framework reduces the accumulation of errors between teacher and student networks while contrastive learning modules promote class separation in the feature space.

Finally, a method combining consistency regularization and adversarial training has been recently proposed [8]. In this case, a data augmentation technique that tries to maintain the image context is proposed. Additionally, a new adversarial dual-student framework is proposed in order to improve the performance of the classical Mean Teacher.

4 Experimental setup

The main obstacle to have a realistic perception of the performance of the different state-of-the-art methods is the non-homogeneity of the comparative experiments presented. As a consequence, a direct comparison of the results obtained by each method is impossible. Among these differences we can find the use of different datasets or partitions of labeled and unlabeled data, different base models on which semi-supervised methods are based or different preprocessing or data augmentation techniques.

That is why the main goal of this experimental section is to offer the reader a comparison with unified, fair and equal conditions for all methods, thus offering a quick and accessible way to know the actual state-of-the-art methods in the field. To this end, we have carried out a series of experiments taking into account some guidelines that try to eradicate the comparison problems described above, on a selection of methods that tries to be representative for all the categories introduced in our taxonomy.

Our experimentation is mainly conducted in two directions. On the one hand, we propose an experiment with exhaustive representation of all categories of methods, on a range of partitions with different ratios of labeled and unlabeled data, with the aim of having quantitative results that allow a direct and fast comparison between the performance of the different methods. On the other hand, we propose another experimentation with some of the most relevant methods in the literature to perform a qualitative and visual comparison of the results obtained.

Datasets. For each experiment described in the previous section we chose datasets that we consider to have the necessary characteristics to carry out the desired comparison. A detailed description of the following datasets can be found in section 2.4. We employ the PASCAL VOC 2012 [29] dataset in the experiment related to the quantitative comparison of state-of-the-art methods. This dataset is the most commonly used in the semi-supervised SS literature. In addition, it has a high number of images which helps to have stability in the results obtained. In this experimental section, we also employ MetalDAM [80], a dataset that we believe has significantly different characteristics from PASCAL VOC 2012. This choice was made with the objective of elucidating the behavior of the methods in different scenarios. Second, for qualitative and visual comparison of the results we considered the Cityscapes dataset [23]. This dataset has higher resolution images, which allows a better visualization. In addition, each of the images in this dataset has representation of many of the classes (unlike PASCAL VOC 2012, where each image focuses on one or a small number of classes). This allows us to see how the trained models perform in situations where there are adjacent areas of several similar classes or with semantic dependencies between them. These two features make Cityscapes an ideal dataset to perform visual and qualitative analysis.

Partition protocol. As discussed above, partitions of labeled and unlabeled data is a key aspect to take into account in semi-supervised experiments. In order to obtain comparable results with other experimental studies, it is important to use the same or similar data partitions. That is why in our experimentation we decided to use the partitions proposed in one of the most recent studies, which present a wide variety of scenarios in terms of labeling ratio. These partitions can be found at [4]. This strategy makes a random sampling of instances, of which we will consider their labels, without taking into account class balancing in order to have a realistic scenario.

Validation strategy. The standard validation strategy in SS on the datasets used in our experiments consists of a simple holdout, with a training set and a validation set. For each of the datasets, the composition of training and validation partitions is standard in the SS literature, so we use these same partitions for better generality. Each of the models have been trained and tested three times and averaged, therefore the resultant metric is the average of the three runs presented along with the standard deviation. The hyperparameters chosen for each model configuration have been obtained from the original code of the proposal, being either the default parameters or the ones recommended by the authors in their papers.

[4] https://github.com/charlesCXK/TorchSemiSeg

Performance metric. The performance metric used in this experimentation, standard in the SS literature, is the mean intersection over union (mean IoU). Unlike accuracy, this metric tries to be robust to the presence of imbalanced classes, which is very common in problems where we have pixel-level labels. Specifically, this metric computes the ratio between the number of true positives and the sum of true positives, false negatives and false positives, for each of the classes and averages these values.

$$meanIoU = \frac{1}{N}\sum_{i=1}^{N}\frac{N_{ii}}{\sum_{j=1}^{N}N_{ij} + \sum_{j=1}^{N}N_{ji} - N_{ii}} \tag{16}$$

where N is the number of classes, N_{ii} is the numbers of true positives for class i, N_{ij} is the numbers of false positives for class i and j and N_{ji} is the number of false negatives for class j and i.

Selection of state-of-the-art methods. We include in the experimental study state-of-the-art methods such that all categories and subcategories defined in the taxonomy presented in section 3 are sufficiently covered. The main criteria taken into account when choosing a method from each category have been popularity of the method, in terms of number of citations, and availability of code. As baseline methods, we include the DeepLabV3+ [15] and SegFormer [146] supervised models, trained only with the labeled partition and the Mean Teacher [127] method, which has a strong influence on most of the proposed methods. The semi-supervised methods included in our experimental study are as follows: s4GAN [87], ClassMix [94], CCT [97], CPS [20], ST [148], DMT [33], ReCO [76] and CAC [65]. All the selected methods for testing have a public implementation available, being this the one selected to be executed.

Base model and backbone. All semi-supervised methods for SS work by supporting a supervised segmentation model. The good performance of the semi-supervised method depends to a large extent on the base model. This is why the choice of this base model is critical. As well, segmentation models rely on a network (i.e., backbone), on which the final performance of the semi-supervised segmentation method also depends. The fact that the different proposals for semi-supervised methods rely on different base models and backbones makes it difficult to compare their performance, which is why in our experimentation we unified this critical aspect. We opt for DeepLabV3+ as the base model and ResNet101 as the backbone, this combination being one of the best performing in the literature.

Hardware and software setup. The entire experimental code has been developed using Python as programming language and PyTorch as Deep Learning framework. The different experiments have been run on a Tesla V100 GPU.

5 Results and discussion

In this section we show and discuss the results obtained. First, in subsection 5.1 and subsection 5.2 we present and discuss the quantitative results obtained on the PASCAL VOC 2012 and MetalDAM datasets, respectively. Secondly, we present the results obtained on Cityscapes in subsection 5.3, carrying out a qualitative and visual analysis of some of the most popular methods, showing some key examples where the performance of these methods can be observed.

5.1 Quantitative results on PASCAL VOC 2012

In this section we show and analyze the performance of the methods that compose our experimentation on PASCAL VOC 2012 dataset, whose quantitative results can be seen in Table 3.

The first aspect to evaluate is the difference in performance between supervised and semi-supervised approaches. It is evident that semi-supervised approaches must show some improvement with respect to the supervised model that justifies the increase in complexity necessary to process the unlabeled data and extract knowledge from them. However, this requirement is not always fulfilled, and sometimes, in certain scenarios, the inclusion

Table 3. Semi-supervised and fully supervised results on the PASCAL VOC 2012 dataset. Each column corresponds to a ratio of labeled/unlabeled images (the number on the left represents the number of labeled images used in each case). In each column the result obtained with the best performing method averaged over three runs is highlighted. Standard deviation is presented next to the metric. (Metric: mean IoU).

Method	**1/100** (106)	**1/50** (212)	**1/20** (529)	**1/8** (1323)
DeepLabV3+	48.8±1.8	57.4±1.4	66.2±1.2	70.2±0.6
SegFormer	43.2±2.7	52.3±2.1	64.1±1.1	69.1±0.8
Mean Teacher	44.9±2.5	58.5±1.8	67.8±1.2	71.6±0.5
ClassMIX	56.5±1.1	67.6±0.8	70.8±0.6	71.9±0.4
CPS	47.7±2.2	56.7±1.7	69.6±0.9	74.7±0.4
CCT	37.1±3.2	52.0±2.2	62.3±1.8	68.6±1.1
s4GAN	50.4±1.9	62.3±1.5	65.3±1.2	71.3±0.6
ST	55.2±1.1	64.8±0.9	71.4±0.7	**74.9±0.4**
DMT	**58.9±0.9**	**70.0±0.7**	**72.3±0.5**	74.4±0.4
ReCo	56.2±1.2	63.2±1.1	68.2±0.8	72.5±0.6
CAC	49.7±2.6	64.3±1.3	70.6±1.1	74.6±0.6
SSDA(DS+US)	57.9±1.1	65.3±0.9	67.1±0.8	71.1±0.5
SemiRoadExNet	41.7±1.1	51.8±1.4	62.7±0.8	65.3±0.5

of unlabeled data in the training process can even harm the performance of the fully supervised model (e.g. CCT). Other methods, such as Mean Teacher, although they do not obtain as notable a deterioration as CCT, also present difficulties in extracting knowledge from unlabeled data, obtaining a gain between 1-2% in all partitions, except in the partition with the least number of label data, in which it obtains worse results than the supervised model.

The next method in performance terms is the s4GAN adversary method. This method obtains performance improvements in almost all the partitions with respect to the supervised model, these improvements varying from 1-5%, except in the 1/20 partition, which does not improve. Although it is true that a 5% improvement could be a desirable improvement in many scenarios, this improvement does not occur in all partitions, presenting some instability in the results depending on the number of labeled data. In addition, this method suffers from an increase in complexity compared to other simpler methods to carry out adversary training, which is hardly justifiable regarding the results.

Other methods show variable results among the different labeling ratios. Some of the best performing methods when we have a very small set of labeled images are the ClassMix and ReCo methods. However, as we increase the number of labeled images, the margin of benefit that this method presents with respect to the supervised baseline is not so wide and there are many other methods that outperform it. Conversely, the CPS and CAC methods are two of the best performers in scenarios where we have many labeled images, and their performance suffers as the size of the labeled partition is reduced, even obtaining worse performance than the supervised baseline in the case of CPS. We can consider these methods as particularly useful in this specific scenario, but not as methods that obtain good overall performance.

Finally, methods based on pseudo-labeling have been shown to be the best performing ones. First, the ST method based on a simple self-training, has obtained the best result in the partition with the highest number of labeled images (1/8) in addition to obtaining competitive results in the rest of the partitions. On the other hand, the DMT method, based on mutual training, obtained the best results in all the partitions, except in the 1/8 partition, which obtained a result less than 1% lower than the best model.

5.2 Quantitative results on MetalDAM

Table 4. Semi-supervised and fully supervised results on the MetalDAM dataset. Each column corresponds to a percentage of labeled images (the number on the left represents the number of labeled images used). In each column the result obtained with the best performing method averaged over three runs is highlighted. Standard deviation is presented next to the metric. (Metric: mean IoU).

Method	**7%** (2)	**25%** (8)	**50%** (14)	**75%**(26)
DeepLabV3+	60.6±2.2	64.1±2.1	64.8±1.6	66.3±1.3
SegFormer	55.8±2.8	58.6±2.9	60.7±2.4	61.5±1.8
Mean Teacher	61.8±1.8	66.3±1.6	68.9±1.1	68.4±0.9
ClassMIX	**64.1±1.5**	67.5±1.6	**70.6±0.8**	69.2±0.6
CPS	61.1±1.6	68.1±1.5	70.5±1.2	71.1±0.9
CCT	58.6±2.1	61.7±1.7	65.2±1.3	66.4±1.4
s4GAN	59.5±1.8	64.9±1.3	66.2±1.1	65.2±1.2
ST	60.5±2.2	67.3±1.8	68.3±1.3	69.1±1.3
DMT	62.8±1.5	**69.4±1.2**	70.5±1.3	**71.3±0.6**
ReCo	62.4±1.6	67.7±1.3	69.5±0.9	69.4±0.5
CAC	60.9±2.2	64.8±1.9	66.9±1.5	67.4±1.4
SSDA(DS+US)	62.1±2.3	65.2±1.7	64.9±1.3	70.2±1.1
SemiRoadExNet	56.2±2.8	59.3±2.5	65.7±1.9	63.8±1.1

In this section we show and analyze the performance of the methods that compose our experimentation on MetalDAM dataset, whose quantitative results can be seen in Table 4.

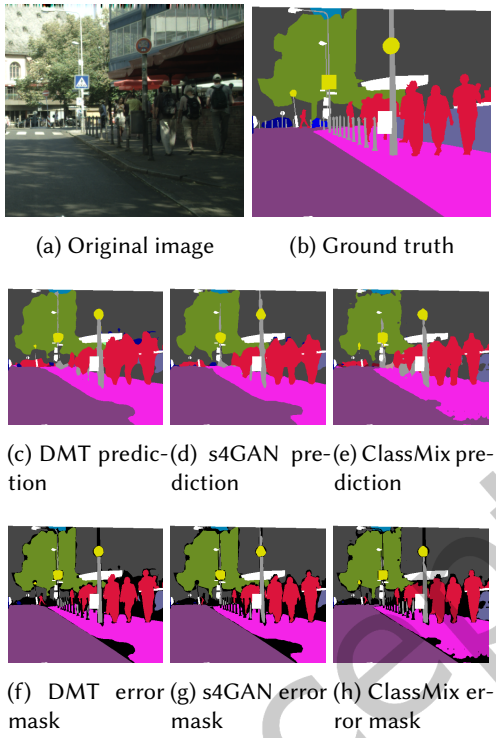
In the results obtained on MetalDAM we can observe similarities in the behavior of the methods with respect to PASCAL VOC 2012. For example, there are some cases where semi-supervised methods perform worse than supervised methods and the worst performing methods in this new scenario are also CCT and s4GAN. Also DMT method remains one of the best performing methods in several cases. Due to the drastic differences between the two datasets, we can generalize with some confidence and expect these results to be repeatable in a multitude of different scenarios.

On the other hand, we see how the ClassMix method is one of the best performers in several experiments (7% and 25% partitions). It makes sense that the Data Augmentation based operation of this method shows its potential in scenarios where the number of training images is very small, as in this case.

Another particular behavior that we observe on this dataset is the minor influence of the number of labeled images that we use. We can see how the maximum difference between the metrics obtained between the largest and smallest partition is about 8%, while in the experimentation on PASCAL VOC 2012 we can observe 30%. This is due to the amount of information provided by each of the images of the datasets in question. While PASCAL VOC 2012 is made up of images containing a reduced number of classes (in many cases only one class), in MetalDAM all the images present appearances of most of the classes in the dataset.

5.3 Qualitative results on Cityscapes

In this subsection we carry out a qualitative and visual analysis of the results obtained on the Cityscapes dataset with some of the most popular state-of-the-art methods. The methods employed in this analysis include DMT, ClassMix, and s4GAN, representing the taxonomy described in this paper through some of the most widely recognized approaches. The popularity of these methods was determined by examining their citation counts and the availability of their corresponding code. Based on the results presented in Tables 3 and 4, it can be reasonably



(a) Original image (b) Ground truth

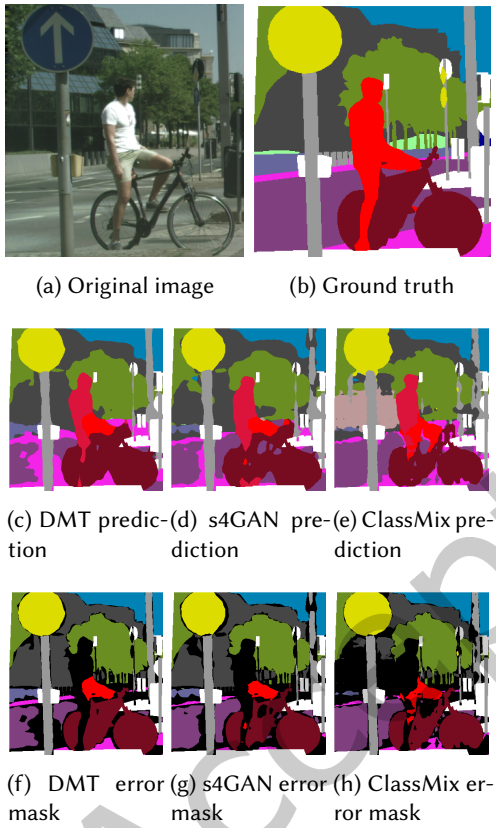
(c) DMT prediction (d) s4GAN prediction (e) ClassMix prediction

(f) DMT error mask (g) s4GAN error mask (h) ClassMix error mask

Fig. 12. Qualitative results obtained with DMT, s4GAN and ClassMix methods in an example of Cityscapes with main representation of the classes person, road, sidewalk, vegetation and building. Black color represents prediction errors.

assumed that other methods within the same taxonomy would exhibit similar behavior. In the following we visually show the segmentation maps predicted with each of these methods on some representative examples of the Cityscapes dataset, comparing them with the ground truth. In addition, in order to clearly and quickly identify the areas where these methods fail, we generate error masks highlighting in black color those areas where the model has predicted an incorrect label. White areas correspond to unlabeled zones in the dataset that are not taken into account in the learning process.

In the first visual example shown in Figure 12 we can see a good and similar performance of the methods used in this qualitative analysis in the classes that predominate in the image, such as the road (●), sidewalk (●), building (●) and vegetation (●) classes. We only see an area in the lower right corner where the three models have a clear confusion between these predominant classes, specifically between the road and sidewalk classes, as we can see in the error masks, due to an irregularity on the sidewalk. Another largely represented class in this image is the person (●) class. Although all models detect the presence of people, they have more difficulties in exactly defining the area belonging to each person. Unlike the previously named classes that usually appear in the image in a single large area, being easy to predict for the models, classes such as person, which present greater fragmentation by appearing in different and smaller areas of the image that do not have to be adjacent, suppose a greater difficulty for the models, obtaining less exact predictions and predicting incorrect classes in the gaps between different instances of the person class.



(a) Original image (b) Ground truth

(c) DMT prediction (d) s4GAN prediction (e) ClassMix prediction

(f) DMT error mask (g) s4GAN error mask (h) ClassMix error mask

Fig. 13. Qualitative results obtained with DMT, s4GAN and ClassMix methods in an example of Cityscapes with main representation of the classes bicycle, rider, pole and traffic sign. Black color represents prediction errors.

In a second example shown in Figure 13, a generalized poor performance on the rider (●) class can be observed. The rider class presents very few differences with respect to the person (●) class. The way in which a model could differentiate a person from a rider would be to look at whether the rider is on a bicycle (●) or motorcycle (●) in the image. However, the results obtained seem to indicate that the models have problems when trying to learn these semantic relationships between classes or contextual information, confusing the instance of the rider class with the person class nearly in its totality. Only some of the parts of the person closest to the bike are segmented with the correct rider class. This indicates that the difficulty of learning semantic relations between objects is even greater as the distance between them increases. Additionally, in this example, we can observe a good generalized performance in the classes traffic sign (●), pole (●), vegetation (●) and sky (●).

5.4 Discussion

We can analyze the methods tested in our experimentation according to various criteria, thereby highlighting certain advantages or disadvantages for each method or family.

Probably the main criterion for comparing the different methods is the performance obtained. Our experiments show that the methods of the Pseudo-labeling family (ST and DMT) generally give the best results.

Another crucial aspect, especially when applying these methods in real-world scenarios, is simplicity or ease of implementation. In this regard, methods like CPS or ST are perhaps the ones with the simplest designs, primarily based on retraining models with previously obtained predictions, without adding additional mechanisms that complicate the process. Mean Teacher is another model that also features a straightforward design and is considered one of the standard baselines in semi-supervised learning.

Considering the trade-off between performance and simplicity, our experimentation reveals a clear winner: ST. However, it's worth noting other methods like CPS, which, despite having a simple design, is close to the best results, or DMT, which, while introducing some more complex elements in the training process, further improves the results, generally obtaining the best results in most experiments.

In contrast, some methods implement complex mechanisms in their training process but do not achieve the performance obtained by the aforementioned methods. Adversarial methods (s4GAN) or methods based on Feature Perturbations (CCT) stand out negatively in this regard.

As we know, the choice of the base segmentation model is a key aspect. Therefore, another advantageous aspect of the design of some methods is the independence from the base model. For example, pseudo-labeling methods (ST and DMT) or some consistency regularization methods (e.g., CPS and ClassMix) have this advantage, allowing the choice of the base model as needed. Thus, these methods can easily adapt to different scenarios, where some base models may be more convenient than others. In contrast, Feature Perturbations methods or some contrastive methods are not independent or pose greater difficulty in switching between different base models.

Another key aspect, especially in semi-supervised scenarios, is the ability to perform well using few labeled data. Hence, we highlight methods like ReCo or ClassMix, which, while not obtaining the best results in partitions with a higher number of labeled images, achieve the best results (along with DMT, which has the best overall performance) in experiments using a lower number of annotations.

The hybrid method used in this study, CAC, presents complex elements in its design (as is common in hybrid methods) but does not generally stand out for obtaining the best results in our comparison. However, due to the diversity among hybrid methods, it is challenging to extrapolate this to all methods belonging to this category.

6 Challenges and future trends

This section presents some of the main challenges related to the semi-supervised SS problem, as well as some of the most promising future research lines.

Evaluation standards. Different studies we found in the semi-supervised SS literature do not present a homogeneous experimental framework (i.e. use of different datasets, different data partitions, different implementations or versions of the base model, etc.). The proposal of a standard and realistic experimental and evaluation framework that all researchers can adopt would be a key point in the development of this field of research.

Diversity in base models. Many of the methods studied employ more than one base model and the diversity of these models can be a key aspect to obtain a good final model. However, these methods are usually limited to choosing the state-of-the-art supervised segmentation model (i.e. DeepLabV3+ [15] at present) obtaining a set of models poor in diversity, and no proposal attempts to go deeper into this decision. A possible future line of research could focus on the study of the implication of inter-model diversity on the final result of semi-supervised segmentation methods.

Evaluation on more realistic scenarios. We have observed that some of the most widely used datasets in both the supervised and semi-supervised segmentation problem are object-centered image datasets (e.g., PASCAL VOC 2012). This type of images represent a very controlled scenario, which we are difficult to find in real-world problems. Models designed to obtain good results in this type of datasets may not be useful in real applications. New emerging datasets (e.g., Cityscapes) present less controlled images and more semantic dependencies between

classes (a clear example of this type of semantic relationships can be seen in Figure 13, between the rider and the bicycle). These types of datasets need new methods capable of dealing with less controlled images and modeling semantic dependencies between classes.

Comparison between different sparse annotation approaches. Annotation sparsity poses a prevalent challenge in SS, primarily attributed to the high cost of pixel-level labeling. Besides semi-supervised approaches, alternative strategies can be considered in such circumstances (Figure 1 showcases a few). These encompass self-supervised learning, few-shot learning, or domain adaptation. Conducting a comparative and experimental study among these diverse SS approaches holds significant potential in advancing this field of research.

New trend: transformers. Despite the fact that transformers have started to be applied in supervised SS, establishing themselves as the state of the art in recent years, they have not yet been successfully introduced in semi-supervised SS scenarios. Consequently, the design of transformer-based methods for semi-supervised SS can be considered one of the most promising future research lines within this field. Within this category, we anticipate the emergence of high-quality foundational models addressing this problem in the near future.

7 Conclusions

This paper seeks to structure the knowledge generated in recent years, as well as to pose challenges and future research trends, around the rise of semi-supervised segmentation methods.

One of the main contributions of this paper is the proposal of a taxonomy, which classifies all previous works (a total of 50 recently published methods related to this field) into five categories: adversarial methods, consistency regularization, pseudo-labeling, constrastive learning and hybrid methods. In this manner, we provide the reader with a quick and precise way to know the state of the art in this field, as well as a detailed description of each method.

The analysis of the state of the art and the defined taxonomy is complemented with an experimental study that compares all taxonomic categories under homogeneous experimental conditions (employing the two most common datasets in the field (PASCAL VOC 2012 and Cityscapes) and a real industrial use case (MetalDAM)). This allows the reader to have an intuition about the performance of each of them.

Finally, we reflect on the current challenges of semi-supervised segmentation and potential future lines of research, highlighting the need for standardization of the experimental and evaluation framework, the convenience of using realistic benchmarks where images are not controlled and are rich in semantic dependencies between classes, and potential application in a semi-supervised scenario of vision transformers.

Acknowledgments

This work was supported by project PID2023-150070NB-I00 granted by Ministerio de Ciencia, Innovación y Universidades. Ignacio Aguilera-Martos was supported by the Ministry of Science of Spain under the FPI programme PRE2021-100169. This work was also supported by the Spanish Ministry of Science and Innovation, the Andalusian Government, and European Regional Development Funds (ERDF) under grants CONFIA (PID2021-122916NB-I00) and FORAGE (B-TIC-456-UGR20). The hardware used in this work is supported by the projects with reference EQC2018-005084-P, granted by the Spain's Ministry of Science and Innovation and European Regional Development Fund (ERDF) and the project with reference SOMM17/6110/UGR, granted by the Andalusian ''Consejería de Conocimiento, Investigación y Universidades" and European Regional Development Funds (ERDF).

References

[1] Iñigo Alonso, Alberto Sabater, David Ferstl, Luis Montesano, and Ana Cristina Murillo. 2021. Semi-Supervised Semantic Segmentation with Pixel-Level Contrastive Learning from a Class-wise Memory Bank. *ICCV* (2021), 8199–8208.

[2] Shan An, Haogang Zhu, Jiaao Zhang, Junjie Ye, Siliang Wang, Jianqin Yin, and Hong Zhang. 2022. Deep Tri-Training for Semi-Supervised Image Segmentation. *IEEE Robotics and Automation Letters* 7, 4 (2022), 10097–10104.

[3] Iro Armeni, Sasha Sax, Amir Roshan Zamir, and Silvio Savarese. 2017. Joint 2D-3D-Semantic Data for Indoor Scene Understanding. *ArXiv* abs/1702.01105 (2017).
[4] Vijay Badrinarayanan, Ignas Budvytis, and Roberto Cipolla. 2013. Semi-Supervised Video Segmentation Using Tree Structured Graphical Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35 11 (2013), 2751–64.
[5] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. 2015. Material recognition in the wild with the Materials in Context Database. *CVPR* (2015), 3479–3487.
[6] Filippo Bergamasco, Andrea Albarelli, and A. Torsello. 2012. A graph-based technique for semi-supervised segmentation of 3D surfaces. *Pattern Recognition Letters* 33 (2012), 2057–2064.
[7] Gabriel J. Brostow, Julien Fauqueur, and Roberto Cipolla. 2009. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters* 30 (2009), 88–97.
[8] Cong Cao, Tianwei Lin, Dongliang He, Fu Li, Huanjing Yue, Jingyu Yang, and Errui Ding. 2022. Adversarial Dual-Student with Differentiable Spatial Warping for Semi-Supervised Semantic Segmentation. *ArXiv* abs/2203.02792, 2 (2022), 793–803.
[9] Olivier Chapelle, Bernhard Schlkopf, and Alexander Zien. 2006. Semi-Supervised Learning. *IEEE Transactions on Neural Networks* 20, 2 (2006), 1.
[10] Changrui Chen, Jungong Han, and Kurt Debattista. 2024. Virtual category learning: A semi-supervised learning method for dense prediction with extremely limited labels. *IEEE transactions on pattern analysis and machine intelligence* 5, 2 (2024), 1.
[11] Chang Wen Chen, Jiebo Luo, and Kevin J. Parker. 1998. Image segmentation via adaptive K-mean clustering and knowledge-based morphological operations with biomedical applications. *IEEE Transactions on Image Processing* 7 12 (1998), 1673–83.
[12] Hao Chen, Zhenghong Li, Jiangjiang Wu, Wei Xiong, and Chun Du. 2023. SemiRoadExNet: A semi-supervised network for road extraction from remote sensing imagery via adversarial learning. *ISPRS Journal of Photogrammetry and Remote Sensing* 198 (2023), 169–183.
[13] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin P. Murphy, and Alan Loddon Yuille. 2018. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40 (2018), 834–848.
[14] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. 2017. Rethinking Atrous Convolution for Semantic Image Segmentation. *ArXiv* abs/1706.05587 (2017).
[15] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. 2018. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *ECCV*. Springer, Milan, Italy, 801–818.
[16] Qi Chen, Lei Wang, Yifan Wu, Guangming Wu, Zhiling Guo, and Steven L. Waslander. 2019. Aerial Imagery for Roof Segmentation: A Large-Scale Dataset towards Automatic Mapping of Buildings. *ArXiv* abs/1807.09532 (2019).
[17] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. *International conference on machine learning* (2020), 1597–1607.
[18] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. 2020. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems* 33 (2020), 22243–22255.
[19] Xinlei Chen and Kaiming He. 2021. Exploring Simple Siamese Representation Learning. *CVPR* (2021), 15745–15753.
[20] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. 2021. Semi-Supervised Semantic Segmentation with Cross Pseudo Supervision. *CVPR* (2021), 2613–2622.
[21] Ying Chen, Xu Ouyang, Kaiyue Zhu, and Gady Agam. 2021. ComplexMix: Semi-supervised semantic segmentation via mask-based data augmentation. *ICIP* (2021), 2264–2268.
[22] Zhenghao Chen, Rui Zhang, Gang Zhang, Zhenhuan Ma, and Tao Lei. 2020. Digging into pseudo label: a low-budget approach for semi-supervised semantic segmentation. *IEEE Access* 8 (2020), 41830–41837.
[23] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The Cityscapes Dataset for Semantic Urban Scene Understanding. *CVPR* (2016), 3213–3223.
[24] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas A. Funkhouser, and Matthias Nießner. 2017. ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes. *CVPR* (2017), 2432–2443.
[25] Brian L. DeCost, Bo Lei, Toby Francis, and Elizabeth A. Holm. 2019. High Throughput Quantitative Metallography for Complex Microstructures Using Deep Learning: A Case Study in Ultrahigh Carbon Steel. *Microscopy and Microanalysis* 25 (2019), 21 – 29.
[26] Terrance Devries and Graham W. Taylor. 2017. Improved Regularization of Convolutional Neural Networks with Cutout. *ArXiv* abs/1708.04552 (2017).
[27] Rui Di and Dan Dan Huang. 2021. Semi-supervised Semantic Segmentation Based on Confrontation Network. *EIECS* (2021), 678–682.
[28] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ArXiv* abs/2010.11929 (2021).
[29] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. 2009. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision* 88 (2009), 303–338.

[30] William Falcon and Kyunghyun Cho. 2020. A Framework For Contrastive Self-Supervised Learning And Designing A New Approach. *ArXiv* abs/2009.00104 (2020).
[31] Pedro F. Felzenszwalb, Ross B. Girshick, David A. McAllester, and Deva Ramanan. 2009. Object Detection with Discriminatively Trained Part Based Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32 (2009), 1627–1645.
[32] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. 2004. Efficient Graph-Based Image Segmentation. *International Journal of Computer Vision* 59 (2004), 167–181.
[33] Zhengyang Feng, Qianyu Zhou, Qiqi Gu, Xin Tan, Guangliang Cheng, Xuequan Lu, Jianping Shi, and Lizhuang Ma. 2022. Dmt: Dynamic mutual training for semi-supervised learning. *Pattern Recognition* (2022), 108777.
[34] Geoff French, Samuli Laine, Timo Aila, Michal Mackiewicz, and Graham Finlayson. 2019. Semi-supervised semantic segmentation needs strong, varied perturbations. *arXiv preprint arXiv:1906.01916* (2019).
[35] Geoff French, Samuli Laine, Timo Aila, Michal Mackiewicz, and Graham D. Finlayson. 2020. Semi-supervised semantic segmentation needs strong, varied perturbations. *BMVC* (2020).
[36] Alberto Garcia-Garcia, Sergio Orts, Sergiu Oprea, Victor Villena-Martinez, Pablo Martinez-Gonzalez, and José García Rodríguez. 2018. A survey on deep learning techniques for image and video semantic segmentation. *Applied Soft Computing* 70 (2018), 41–65.
[37] Yuying Ge, Ruimao Zhang, Lingyun Wu, Xiaogang Wang, Xiaoou Tang, and Ping Luo. 2019. DeepFashion2: A Versatile Benchmark for Detection, Pose Estimation, Segmentation and Re-Identification of Clothing Images. *CVPR* (2019), 5332–5340.
[38] Andreas Geiger, Philip Lenz, and Raquel Urtasun. 2012. Are we ready for autonomous driving? The KITTI vision benchmark suite. *CVPR* (2012), 3354–3361.
[39] Josep Maria Gonfaus, Xavier Boix, Joost van de Weijer, Andrew D. Bagdanov, Joan Serrat Gual, and Jordi Gonzàlez. 2010. Harmony potentials for joint classification and segmentation. *CVPR* (2010), 3280–3287.
[40] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. 2017. Look into Person: Self-Supervised Structure-Sensitive Learning and a New Benchmark for Human Parsing. *CVPR* (2017), 6757–6765.
[41] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep learning*. MIT press.
[42] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. *NeurIPS* 27 (2014).
[43] Simon Graham, Quoc Dang Vu, Shan e Ahmed Raza, Ayesha Azam, Yee-Wah Tsang, Jin Tae Kwak, and Nasir M. Rajpoot. 2019. Hover-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical image analysis* 58 (2019), 101563.
[44] Ivan Grubišić, Marin Oršić, and Siniša Šegvić. 2021. A baseline for semi-supervised learning of efficient semantic segmentation models. *ICMVA* (2021), 1–5.
[45] Ritwik Gupta, Bryce Goodman, Nirav Patel, Ricky Hosfelt, Sandra Sajeev, Eric Heim, Jigar Doshi, Keane Lucas, Howie Choset, and Matthew Gaston. 2019. Creating xBD: A dataset for assessing building damage from satellite imagery. *CVPR* (2019), 10–17.
[46] Kai Han, Victor S Sheng, Yuqing Song, Yi Liu, Chengjian Qiu, Siqi Ma, and Zhe Liu. 2024. Deep semi-supervised learning for medical image segmentation: A review. *Expert Systems with Applications* 245 (2024), 123052.
[47] Shijie Hao, Yuanen Zhou, and Yanrong Guo. 2020. A Brief Survey on Semantic Segmentation with Deep Learning. *Neurocomputing* 406 (2020), 302–321.
[48] Bharath Hariharan, Pablo Arbeláez, Lubomir D. Bourdev, Subhransu Maji, and Jitendra Malik. 2011. Semantic contours from inverse detectors. *ICCV* (2011), 991–998.
[49] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. *CVPR* (2016), 770–778.
[50] Ruifei He, Jihan Yang, and Xiaojuan Qi. 2021. Re-distributing biased pseudo labels for semi-supervised semantic segmentation: A baseline investigation. *ICCV* (2021), 6930–6940.
[51] Patrick Helber, Benjamin Bischke, Andreas R. Dengel, and Damian Borth. 2019. EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 12 (2019), 2217–2226.
[52] Mohammad Hesam Hesamian, Wenjing Jia, Xiangjian He, and Paul J. Kennedy. 2019. Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges. *Journal of Digital Imaging* 32 (2019), 582 – 596.
[53] Lukas Hoyer, Dengxin Dai, Qin Wang, Yuhua Chen, and Luc Van Gool. 2023. Improving semi-supervised and domain-adaptive semantic segmentation with self-supervised depth estimation. *International Journal of Computer Vision* 131, 8 (2023), 2070–2096.
[54] Lukas Hoyer, David Joseph Tan, Muhammad Ferjad Naeem, Luc Van Gool, and Federico Tombari. 2024. SemiVL: semi-supervised semantic segmentation with vision-language guidance. In *European Conference on Computer Vision*. Springer, 257–275.
[55] Hanzhe Hu, Fangyun Wei, Han Hu, Qiwei Ye, Jinshi Cui, and Liwei Wang. 2021. Semi-supervised semantic segmentation via adaptive equalization learning. *Advances in Neural Information Processing Systems* 34 (2021), 22106–22118.
[56] Wei-Chih Hung, Yi-Hsuan Tsai, Yan-Ting Liou, Yen-Yu Lin, and Ming-Hsuan Yang. 2018. Adversarial Learning for Semi-supervised Semantic Segmentation. *ArXiv* abs/1802.07934 (2018).

[57] Rushi Jiao, Yichi Zhang, Le Ding, Bingsen Xue, Jicong Zhang, Rong Cai, and Cheng Jin. 2024. Learning with limited annotations: a survey on deep semi-supervised learning for medical image segmentation. *Computers in Biology and Medicine* 169 (2024), 107840.
[58] Ge Jin, Chuancai Liu, and Xu Chen. 2021. Adversarial network integrating dual attention and sparse representation for semi-supervised semantic segmentation. *Information Processing and Management* 58 (2021), 102680.
[59] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and Improving the Image Quality of StyleGAN. *CVPR* (2020), 8107–8116.
[60] Isinsu Katircioglu, Helge Rhodin, Victor Constantin, Jörg Spörri, Mathieu Salzmann, and Pascal Fua. 2021. Self-supervised human detection and segmentation via background inpainting. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 12 (2021), 9574–9588.
[61] Rihuan Ke, Angelica I. Avilés-Rivero, Saurabh Pandey, Saikumar Reddy, and Carola-Bibiane Schönlieb. 2022. A Three-Stage Self-Training Framework for Semi-Supervised Semantic Segmentation. *IEEE Transactions on Image Processing* 31 (2022), 1805–1815.
[62] Zhanghan Ke, Di Qiu, Kaican Li, Qiong Yan, and Rynson WH Lau. 2020. Guided collaborative training for pixel-wise semi-supervised learning. *ECCV* (2020), 429–445.
[63] Jongmok Kim, Jooyoung Jang, and Hyunwoo Park. 2020. Structured Consistency Loss for semi-supervised semantic segmentation. *ArXiv* abs/2001.04647 (2020).
[64] Alex Krizhevsky. 2009. Learning Multiple Layers of Features from Tiny Images. (2009).
[65] Xin Lai, Zhuotao Tian, Li Jiang, Shu Liu, Hengshuang Zhao, Liwei Wang, and Jiaya Jia. 2021. Semi-supervised Semantic Segmentation with Directional Context-aware Consistency. *CVPR* (2021), 1205–1214.
[66] Vuong Le, Jonathan Brandt, Zhe L. Lin, Lubomir D. Bourdev, and Thomas S. Huang. 2012. Interactive Facial Feature Localization. *ECCV* (2012), 679–692.
[67] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521, 7553 (2015), 436–444.
[68] Dong-Hyun Lee. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. *ICML* 3, 2 (2013), 896.
[69] Daiqing Li, Junlin Yang, Karsten Kreis, Antonio Torralba, and Sanja Fidler. 2021. Semantic Segmentation with Generative Models: Semi-Supervised Learning and Strong Out-of-Domain Generalization. *CVPR* (2021), 8296–8307.
[70] Haoliang Li and Huicheng Zheng. 2021. A Residual Correction Approach for Semi-supervised Semantic Segmentation. *PRCV* (2021), 90–102.
[71] Xiaoqiang Li, Qin He, Songmin Dai, Pin Wu, and Weiqin Tong. 2020. Semi-Supervised Semantic Segmentation Constrained by Consistency Regularization. *ICME* (2020), 1–6.
[72] Bingyuan Liu, Christian Desrosiers, Ismail Ben Ayed, and Jose Dolz. 2023. Segmentation with mixed supervision: Confidence maximization helps knowledge distillation. *Medical Image Analysis* 83 (2023), 102670.
[73] Beibei Liu and Bei Hua. 2019. Semi-supervised semantic image segmentation using dual discriminator adversarial networks. *ICDIP)* 11179 (2019), 36–41.
[74] Ce Liu, Jenny Yuen, and Antonio Torralba. 2011. Nonparametric Scene Parsing via Label Transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33 (2011), 2368–2382.
[75] Li Liu, Wanli Ouyang, Xiaogang Wang, Paul W. Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. 2019. Deep Learning for Generic Object Detection: A Survey. *International Journal of Computer Vision* 128 (2019), 261–318.
[76] Shikun Liu, Shuaifeng Zhi, Edward Johns, and Andrew J. Davison. 2021. Bootstrapping Semantic Segmentation with Regional Contrast. *ArXiv* abs/2104.04465 (2021).
[77] Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. 2019. Structured Knowledge Distillation for Semantic Segmentation. *CVPR* (2019), 2599–2608.
[78] Yuyuan Liu, Yu Tian, Yuanhong Chen, Fengbei Liu, Vasileios Belagiannis, and Gustavo Carneiro. 2022. Perturbed and strict mean teachers for semi-supervised semantic segmentation. *CVPR* (2022), 4258–4267.
[79] Pauline Luc, Camille Couprie, Soumith Chintala, and Jakob Verbeek. 2016. Semantic Segmentation using Adversarial Networks. *ArXiv* abs/1611.08408 (2016).
[80] Julián Luengo, Raúl Moreno, Iván Sevillano-García, David Charte, Adrián Peláez-Vegas, Marta Fernández-Moreno, Pablo Mesejo, and Francisco Herrera. 2022. A tutorial on the segmentation of metallographic images: Taxonomy, new MetalDAM dataset, deep learning-based ensemble model, experimental analysis and challenges. *Information Fusion* 78 (2022), 232–253.
[81] Dwarikanath Mahapatra, Peter J. Schüffler, Jeroen A. W. Tielbeek, Frans Vos, and Joachim M. Buhmann. 2013. Semi-Supervised and Active Learning for Automatic Segmentation of Crohn's Disease. *MICCAI* 16 Pt 2 (2013), 214–21.
[82] Marcin Marszalek and Cordelia Schmid. 2007. Accurate Object Localization with Shape Masks. *CVPR* (2007), 1–8.
[83] Daniela O Medley, Carlos Santiago, and Jacinto C Nascimento. 2021. Cycoseg: a cyclic collaborative framework for automated medical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 11 (2021), 8167–8182.
[84] Robert Mendel, Luis Antonio De Souza, David Rauber, João Paulo Papa, and Christoph Palm. 2020. Semi-supervised Segmentation Based on Error-Correcting Supervision. *ECCV* (2020), 141–157.

[85] Yanda Meng, Hongrun Zhang, Yitian Zhao, Dongxu Gao, Barbra Hamill, Godhuli Patri, Tunde Peto, Savita Madhusudhan, and Yalin Zheng. 2022. Dual consistency enabled weakly and semi-supervised optic disc and cup segmentation with dual adaptive graph convolutional networks. *IEEE transactions on medical imaging* 42, 2 (2022), 416–429.
[86] Shervin Minaee, Yuri Boykov, Fatih Murat Porikli, Antonio J. Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. 2022. Image Segmentation Using Deep Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44 (2022), 3523–3542.
[87] Sudhanshu Mittal, Maxim Tatarchenko, and Thomas Brox. 2021. Semi-Supervised Semantic Segmentation With High- and Low-Level Consistency. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43 (2021), 1369–1379.
[88] Takeru Miyato, Shin ichi Maeda, Masanori Koyama, and Shin Ishii. 2019. Virtual Adversarial Training: A Regularization Method for Supervised and Semi-Supervised Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41 (2019), 1979–1993.
[89] Gabriele Moser, Sebastiano Bruno Serpico, and Jón Atli Benediktsson. 2012. Markov random field models for supervised land cover classification from very high resolution multispectral remote sensing images. *TyWRRS* (2012), 235–242.
[90] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Loddon Yuille. 2014. The Role of Context for Object Detection and Semantic Segmentation in the Wild. *CVPR* (2014), 891–898.
[91] Lichao Mou, Yuansheng Hua, and Xiaoxiang Zhu. 2019. A Relation-Augmented Fully Convolutional Network for Semantic Segmentation in Aerial Scenes. *CVPR* (2019), 12408–12417.
[92] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulò, and Peter Kontschieder. 2017. The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes. *ICCV* (2017), 5000–5009.
[93] Munan Ning, Donghuan Lu, Yujia Xie, Dongdong Chen, Dong Wei, Yefeng Zheng, Yonghong Tian, Shuicheng Yan, and Li Yuan. 2023. Madav2: Advanced multi-anchor based active domain adaptation segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).
[94] Viktor Olsson, Wilhelm Tranheden, Juliano Pinto, and Lennart Svensson. 2021. ClassMix: Segmentation-Based Data Augmentation for Semi-Supervised Learning. *WACV* (2021), 1368–1377.
[95] Marin Orsic, Ivan Kreso, Petra Bevandic, and Sinisa Segvic. 2019. In Defense of Pre-Trained ImageNet Architectures for Real-Time Semantic Segmentation of Road-Driving Images. *CVPR* (2019), 12599–12608.
[96] Yassine Ouali, Céline Hudelot, and Myriam Tami. 2020. An Overview of Deep Semi-Supervised Learning. *ArXiv* abs/2006.05278 (2020).
[97] Yassine Ouali, Céline Hudelot, and Myriam Tami. 2020. Semi-Supervised Semantic Segmentation With Cross-Consistency Training. *CVPR* (2020), 12671–12681.
[98] Jizong Peng, Guillermo Estrada, Marco Pedersoli, and Christian Desrosiers. 2020. Deep co-training for semi-supervised image segmentation. *Pattern Recognition* 107 (2020), 107269.
[99] Dong ping Tian. 2014. Semi-supervised learning for refining image annotation based on random walk model. *Knowledge-Based Systems* 72 (2014), 72–80.
[100] Alec Radford, Luke Metz, and Soumith Chintala. 2017. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *ICIG* (2017), 97–108.
[101] Maryam Rahnemoonfar, Tashnim Chowdhury, Argho Sarkar, Debvrat Varshney, Masoud Yari, and Robin Murphy. 2021. FloodNet: A High Resolution Aerial Imagery Dataset for Post Flood Scene Understanding. *IEEE Access* 9 (2021), 89644–89654.
[102] Lingyan Ran, Yali Li, Guoqiang Liang, and Yanning Zhang. 2024. Semi-supervised semantic segmentation based on pseudo-labels: A survey. *arXiv preprint arXiv:2403.01909* (2024).
[103] Waseem Rawat and Zenghui Wang. 2017. Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation* 29, 9 (2017), 2352–2449.
[104] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. 2016. You Only Look Once: Unified, Real-Time Object Detection. *CVPR* (2016), 779–788.
[105] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. *MICCAI* (2015), 234–241.
[106] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. 2022. Map-guided curriculum domain adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 6 (2022), 3139–3153.
[107] Jürgen Schmidhuber. 2015. Deep learning in neural networks: An overview. *Neural networks* 61 (2015), 85–117.
[108] Florian Schroff, Antonio Criminisi, and Andrew Zisserman. 2008. Object Class Segmentation using Random Forests. *BMVC* (2008), 1–10.
[109] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. 2019. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision* 128 (2019), 336–359.
[110] Mehmet Sezgin and Bülent Sankur. 2004. Survey over image thresholding techniques and quantitative performance evaluation. *Journal of Electronic Imaging* 13 (2004), 146–168.
[111] Evan Shelhamer, Jonathan Long, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. *CVPR* (2015), 3431–3440.

[112] Xiaoyong Shen, Aaron Hertzmann, Jiaya Jia, Sylvain Paris, Brian L. Price, Eli Shechtman, and Ian Sachs. 2016. Automatic Portrait Segmentation for Image Stylization. *Computer Graphics Forum* 35, 2 (2016), 93–102.
[113] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. 2012. Indoor Segmentation and Support Inference from RGBD Images. *ECCV* (2012), 746–760.
[114] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. *ICLR* (2015).
[115] Amber L Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram Van Ginneken, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, et al. 2019. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *ArXiv* abs/1902.09063 (2019).
[116] Korsuk Sirinukunwattana, Josien PW Pluim, Hao Chen, Xiaojuan Qi, Pheng-Ann Heng, Yun Bo Guo, Li Yang Wang, Bogdan J Matuszewski, Elia Bruni, Urko Sanchez, et al. 2017. Gland segmentation in colon histology images: The glas challenge contest. *Medical image analysis* 35 (2017), 489–502.
[117] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems* 33 (2020), 596–608.
[118] Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. 2015. SUN RGB-D: A RGB-D scene understanding benchmark suite. *CVPR* (2015), 567–576.
[119] Nasim Souly, Concetto Spampinato, and Mubarak Shah. 2017. Semi Supervised Semantic Segmentation Using Generative Adversarial Network. *ICCV* (2017), 5689–5697.
[120] Joes Staal, Michael David Abràmoff, Meindert Niemeijer, Max A. Viergever, and Bram van Ginneken. 2004. Ridge-based vessel segmentation in color images of the retina. *IEEE Transactions on Medical Imaging* 23 (2004), 501–509.
[121] Hamil Stanly, Mercy Shalinie S., and Riji Paul. 2023. A review of generative and non-generative adversarial attack on context-rich images. *Engineering Applications of Artificial Intelligence* 124 (2023), 106595.
[122] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. 2021. Segmenter: Transformer for semantic segmentation. (2021), 7262–7272.
[123] Chun-Yu Sun, Yu-Qi Yang, Hao-Xiang Guo, Peng-Shuai Wang, Xin Tong, Yang Liu, and Heung-Yeung Shum. 2023. Semi-supervised 3D shape segmentation with multilevel consistency and part substitution. *Computational Visual Media* 9, 2 (2023), 229–247.
[124] Changki Sung, Wanhee Kim, Jungho An, Wooju Lee, Hyungtae Lim, and Hyun Myung. 2024. Contextrast: Contextual Contrastive Learning for Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3732–3742.
[125] Richard Szeliski. 2010. *Computer vision: algorithms and applications*. Springer Science & Business Media.
[126] Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. *ICML* (2019), 6105–6114.
[127] Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *NeurIPS* 30 (2017).
[128] Eu Wern Teh, Terrance Devries, Brendan Duke, Ruowei Jiang, Parham Aarabi, and Graham W. Taylor. 2022. The GIST and RIST of Iterative Self-Training for Semi-Supervised Segmentation. *CRV* (2022), 58–66.
[129] Demetri Terzopoulos and Kurt W. Fleischer. 2005. Deformable models. *The Visual Computer* 4 (2005), 306–331.
[130] Xin-Yi Tong, Guisong Xia, Qikai Lu, Huanfeng Shen, Shengyang Li, Shucheng You, and Liangpei Zhang. 2018. Land-cover classification with high-resolution remote sensing images using transferable deep models. *Remote Sensing of Environment* 237 (2018), 111322.
[131] Alain Trémeau and Nathalie Borel. 1997. A region growing and merging algorithm to color segmentation. *Pattern Recognition* 30 (1997), 1191–1203.
[132] Isaac Triguero, Salvador García, and Francisco Herrera. 2013. Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study. *Knowledge and Information Systems* 42 (2013), 245–284.
[133] Peng Tu, Yawen Huang, Rongrong Ji, Feng Zheng, and Ling Shao. 2021. GuidedMix-Net: Learning to Improve Pseudo Masks Using Labeled Images as Reference. *ArXiv* abs/2106.15064 (2021).
[134] Abhinav Valada, Gabriel L. Oliveira, Thomas Brox, and Wolfram Burgard. 2016. Deep Multispectral Semantic Scene Understanding of Forested Environments Using Multimodal Fusion. *ISER* (2016), 465–477.
[135] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation Learning with Contrastive Predictive Coding. *ArXiv* abs/1807.03748 (2018).
[136] Jesper E. van Engelen and Holger H. Hoos. 2019. A survey on semi-supervised learning. *Machine Learning* 109 (2019), 373–440.
[137] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
[138] Mei Wang and Weihong Deng. 2021. Deep Face Recognition: A Survey. *Neurocomputing* 429 (2021), 215–244.
[139] Wei Wang and Zhi-Hua Zhou. 2010. A New Analysis of Co-Training. *ICML* (2010).
[140] X. Wang, Abhinav Shrivastava, and Abhinav Kumar Gupta. 2017. A-Fast-RCNN: Hard Positive Generation via Adversary for Object Detection. *CVPR* (2017), 3039–3048.

[141] Zhihao Wang, Jian Chen, and Steven C. H. Hoi. 2021. Deep Learning for Image Super-Resolution: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43 (2021), 3365–3387.
[142] Yulin Wu, Chang Liu, Lei Chen, Dong Zhao, Qinghe Zheng, and Hongchao Zhou. 2022. Perturbation consistency and mutual information regularization for semi-supervised semantic segmentation. *Multimedia Systems* (2022), 1–13.
[143] Yunyang Wu, Xiaobo Zhang, Xiaole Zhao, Yimin Sun, and Tianrui Li. 2025. segWCD: A new segmentation-based weak supervision neural network for building change detection. *Applied Intelligence* 55, 2 (2025), 147.
[144] Hui Xiao, Dong Li, Hao Xu, Shuibo Fu, Diqun Yan, Kangkang Song, and Chengbin Peng. 2022. Semi-supervised semantic segmentation with cross teacher training. *Neurocomputing* 508 (2022), 36–46.
[145] Zhu Xiaojin. 2008. Semi-supervised learning literature survey. *Computer Sciences TR* 1530 (2008).
[146] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems* 34 (2021), 12077–12090.
[147] Di Xu and Zhili Wang. 2021. Semi-supervised semantic segmentation using an improved generative adversarial network. *Journal of Intelligent & Fuzzy Systems* 40, 5 (2021), 9709–9719.
[148] Lihe Yang, Wei Zhuo, Lei Qi, Yinghuan Shi, and Yang Gao. 2022. St++: Make self-training work better for semi-supervised semantic segmentation. *CVPR* (2022), 4268–4277.
[149] Xiangli Yang, Zixing Song, Irwin King, and Zenglin Xu. 2021. A Survey on Deep Semi-supervised Learning. *ArXiv* abs/2103.00550 (2021).
[150] David Yarowsky. 1995. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. *ACL* (1995), 189–196.
[151] Jianlong Yuan, Yifan Liu, Chunhua Shen, Zhibin Wang, and Hao Li. 2021. A Simple Baseline for Semi-supervised Semantic Segmentation with Strong Data Augmentation. *ICCV* (2021), 8209–8218.
[152] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Young Joon Yoo. 2019. CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features. *ICCV* (2019), 6022–6031.
[153] Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Hameed Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Guisong Xia, and Xiang Bai. 2019. iSAID: A Large-scale Dataset for Instance Segmentation in Aerial Images. *CVPR* (2019).
[154] Matthew D. Zeiler, Graham W. Taylor, and Rob Fergus. 2011. Adaptive deconvolutional networks for mid and high level feature learning. *ICCV* (2011), 2018–2025.
[155] Jia Zhang, Zhixin Li, Canlong Zhang, and Huifang Ma. 2020. Robust Adversarial Learning For Semi-Supervised Semantic Segmentation. *ICIP* (2020), 728–732.
[156] Jia Zhang, Zhixin Li, Canlong Zhang, and Huifang Ma. 2021. Stable self-attention adversarial learning for semi-supervised semantic image segmentation. *Journal of Visual Communication and Image Representation* 78 (2021), 103170.
[157] Man Zhang, Yong Zhou, Jiaqi Zhao, Yiyun Man, Bing Liu, and Rui Yao. 2020. A survey of semi- and weakly supervised semantic segmentation of images. *Artificial Intelligence Review* 53 (2020), 4259 – 4288.
[158] Ying Zhang, Tao Xiang, Timothy M. Hospedales, and Huchuan Lu. 2018. Deep Mutual Learning. *CVPR* (2018), 4320–4328.
[159] Yuanyi Zhong, Bodi Yuan, Hong Wu, Zhiqiang Yuan, Jian Peng, and Yu-Xiong Wang. 2021. Pixel Contrastive-Consistent Semi-Supervised Semantic Segmentation. *ICCV* (2021), 7253–7262.
[160] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2017. Scene Parsing through ADE20K Dataset. *CVPR* (2017), 5122–5130.
[161] Yan Zhou, Ruyi Jiao, Dongli Wang, Jinzhen Mu, and Jianxun Li. 2022. Catastrophic Forgetting Problem in Semi-Supervised Semantic Segmentation. *IEEE Access* 10 (2022), 48855–48864.
[162] Yanning Zhou, Hang Xu, Wei Zhang, Bin Gao, and Pheng-Ann Heng. 2021. C3-semiseg: Contrastive semi-supervised segmentation via cross-set learning and dynamic class-balancing. *ICCV* (2021), 7036–7045.
[163] Zhi-Hua Zhou and Ming Li. 2010. Semi-supervised learning by disagreement. *Knowledge and Information Systems* 24, 3 (2010), 415–439.
[164] Yi Zhu, Zhongyue Zhang, Chongruo Wu, Zhi Zhang, Tong He, Hang Zhang, R Manmatha, Mu Li, and Alexander J Smola. 2021. Improving Semantic Segmentation via Efficient Self-Training. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021), 1–1.
[165] Yuliang Zou, Zizhao Zhang, Han Zhang, Chun-Liang Li, Xiao Bian, Jia-Bin Huang, and Tomas Pfister. 2021. PseudoSeg: Designing Pseudo Labels for Semantic Segmentation. *ArXiv* abs/2010.09713 (2021).

Received 23 April 2025; revised 17 December 2025; accepted 24 March 2026