SemiCVT: Semi-Supervised Convolutional Vision Transformer for Semantic Segmentation

Huimin Huang[1], Shiao Xie[1†], Lanfen Lin[1*], Ruofeng Tong[1,2], Yen-Wei Chen[3], Yuexiang Li[4], Hong Wang[4], Yawen Huang[4*], Yefeng Zheng[4]
[1] Zhejiang University, [2] Zhejiang Lab, [3] Ritsumeikan University, [4] Tencent Jarvis Lab

Abstract

Semi-supervised learning improves data efficiency of deep models by leveraging unlabeled samples to alleviate the reliance on a large set of labeled samples. These successes concentrate on the pixel-wise consistency by using convolutional neural networks (CNNs) but fail to address both global learning capability and class-level features for unlabeled data. Recent works raise a new trend that Transformer achieves superior performance on the entire feature map in various tasks. In this paper, we unify the current dominant Mean-Teacher approaches by reconciling intra-model and inter-model properties for semi-supervised segmentation to produce a novel algorithm, SemiCVT, that absorbs the quintessence of CNNs and Transformer in a comprehensive way. Specifically, we first design a parallel CNN-Transformer architecture (CVT) with introducing an intra-model local-global interaction schema (LGI) in Fourier domain for full integration. The inter-model class-wise consistency is further presented to complement the class-level statistics of CNNs and Transformer in a cross-teaching manner. Extensive empirical evidence shows that SemiCVT yields consistent improvements over the state-of-the-art methods in two public benchmarks.

1. Introduction

Semantic segmentation [4, 25, 41, 44] is a foundational problem in computer vision and has attracted tremendous interests for assigning pixel-level semantic labels in an image. Despite remarkable successes of convolutional neural network (CNN), collecting a large quantity of pixel-level annotations is quite expensive and time-consuming. Recently, semi-supervised learning (SSL) provides an alternative way to infer labels by learning from a small number of images annotated to fully explore those unlabeled data.

The main stream of semi-supervised learning relies on consistency regularization [13, 30, 33, 38], pseudo labeling [29, 32], entropy minimization [3, 14] and bootstrapping [15]. For semantic segmentation, a typical approach is to build a Mean-Teacher (MT) model [33] (in Fig. 2 (a)), which allows the predictions generated from either teacher or student model as close as possible. However, such a classic structure still suffers from two limitations: **1)** Most of MT-based frameworks are built upon stacking convolutional layers, while the dilemma of CNNs is to capture global representations in the limited receptive field [11]. It results in the neglected ability of aggregating global context and local features, as depicted in Figs. 1 (a) and (c). **2)** These approaches usually leverage the pixel-wise predictions from CNNs to enforce consistency regularization with their focuses on fine-level pair-wise similarity. It may fail to explore rich information in feature space and also overlook the global feature distribution, as shown in Figs. 1 (e), (g).

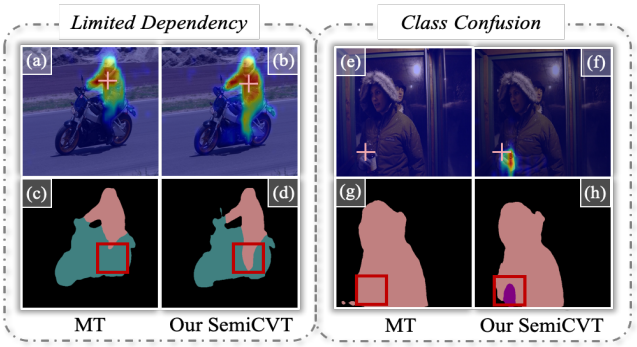


Figure 1. Visualizations of class activation maps generated by Grad-CAM [31] and segmentation results of MT [33] and Our SemiCVT. Current MT-based SSLs suffer from limited global dependency (*e.g.,* incomplete *human leg* in (a), (c)) and class confusion (*e.g.,* mis-classify *bottle* as *person* in (e), (g)). Our SemiCVT improves the performance from intra-model and inter-model perspective, achieving better compactness and accurate localization.

On the other hand, Transformer [34] has achieved notable performance on vision tasks [2, 24, 37], owing to their strong capability in multi-head self-attention for capturing long-distance dependencies. However, pure Transformer-based architectures cannot achieve satisfactory perfor-

*Corresponding Authors: Lanfen Lin (llf@zju.edu.cn), Yawen Huang (yawenhuang@tencent.com).
†Huimin Huang and Shiao Xie are co-first authors, and this work is done during the internship at Tencent Jarvis Lab.

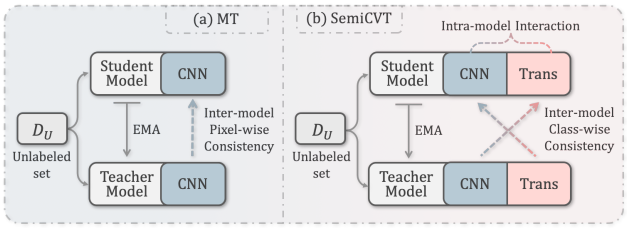


Figure 2. Comparison of (a) MT-based SSL with (b) SemiCVT.

mance, due to lack of spatial inductive-bias in modelling local cues [27]. Despite the combination of Transformer and CNN has proven to be effective [7, 16, 27, 36], the integration of Transformer with CNN-based Mean-Teacher SSLs remains the fundamental problem for several reasons: **(1) Intra-model problem**: The feature paradigm of Transformer is heterogeneous compared to CNNs. Additionally, Transformer relies on the large-scale pre-trained model with customized fine-tuning for different downstream tasks, which consumes enormous time and energy. How to efficiently combine the complementary of the two-style features and train the Transformer with relatively little labeled data from scratch remains an open question. **(2) Inter-model problem**: The existing MT-based SSLs merely leverage the pixel-wise predictions from the teacher model for guiding the student model to approximate, which ignores rich class-level information. How to make CNN and Transformer learn from each other on unlabeled data in class-level is a problem worthy of exploring.

To tackle these problems, we propose Semi-Supervised Convolutional Vision Transformer (termed as SemiCVT in Fig. 2. (b)), which fully combines CNN and Transformer for semi-supervised segmentation motivated by **(1) Intra-model local-global interaction**: Considering the heterogeneous paradigm of CNN and Transformer in the spatial domain, we alternatively investigate the interaction of CNN and Transformer in the Fourier domain [28], since learning on frequency spectrum is able to steer all the frequencies to capture both long-term and short-term interactions. In this way, both contextual details in CNN and long-range dependency in Transformer can be extracted with better local-global interaction. **(2) Inter-model class-wise consistency**: CNN and Transformer have different inner feature flow forms, in which their feature maps with complementary class-wise statistics creates a potential opportunity for incorporation. Inspired by such an observation, we utilize the class-wise statistics of unlabeled data generated by CNN/Transformer (from teacher) to update the parameters of the Transformer/CNN (from student), respectively. In a cross-teaching manner, we learn an implicit consistency regularization with complementary cues in graph domain, which can produce more stable and accurate pseudo labels.

The ability of SemiCVT in capturing local-global cues and class-specific characteristics is shown in Fig. 1. Compared with the MT-based SSL, SemiCVT can attend to full object extent with in various sizes and long-range scenarios (*e.g.,* full extent of the people's *leg* in Fig. 1. (b), as well as the feature discriminability between different classes (*e.g.,* activated small-size *bottle* in Fig. 1. (f)), achieving accurate segmentation shown in Fig. 1 (d) and (h). In summary, the main contributions of this work are fourfold: (**i**) We analyze the intra-model and inter-model problems faced by the existing CNN-based Mean-Teacher methods for semi-supervised segmentation, and propose a novel scheme, named SemiCVT, to fully capitalize the unlabeled data. (**ii**) We introduce an intra-model local-global interaction strategy for chaining both CNN and Transformer in the Fourier domain. (**iii**) We propose an inter-model class-wise consistency to learn complementary class-level statistics in a cross-teaching manner. (**iv**) Extensive experiments are performed on two public datasets, resulting in the new state-of-the-art performances consistently.

2. Related Work

Semi-supervised Learning in Segmentation. Semi-supervised learning improves the representation learning by leveraging numerous unlabeled data. Consistency regularization methods [1, 13, 38] refer to the consistent outputs of model under different perturbations, which is usually adopted by the Mean-Teacher [33] scheme. Based on the MT-SSL, U^{2}PL [35] further makes full use of the unreliable pseudo labels. However, these CNN-based SSLs experience difficulty to capture long-range dependencies, which is critical for accurate semantic segmentation.

Hybrid Modeling of CNNs and Transformers. Recent works [7, 16, 27, 36] show the advantages of combining CNN and Transformer. CMT [16] designs a local perception unit with a light-weight Transformer block. By considering the interactions, Mixformer [5] builds spatial and channel attentions between local window attention and depthwise convolution, while Conformer [27] designs a simple down/up sampling technique to eliminate the feature misalignment. Different from existing methods that learn the complementary information in spatial domain, we explore the proper interaction design in Fourier domain.

Fourier Transform in Computer Vision. Recent years have witnessed increasing research introducing Fourier transform into deep learning method for vision tasks [8, 23, 28, 40]. For example, FFC [8] proposes a local fourier unit that utilized both spatial and spectral information for achieving mixed receptive fields. GFNet [28] explores the long-term spatial dependencies in the Fourier domain with log-linear complexity. Inspired by that Fourier transform has no learnable parameters and its capacity in focusing on the all locations in spatial domain, we design a simple yet effective interactive module in Fourier domain.

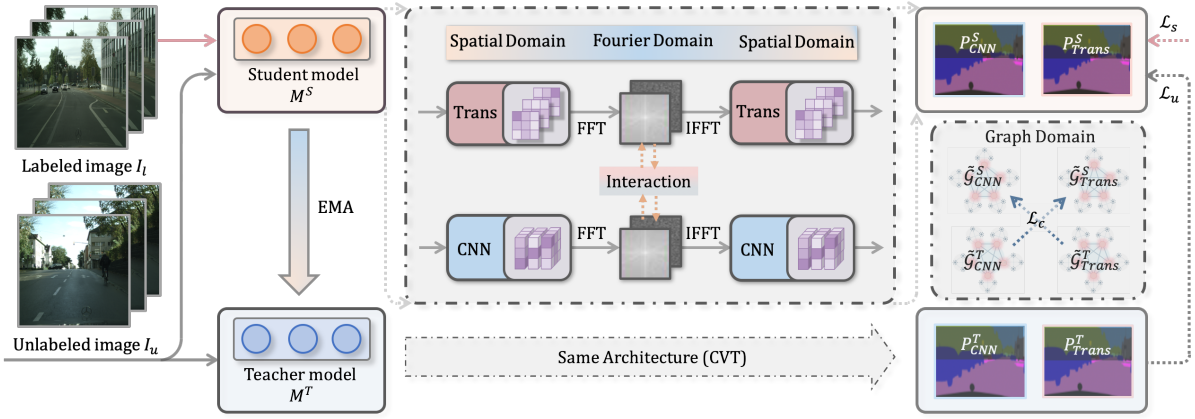


Figure 3. **An overview of our SemiCVT.** The student model takes both labeled images (I_l) and unlabeled images (I_u) as input, and obtains two-style predictions (P^S_{CNN} and P^S_{Trans}) after feature extraction by a parallel CNN-Transformer architecture (CVT), embedded with the intra-model local-global interaction module (LGI) in each stage for full integration. The student model and teacher model share the same architecture, and the student learns from the teacher by minimizing the supervised loss (L_s) on labeled data (I_l), the pixel-wise unsupervised loss (L_u) and the inter-model class-wise consistency loss (L_c) on unlabeled data (I_u).

3. Method

In the setting of a semi-supervised segmentation task, two sources of data are available: a labeled set $D_L = (I_l, Y_l)_{l=1}^{N}$, and a relatively larger unlabeled set $D_U = (I_u)_{u=N+1}^{N+M}$. Our goal is to design a simple yet effective semi-supervised learning strategy by leveraging both the labeled and unlabeled data. The overview of our SemiCVT is shown in Fig. 3, which consists of two key designs upon the standard MT-based SSL: (**1**) A parallel CNN-Transformer architecture, namely CVT (in Sec. 3.1), is proposed with the intra-model local-global interaction (LGI) in Fourier domain for full integration (in Sec. 3.2); (**2**) The inter-model consistency is improved by extracting complementary class statistics in graph domain (in Sec. 3.3). These two components are proposed to explore global learning capability and class-level features from unlabeled data, respectively.

Our CVT is composed of two branches: a CNN branch (f_{CNN}) with local-level details and a Transformer branch (f_{Trans}) with global-level information. Given the input images I_l and I_u, f_{CNN} and f_{Trans} extract the feature x_{CNN} and x_{Trans}, respectively. To learn the complementary information between the two branches, LGI is introduced into each stage. The module firstly converts x_{CNN} and x_{Trans} from spatial domain into X_{CNN} and X_{Trans} in Fourier domain for full interaction. After that, the enhanced features are transformed back to the spatial domain. Eventually, each branch is appended with an individual segmentation head to yield predictions (P_{CNN} and P_{Trans}).

To fully exploit the complementary class statistics embedded in CNN and Transformer, we propose a class-wise consistency loss $\mathcal{L}_c$ in a cross-teaching manner. It enforces the class-level distribution of CNN/Transformer from student and Transformer/CNN from teacher ($\widetilde{\mathcal{G}}^T_{CNN} \rightarrow \widetilde{\mathcal{G}}^S_{Trans}$, $\widetilde{\mathcal{G}}^T_{Trans} \rightarrow \widetilde{\mathcal{G}}^S_{CNN}$) to be closed in graph domain, which effectively improves the robustness of our model.

3.1. Architecture of CVT

The proposed CVT exploits the complementary characteristics of CNN and Transformer, as shown in Fig. 4. Each component of CVT will be introduced in detail as follows.

Stem Module and Segmentation Heads. Our CVT has a stem module, which consists of three successive 3×3 convolution layers followed by a batch normalization (BN) [20], a ReLU activation function, and a max pooling with stride of 2, to extract initial features which are fed to the dual branches. After feature extraction, the segmentation head proposed by DeepLabV3+ [4] is implemented for each branch. For inference, we simply average the predictions from two branches as our final results.

CNN Branch. For CNN branch, we utilize the ResNet-101 [18] as backbone to extract local-level features with detailed contexture. According to the architecture of ResNet-101, there are four stages with stacking convolution blocks. When CNN goes deeper, the number of channel dimensions gradually increases ($256, 512, 1024, 2048$ for four stages, respectively) while the resolution of feature maps decreases $1/4, 1/8, 1/16, 1/16$, respectively.

Transformer Branch. The Transformer branch is parallel to the CNN branch, where the size of Transformer feature is consistent to CNN feature in each stage. In particular, we employ the SwinUNet [**?**] (without pre-training) as the Transformer branch, which is composed of the light-weight Multi-Head Self-Attention (MHSA) and the Multi-Layer Perceptron (MLP) for feature extraction. To quickly train Transformer from scratch, we add the outputs of CNN and Transformer as the input of the following stage. The

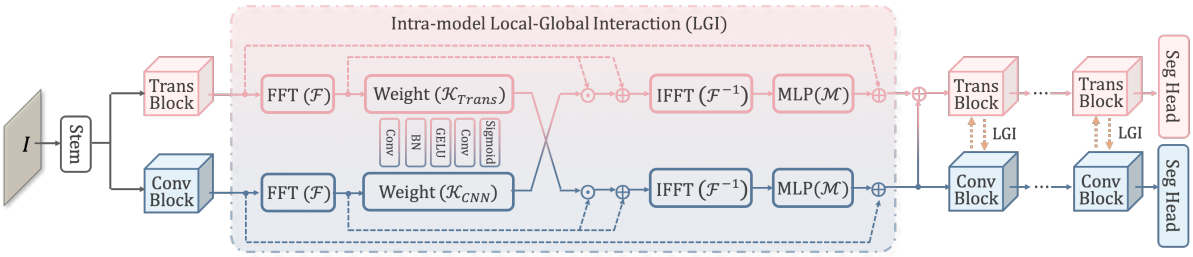


Figure 4. **The overall architecture of CVT.** There are four parts in CVT: a **Stem Module** with successive convolutions to increase the channels of input image, **Dual Branches** with Convolution (Conv) and Transformer (Trans) blocks, **LGI** to interact dual branches in each stage, and two **Segmentation Heads** for the dual branches.

stack of feature maps x^i_{Trans} can be represented as:

$$x^i_{Trans} = \begin{cases} Proj\left(x^{(i-1)}_{CNN}\right), i = n_s, \\ \mathcal{P}\left(x^{(i-1)}_{CNN} + x^{(i-1)}_{Trans}\right), i = n_s + 1, \cdots, 4, \end{cases} \tag{1}$$

where $Proj(\cdot)$ is the projection function, and $\mathcal{P}(\cdot)$ is the patch embedding mechanism. Both of them are realized by a 1×1 convolution followed by BN and ReLU function. n_s denotes the stage that we start to incorporate the Transformer block. Fig. 4 illustrates that we start at the first stage, *i.e.*, $n_s = 1$. In the experiment, our CVT achieves the best performance when $n_s = 2$ (in Sec. 4.3).

3.2. Intra-model Local-Global Interaction

According to spectral convolution theorem in Fourier theory, point-wise update in Fourier domain globally affects all input features [8]. Hence, learning on frequency spectrum is able to steer all the frequencies to capture both long-term and short-term interactions [28]. Inspired by this, we explore the interaction in Fourier domain, which is different from the previous interactions [7, 16, 27] devoted to the spatial domain. The overall architecture of LGI is depicted in Fig. 4, and we will describe each step in the following.

Fast Fourier Transform (FFT). FFT plays an important role in digital signal processing and is the first step of our LGI. The original spatial features x_{CNN} (x_{Trans}) are converted into Fourier domain X_{CNN} (X_{Trans}) via:

$$X_{CNN} = \mathcal{F}[x_{CNN}], X_{Trans} = \mathcal{F}[x_{Trans}], \tag{2}$$

where X_{CNN} (X_{Trans}) is the spectrum of x_{CNN} (x_{Trans}) with real and imaginary part. Benefiting from the conjugate symmetric property of Discrete Fourier Transform (DFT), we can keep only half of the values in X and trivially remove the other half without losing important information [28]. In this way, X_{CNN} (X_{Trans}) has the reduced size of $H \times \lceil W/2 \rceil \times D$ for computational efficiency.

Interactive Weights ($\mathcal{K}$). As proved in [28], multiplying the spectrum with a global weight can efficiently exchange spatial information. Our LGI follows this trend and learns the interactive weight $\mathcal{K}_{CNN}$/$\mathcal{K}_{Trans}$ from CNN/Transformer branch to modulate the representation of Transformer/CNN branch, respectively. As seen in Fig. 4, we adopt a simple design to generate two interactive weights $\mathcal{K}_{CNN}, \mathcal{K}_{Trans} \in \mathbb{R}^{H \times \lceil W/2 \rceil \times 1}$, which consists of two 1×1 convolution layers, BN [20], GELU [19] and sigmoid layers. Following the practice in [8], we only utilize the real part of the complex to calculate the weight for computational efficiency. We can then modulate the spectrum X_{CNN} (X_{Trans}) by multiplying $\mathcal{K}_{Trans}$ and ($\mathcal{K}_{CNN}$):

$$\tilde{X}_{CNN} = \mathcal{K}_{Trans} \odot X_{CNN}, \ \tilde{X}_{Trans} = \mathcal{K}_{CNN} \odot X_{Trans}. \tag{3}$$

Inverse FFT (IFFT) and MLP. IFFT ($\mathcal{F}^{-1}$) converts the enhanced features $\widetilde{X}_{CNN}$ ($\widetilde{X}_{Trans}$) with their original spectrum back into $\widetilde{x}_{CNN}$ ($\widetilde{x}_{Trans}$) in the spatial domain:

$$\tilde{x}_{CNN} = \mathcal{F}^{-1}[\tilde{X}_{CNN} + X_{CNN}] \in \mathbb{R}^{H \times W \times D}. \tag{4}$$

Similarly, we can obtain the transferred $\tilde{x}_{Trans}$. Finally, MLP ($\mathcal{M}$) is used as the channel mixer and combines with the original input x_{CNN} (x_{Trans}) to form a residual path.

Discussion. We give deep insight of our LGI with Fourier transform. According to convolutional theorem [28], multiplication in Fourier domain is equivalent to the *circular convolution* with the filter size of $H \times W$ in spatial domain. Inspired by this theorem, our interactive weight in Fourier domain can be treated as a global convolution filter with learned complementary cues in spatial domain, while multiplication is easier to be implemented brought in less computational cost than the *circular convolution*.

3.3. Inter-model Class-wise Consistency

In this section, we investigate the class-wise distribution in a global level, and thus build a bi-level graph-based prototype, which explores the class-patch and class-class relations in the graph domain. Equipped with different feature extractors, CNN and Transformer have distinct class-level statistics. Hence, we employ the cross teaching strategy with implicit consistency regularization, which can produce more stable and accurate pseudo labels [26]. In the following, we describe the process in details.

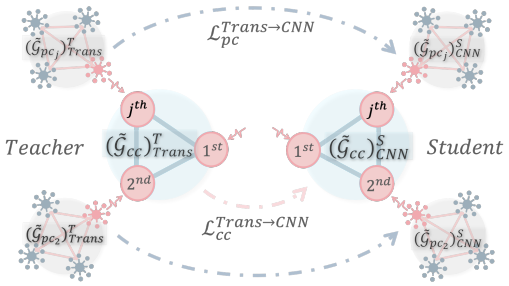


Figure 5. An illustration of the cross-guided consistency between Transformer of teacher and CNN of student with the bi-level graph as the prototype. $\widetilde{\mathcal{G}}_{pc}$ denotes the class-patch graphs, $\widetilde{\mathcal{G}}_{cc}$ represents the class-class graphs built upon the enhanced $\widetilde{\mathcal{G}}_{pc}$.

Class Nodes ($\mathcal{N}_c$) and Patch Nodes ($\mathcal{N}_p$). The center of each class can be calculated by taking the average of all feature pixels of the same class. Given an unlabeled image I_u, the student network can generate two features x^S_{CNN}, x^S_{Trans} from the last stage (before the segmentation head), and their corresponding predictions P^S_{CNN}, P^S_{Trans}. Then, we aggregate the representations of all pixels weighted by their probabilities belonging to the j-th object:

$$\mathcal{N}_{c_j} = \sum_{i \in I_u} x^i P^i_j, \tag{5}$$

where x^i is the feature of the i-th pixel and P^i_j is the normalized probability for the i-th pixel belonging to the j-th class. Note that $\|\mathcal{N}_c\|$ is the number of class nodes, which equals to the number of semantic categories. In this way, we can obtains the class node $(\mathcal{N}_c)^S_{CNN}$, $(\mathcal{N}_c)^S_{Trans}$ from the student, and $(\mathcal{N}_c)^T_{CNN}$, $(\mathcal{N}_c)^T_{Trans}$ from the teacher.

Instead of directly regarding each pixel as a patch node, we use the parameter-free average-pooling with stride r on the feature map (*e.g.,* x^S_{CNN}). Hence, the number of patch nodes $\|\mathcal{N}_p\|$ is reduced into $H/r \times W/r$. Accordingly, we can obtain the patch nodes $(\mathcal{N}_p)^S_{CNN}$, $(\mathcal{N}_p)^S_{Trans}$ from the student, and $(\mathcal{N}_p)^T_{CNN}$, $(\mathcal{N}_p)^T_{Trans}$ from the teacher.

Class-Patch Graph Construction. To capture the correlations among class centers and patch nodes, we construct a graph $\mathcal{G}_{pc}$ to model the class-patch dependencies, which is a flexible way to capture the topological structure. Specifically, for each class (*e.g.,* the j-th class), we combine their class node $\mathcal{N}_{c_j}$ with all patch nodes $\mathcal{N}_p$ to diffuse the class-specific information (as depicted in the top left corner of Fig. 5). Hence, a lightweight fully-connected graph $\mathcal{G}_{pc_j} = \left\{\left[\mathcal{N}_p, \mathcal{N}_{c_j}\right]\right\}$ with an adjacency matrix $\mathcal{A}_{pc_j} \in \mathbb{R}^{(\|\mathcal{N}_p\|+1)\times(\|\mathcal{N}_p\|+1)}$ is generated for the j-th class, where $[\cdot]$ is the concatenation operation. The adjacency matrix is defined as the similarity between nodes [43] by:

$$\mathcal{A}_{pc_j} = softmax\left(x_{pc_j} \otimes x^T_{pc_j}\right), \tag{6}$$

where x_{pc_j} is the representation of graph $\mathcal{G}_{pc_j}$ and $\otimes$ is the matrix multiplication. The $softmax(\cdot)$ operation is utilized to yield a normalized adjacency matrix. Then we conduct graph convolution [22] to diffuse information by:

$$\widetilde{\mathcal{G}}_{pc_j} = \mathcal{A}_{pc_j} x_{pc_j} \mathcal{W}_{pc_j}, \tag{7}$$

where $\mathcal{W}_{pc_j} \in \mathbb{R}^{D\times D}$ is a weight matrix. The $\widetilde{\mathcal{G}}_{pc_j}$ is the enhanced feature map after graph convolution.

Class-Class Graph Construction. Since the dependencies among semantic categories are essential for context modeling, we further construct a graph $\mathcal{G}_{cc}$ to explore the correlations among class centers. As shown in Fig. 5, the graph $\mathcal{G}_{cc}$ is built over the enhanced class nodes $\widetilde{\mathcal{N}}_c$ extracted from graph $\widetilde{\mathcal{G}}_{pc}$. Similarly, we also calculate the adjacency matrix $\mathcal{A}_{cc}$ in Eq. (6) and adopt the graph convolution in Eq. (7) to learn the relations and thus obtain an enhance graph $\widetilde{\mathcal{G}}_{cc}$.

Cross-guided Class-wise Consistency Loss $\mathcal{L}_c$. Based on the enhanced graphs from either student or teacher, we simply employ the Mean Squared Error (MSE) loss to learn the complementary information as follows:

$$\mathcal{L}^{Trans\to CNN}_{pc} = MSE\left((\widetilde{\mathcal{G}}_{pc})^T_{Trans}, (\widetilde{\mathcal{G}}_{pc})^S_{CNN}\right), \tag{8}$$

$$\mathcal{L}^{Trans\to CNN}_{cc} = MSE\left((\widetilde{\mathcal{G}}_{cc})^T_{Trans}, (\widetilde{\mathcal{G}}_{cc})^S_{CNN}\right). \tag{9}$$

In this process, the gradient of the teacher model is restrained. After that, we obtain $\mathcal{L}^{Trans\to CNN}_c$ by adding $\mathcal{L}^{Trans\to CNN}_{pc}$ and $\mathcal{L}^{Trans\to CNN}_{cc}$. Similarly, we can get $\mathcal{L}^{CNN\to Trans}_c$. The final $\mathcal{L}_c$ can be further calculated by averaging $\mathcal{L}^{Trans\to CNN}_c$ and $\mathcal{L}^{CNN\to Trans}_c$.

3.4. SemiCVT Framework

The overview of our SemiCVT is presented in Fig. 3, which follows the Mean-Teacher scheme and is optimized using the following loss: $\mathcal{L} = \lambda_s\mathcal{L}_s + \lambda_u\mathcal{L}_u + \lambda_c\mathcal{L}_c$, where $\mathcal{L}_s$ is a supervised loss on labeled data, $\mathcal{L}_u$ is a unsupervised loss and $\mathcal{L}_c$ is our proposed inter-model class-wise consistent loss (in Sec. 3.3). Both $\mathcal{L}_u$ and $\mathcal{L}_c$ are applied on unlabeled data. λ_s, λ_u and λ_c are corresponding constraints for balancing the training. We employ the Cross-Entropy (CE) loss to calculate $\mathcal{L}_s$ and $\mathcal{L}_u$, which can be formulated as:

$$\mathcal{L}_s = CE\left(M(I_l), Y_l\right), \ \mathcal{L}_u = CE\left(M(I_u), \hat{Y}_u\right). \tag{10}$$

where Y_l is the ground truth of labeled data, and $\hat{Y}_u$ is the pseudo-label of unlabeled data. We apply $\mathcal{L}_s$ and $\mathcal{L}_u$ on both two predictions from CNN and Transformer branches.

4. Experiments

4.1. Setup

Datasets. Experiments are conducted on two public datasets: **(1) PASCAL VOC** [12] is a natural scene dataset with 21 semantic classes. It originally consists of $1,464$ training and $1,449$ validation images. The training set can be augmented via adopting coarsely annotated $9,118$ images from the SBD dataset [17], resulting in $10,582$ training

Table 1. Comparison with SOTA methods on *classic* PASCAL VOC dataset under different partitions.

Method		1/16(92)	1/8(183)	1/4(366)	1/2(732)	Full(1464)
Suponly (Sup)	Baseline	45.77	54.92	65.88	71.69	72.50
	CVT	46.96	59.74	68.38	73.68	74.55
Semi-supervised (SSL)	MT [33]	51.72	58.93	63.86	69.51	70.96
	CutMix [42]	52.16	63.47	69.46	73.73	76.54
	PseudoSeg [47]	57.60	65.50	69.14	72.41	73.23
	PC^2Seg [45]	57.00	66.28	69.78	73.05	74.15
	ST++ [39]	65.22	67.45	72.33	75.37	78.06
	U^2PL [35]	67.98	69.15	73.66	76.16	79.49
	$SemiCVT^-$	68.21	70.32	74.02	77.23	79.96
	SemiCVT	**68.56**	**71.26**	**74.99**	**78.54**	**80.32**

Table 2. Comparison with SOTA methods on *blender* dataset.

Method		1/16 (662)	1/8 (1323)	1/4 (2646)	1/2 (5291)
Sup	Baseline	67.87	71.55	75.80	77.13
	CVT	72.36	74.32	77.88	78.89
SSL	MT [33]	70.51	71.53	73.02	76.58
	CutMix [42]	71.66	75.51	77.33	78.21
	CCT [26]	71.86	73.68	76.51	77.40
	GCT [21]	70.90	73.29	76.66	77.98
	CPS [6]	74.48	76.44	77.68	78.64
	ST++ [39]	73.85	76.55	78.23	79.51
	U^2PL [35]	77.21	79.01	79.30	80.50
	$SemiCVT^-$	77.50	79.43	79.77	80.69
	SemiCVT	**78.20**	**79.95**	**80.20**	**80.92**

Table 3. Comparison with SOTA methods on Cityscapes dataset.

Method		1/16 (186)	1/8 (372)	1/4 (744)	1/2 (1488)
Sup	Baseline	65.74	72.53	74.43	77.83
	CVT	67.17	73.10	75.12	78.55
SSL	MT [33]	69.03	72.06	74.20	78.15
	CutMix [42]	67.06	71.83	76.36	78.25
	CCT [26]	69.32	74.12	75.99	78.10
	GCT [21]	66.75	72.66	76.11	78.34
	CPS [6]	69.78	74.31	74.58	76.81
	ST++ [39]	67.64	73.43	74.64	77.78
	U^2PL [35]	70.30	74.37	76.47	79.05
	$SemiCVT^-$	71.24	74.94	76.82	79.18
	SemiCVT	**72.19**	**75.41**	**77.17**	**79.55**

images. Following [35], we evaluate our method on both the *classic* dataset ($1,464$ labeled images with $9,118$ unlabeled images), and the *blender* dataset ($10,582$ labeled images). **(2) Cityscapes dataset** [9] finely annotates 19 object categories in real urban scenes, with $2,975$ and 500 images for training and validation, respectively.

Each dataset is divided into $1/16$, $1/8$, $1/4$, $1/2$ images as labeled, and the rest of training images as the unlabeled data, following the same settings as U^2PL [35]. Note that in the *classic* dataset, the labeled data is only sampled from $1,464$ labeled images, while the rest of images is combined with $9,118$ additional images as the unlabeled data.

Implementation Details. We itemize the crop size (cs), batch size (bs), learning rate (lr), training epochs (ep), optimizer (opt) for each dataset: (1) *Classic* and *blender* PASCAL VOC: cs=513 $\times$ 513; bs=16; lr=0.001; ep=80; opt=SGD. (2) Cityscapes dataset: cs=769 $\times$ 769; bs=16; lr=0.01; ep=200; opt=SGD. Following the common practice [35], the base learning rate of the decoder is ten times that of the backbone, and we also utilize the CutMix [42] strategy. The weight of supervised loss λ_s and unsupervised loss λ_u are set to 1. The number of patch nodes $\|\mathcal{N}_p\|$ is set to 100, while the number of class nodes $\|\mathcal{N}_c\|$ is the same as the semantic categories in each dataset.

Evaluation. We leverage the mean Intersection-over-Union (mIOU) as our evaluation metric. Following the previous methods [13,26,35,46], our models are evaluated on the validation set of PASCAL VOC and Cityscapes. To reduce the randomness, all hyper-parameters (in Sec. 4.3) and ablation studies (in Sec. 4.4) are conducted under two proportions, i.e., 1/2 and 1/4, on the *classic* PASCAL VOC.

4.2. Comparison with the State-of-the-Arts

We compare our SemiCVT with the following SOTA methods: MT [33], CCT [26], GCT [21], PseudoSeg [47], Cut-Mix [42], CPS [6], PC^2Seg [45], ST++ [39] and U^2PL [35]. We re-implement ST++ for a fair comparison. In these methods, we use the same ResNet-101 backbone pretrained on ImageNet [10] with DeepLabV3+ as decoder, termed as Baseline, which is consistent with our CNN branch.

Results on *classic* PASCAL VOC Dataset. Table 1 lists the performance of different comparison methods on the *classic* dataset. We first show the fully-supervised performance achieved by training on the limited labeled data, namely SupOnly (Sup). As seen, our CVT outperforms the CNN-based supervised baseline by +1.19%, +4.82%, +2.5%, +1.99% and +2.05% under 1/16, 1/8, 1/4, 1/2 partitions and full supervision, respectively. Interestingly, CVT even surpasses the MT-based SSL (trained with additional unlabeled data), under 1/16, 1/8, 1/4 and 1/2 partitions. It demonstrates that our CVT can extract the complementary information from CNN and Transformer, which achieves satisfactory performance even with limited labeled data.

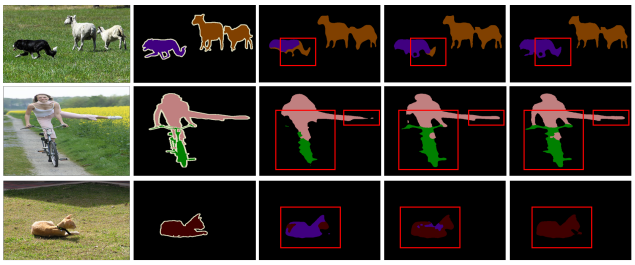


Figure 6. Visual comparisons on **PASCAL VOC 2012** dataset.

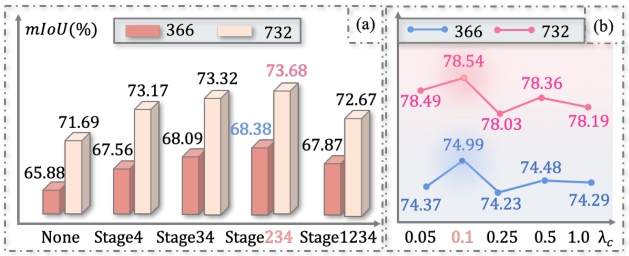


Figure 7. (a) Impact of the insertion location of Transformer block at different stages. (b) Performance of SemiCVT w.r.t. λ_c.

In the semi-supervised setting, we compare with the methods that were evaluated on the *classic* dataset in their paper. As seen, our SemiCVT without inter-model class-wise consistency (SemiCVT$^-$) already outperforms existing methods in all partitions. Moreover, by considering the class statistics, SemiCVT further boosts the performance consistently, establishing a new SOTA. In particular, compared with the MT-based SSL, our SemiCVT achieves impressive improvements in five partitions: +16.84%, +12.33%, +11.13%, +9.03% and +9.36%, respectively.

Results on *blender* PASCAL VOC Dataset. Table 2 reports the comparison results on *blender* PASCAL VOC. More semi-supervised approaches are compared, for example, CCT [26], GCT [21], CPS [6], and ST++ [39]. We can see that our method still achieves the best segmentation performance, with improvements of +0.99%, +0.94%, +0.90%, and +0.42% over the previous best model U^{2}PL.

Results on Cityscapes Dataset. Table 3 presents the comparison results on the Cityscapes dataset. Benefited from the powerful capability of feature extraction, SemiCVT achieves the highest performance. The improvements of our method over the previous best U^{2}PL are 1.89%, 1.04%, 0.70% and 0.50% under the partition of 1/16, 1/8, 1/4 and 1/2, respectively. These quantitative results on three datasets substantiate the fine robustness of our SemiCVT.

Visual Comparison. Fig. 6 illustrates the qualitative results on PASCAL VOC. As seen, our SemiCVT$^-$ is able to segment both tiny objects with touching boundaries and large object with fine structures, by learning both local and global information. By considering the class statistics, our SemiCVT can distinguish cluttered foreground and background, which achieves much neater and cleaner results.

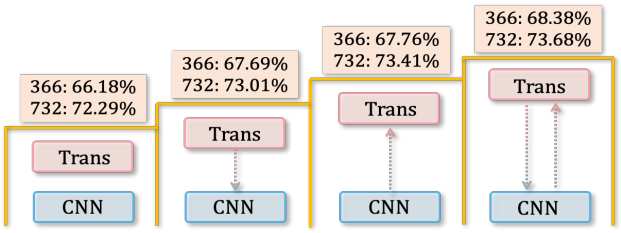


Figure 8. Ablation of the interactions in the LGI Module.

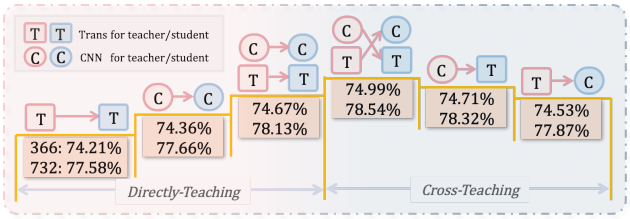


Figure 9. Comparison of directly-teaching strategy with cross-teaching strategy between Transformer (Trans) and CNN.

4.3. Hyper-Parameters

Start Stage of Transformer n_s. We explore the influence of integrating CNN and Transformer blocks at different stages under two partitions on the *classic* dataset. In Fig. 7 (a), the accuracy is improved by stacking the Transformer blocks with interacted LGI, which achieves the highest results when extracting complementary cues on three successive stages (Stage 2, 3 and 4). However, applying the Transformer block on Stage 1 may affect the genuine characteristics and increase the computational cost. Therefore, in our experiments, we start by introducing the Transformers block on Stage 2, which provides the best results within a reasonable computational complexity.

Adjustment of λ_c. As mentioned in Sec. 3.4, λ_c is a coefficient that controls the inter-model class-wise consistency. In Fig. 7 (b), mIoU is improved with the increase of λ_c. However, the accuracy starts to decline at λ_c=0.1 under both partitions. Such results indicate that the larger λ_c may hinder the feature learning in semi-supervised segmentation.

4.4. Ablation Study

Ablation of Interactions in LGI Module. We begin by assessing the influence of interactions in our LGI module. In Fig. 8, the consistent improvements (366: +1.51%, 732: +0.72%) are achieved when introducing the interaction from Transformer to CNN. Interestingly, a better performance is achieved when using the interaction from CNN to Transformer. By combining two interactions, we obtain the best results (366: 68.38%, 732: 73.68%) through interacting fine-grained details and coarse-grained semantics.

Effectiveness of Our LGI Module. We compare our LGI module with two interaction modules proposed by Con-

Table 4. Comparison of our LGI module (in Fourier domain) with other interaction methods (in spatial domain).

Methods (Supervised)	1/4(366)	1/2(732)
baseline (ResNet101)	65.88	71.69
+ Interaction in Conformer [27]	66.35	72.37
+ Interaction in MixFormer [5]	66.63	72.56
+ our LGI	**68.38**	**73.68**

Table 5. Ablation study of bi-level graph-based prototypes in $\mathcal{L}_c$.

#	class-patch		class-class		1/4	1/2
	$\mathcal{G}_{pc}$	$\tilde{\mathcal{G}}_{pc}$	$\mathcal{G}_{cc}$	$\tilde{\mathcal{G}}_{cc}$	(366)	(732)
1	×	×	×	×	74.02	77.23
2	✓	×	×	×	74.42	77.72
3	×	✓	×	×	74.63	78.04
4	×	×	✓	×	74.34	77.58
5	×	×	×	✓	74.46	77.91
6	×	✓	×	✓	**74.99**	**78.54**

Former and MixFormer in Table 4. These are two typical works that perform feature interaction in the spatial domain. Comparatively, our method achieves the best accuracy on both partitions. It indicates the superiority of LGI Module via interacting two-style features in Fourier domain.

Ablation of Bi-level Graphs in $\mathcal{L}_c$. In Table. 5, we explore the influence of bi-level graph-based prototypes in the class-wise consistency loss $\mathcal{L}_c$. We observe that using the original class-patch graphs $\mathcal{G}_{pc}$ (#2) and class-class graphs $\mathcal{G}_{cc}$ (#4) to SemiCVT can achieve consistent improvements, due to the introduction of the class-wise statistics. In addition, the enhanced $\tilde{\mathcal{G}}_{pc}$ (#3) and $\tilde{\mathcal{G}}_{cc}$ (#5) with the assistance of graph convolution can further boost the performance. By combining both $\tilde{\mathcal{G}}_{pc}$ and $\tilde{\mathcal{G}}_{cc}$ (#6), the best results are achieved with +1.31% and +0.97% under 1/2 and 1/4 partitions.

Effectiveness of Cross-Teaching Strategy. We provide a comprehensive ablation study to assess the effectiveness of the proposed cross-teaching strategy, which consists of $T \rightarrow C$ (from Transformer to CNN), $C \rightarrow T$ and both. Different from the cross-teaching strategy, the direct-teaching strategy performs the explicit consistency regularization, including $C \rightarrow C$, $T \rightarrow T$ and both. As a result, the best performance can be achieved when cross-guidance implemented in both $C \rightarrow T$ and $T \rightarrow C$, which demonstrates that CNN and Transformer with different learning paradigm can compensate each other in the training stage.

4.5. Interpretation of SemiCVT

Distribution of Deeply Learned Features. As shown in Fig. 10, the learned pixel embeddings by SemiCVT become more compact and well separated, which indicates that the designed class-wise consistency benefits the discriminative power of deeply learned features, which is crucial for semi-supervised segmentation.

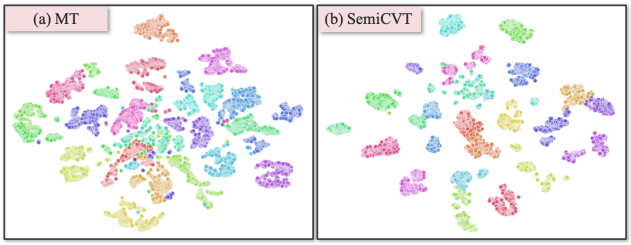
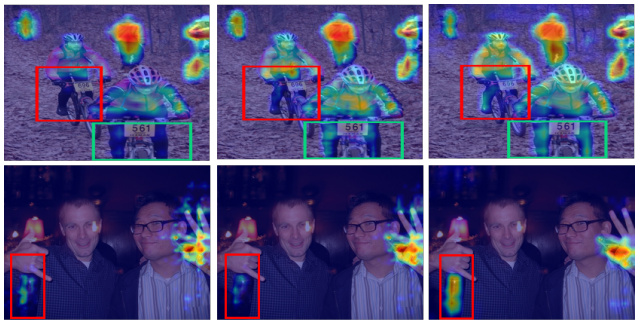


Figure 10. t-SNE visualization of deep feature representations extracted from MT (a) and Our SemiCVT(b) on PASCAL VOC.



(a) MT (b) U^{2}PL (c) SemiCVT

Figure 11. Visualizations of class activation maps on different methods by using Grad-CAM [31].

Visualization of Feature Maps. We visualize the class activation maps of different methods in Fig. 11. Compared with CNN-based MT and U^{2}PL, SemiCVT inherits the advantages of both retaining local features and capturing global dependency (*e.g.,* fully extent of rider) and tiny objects (*e.g.,* activated bottle in the last row).

5. Conclusion

In this paper, we proposed a novel semi-supervised learning scheme, termed as SemiCVT, which tackles intra-model and inter-model problems faced by MT-based SSL. Specifically, we designed a novel interaction module by incorporating both local representations and global cues in Fourier domain. Further, an implicit class-wise consistency regularization modeled in graph domain was introduced to make the pseudo label more accurate and stable. Our method was extensively evaluated on two public datasets and consistently outperformed other SSL approaches.

6. Acknowledgments

This work was supported in part by Zhejiang Provincial Natural Science Foundation of China (No. LZ22F020012), Major Technological Innovation Project of Hangzhou (No. 2022AIZD0147), the National Key Research and Development Project (No. 2022YFC2504605), Major Scientific Research Project of Zhejiang Lab (No. 2020ND8AD01), and Japanese Ministry for Education, Science, Culture and Sports (No. 20KK0234, No. 21H03470 and No. 20K21821).

References

[1] Philip Bachman, Ouais Alsharif, and Doina Precup. Learning with pseudo-ensembles. *Advances in Neural Information Processing Systems*, 27, 2014. 2

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 1

[3] Huaian Chen, Yi Jin, Guoqiang Jin, Changan Zhu, and Enhong Chen. Semisupervised semantic segmentation by improving prediction confidence. *IEEE Transactions on Neural Networks and Learning Systems*, 2021. 1

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 801–818, 2018. 1, 3

[5] Qiang Chen, Qiman Wu, Jian Wang, Qinghao Hu, Tao Hu, Errui Ding, Jian Cheng, and Jingdong Wang. Mixformer: Mixing features across windows and dimensions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5249–5259, 2022. 2, 8

[6] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2613–2622, 2021. 6, 7

[7] Yinpeng Chen, Xiyang Dai, Dongdong Chen, Mengchen Liu, Xiaoyi Dong, Lu Yuan, and Zicheng Liu. Mobile-former: Bridging mobilenet and transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5270–5279, 2022. 2, 4

[8] Lu Chi, Borui Jiang, and Yadong Mu. Fast fourier convolution. *Advances in Neural Information Processing Systems*, 33:4479–4488, 2020. 2, 4

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016. 6

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009. 6

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[12] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88:303–338, 2010. 5

[13] Geoff French, Samuli Laine, Timo Aila, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, varied perturbations. *arXiv preprint arXiv:1906.01916*, 2019. 1, 2, 6

[14] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. *Advances in Neural Information Processing Systems*, 17, 2004. 1

[15] Frantiseli Grezl and Martin Karafiát. Semi-supervised bootstrapping approach for neural network feature extractor training. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 470–475. IEEE, 2013. 1

[16] Jianyuan Guo, Kai Han, Han Wu, Yehui Tang, Xinghao Chen, Yunhe Wang, and Chang Xu. Cmt: Convolutional neural networks meet vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12175–12185, 2022. 2, 4

[17] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *2011 International Conference on Computer Vision*, pages 991–998. IEEE, 2011. 5

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 3

[19] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 4

[20] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456. PMLR, 2015. 3, 4

[21] Zhanghan Ke, Di Qiu, Kaican Li, Qiong Yan, and Rynson WH Lau. Guided collaborative training for pixel-wise semi-supervised learning. In *European Conference on Computer Vision*, pages 429–445. Springer, 2020. 6, 7

[22] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 5

[23] James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontanon. Fnet: Mixing tokens with fourier transforms. *arXiv preprint arXiv:2105.03824*, 2021. 2

[24] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 1

[25] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015. 1

[26] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12674–12684, 2020. 4, 6, 7

[27] Zhiliang Peng, Wei Huang, Shanzhi Gu, Lingxi Xie, Yaowei Wang, Jianbin Jiao, and Qixiang Ye. Conformer: Local features coupling global representations for visual recognition.

In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 367–376, 2021. 2, 4, 8

[28] Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. Global filter networks for image classification. *Advances in Neural Information Processing Systems*, 34:980–993, 2021. 2, 4

[29] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. *arXiv preprint arXiv:2101.06329*, 2021. 1

[30] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Advances in Neural Information Processing Systems*, 29, 2016. 1

[31] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017. 1, 8

[32] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems*, 33:596–608, 2020. 1

[33] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, 2017. 1, 2, 6

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017. 1

[35] Yuchao Wang, Haochen Wang, Yujun Shen, Jingjing Fei, Wei Li, Guoqiang Jin, Liwei Wu, Rui Zhao, and Xinyi Le. Semi-supervised semantic segmentation using unreliable pseudo-labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4248–4257, 2022. 2, 6

[36] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22–31, 2021. 2

[37] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 1

[38] Yi Xu, Lei Shang, Jinxing Ye, Qi Qian, Yu-Feng Li, Baigui Sun, Hao Li, and Rong Jin. Dash: Semi-supervised learning with dynamic thresholding. In *International Conference on Machine Learning*, pages 11525–11536, 2021. 1, 2

[39] Lihe Yang, Wei Zhuo, Lei Qi, Yinghuan Shi, and Yang Gao. St++: Make self-training work better for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4268–4277, 2022. 6, 7

[40] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4085–4095, 2020. 2

[41] Yuhui Yuan, Lang Huang, Jianyuan Guo, Chao Zhang, Xilin Chen, and Jingdong Wang. Ocnet: Object context network for scene parsing. *arXiv preprint arXiv:1809.00916*, 2018. 1

[42] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6023–6032, 2019. 6

[43] Li Zhang, Xiangtai Li, Anurag Arnab, Kuiyuan Yang, Yunhai Tong, and Philip HS Torr. Dual graph convolutional network for semantic segmentation. *arXiv preprint arXiv:1909.06121*, 2019. 5

[44] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2881–2890, 2017. 1

[45] Yuanyi Zhong, Bodi Yuan, Hong Wu, Zhiqiang Yuan, Jian Peng, and Yu-Xiong Wang. Pixel contrastive-consistent semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7273–7282, 2021. 6

[46] Yuanyi Zhong, Bodi Yuan, Hong Wu, Zhiqiang Yuan, Jian Peng, and Yu-Xiong Wang. Pixel contrastive-consistent semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7273–7282, 2021. 6

[47] Yuliang Zou, Zizhao Zhang, Han Zhang, Chun-Liang Li, Xiao Bian, Jia-Bin Huang, and Tomas Pfister. Pseudoseg: Designing pseudo labels for semantic segmentation. *arXiv preprint arXiv:2010.09713*, 2020. 6