2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Transformer-CNN Cohort: Semi-supervised Semantic Segmentation by the Best of Both Students

Xu Zheng[1], *IEEE Student Member*, Yunhao Luo[4], Chong Fu[3], Kangcheng Liu[1], *IEEE Member*, Lin Wang[1,2†], *IEEE Member*

***Abstract*— The popular methods for semi-supervised semantic segmentation mostly adopt a unitary network model using convolutional neural networks (CNNs) and enforce consistency of the model's predictions over perturbations applied to the inputs or model. However, such a learning paradigm suffers from two critical limitations: a) learning the discriminative features for the unlabeled data; b) learning both global and local information from the whole image. In this paper, we propose a novel Semi-supervised Learning (SSL) approach, called Transformer-CNN Cohort (TCC), that consists of two students with one based on the vision transformer (ViT) and the other based on the CNN. Our method subtly incorporates the multi-level consistency regularization on the predictions and the heterogeneous feature spaces via pseudo-labeling for the unlabeled data. First, as the inputs of the ViT student are image patches, the feature maps extracted encode crucial class-wise statistics. To this end, we propose class-aware feature consistency distillation (CFCD) that first leverages the outputs of each student as the pseudo labels and generates class-aware feature (CF) maps for knowledge transfer between the two students. Second, as the ViT student has more uniform representations for all layers, we propose consistency-aware cross distillation (CCD) to transfer knowledge between the pixel-wise predictions from the cohort. We validate the TCC framework on Cityscapes and Pascal VOC 2012 datasets, which outperforms existing SSL methods by a large margin. Project page: https://vlislab22.github.io/TCC/.**

I. INTRODUCTION

Semantic segmentation [1], [2] is a crucial scene understanding task in computer and robotic vision, aiming to generate pixel-wise category prediction of an image. Most of the state-of-the-art (SoTA) methods focus on exploring the potential of convolutional neural networks (CNNs) and learning strategies [3], [4]. However, a hurdle of training these models is the lack of large-scale and high-quality annotated datasets, imposing much burden for real applications, *e.g.*, autonomous driving [5]. Consequently, growing attention has been paid to deep semi-supervised learning (SSL) for semantic segmentation [6] using the labeled data and additional unlabeled data.

The dominant deep SSL methods rely on consistency regularization [7], [8], pseudo labeling [9], entropy minimization [10] and bootstrapping [11], etc. However, these methods are only limited to classification, and their applications to semantic segmentation are still restricted [12]. Only recently, attempts have been made focusing on consistency-based SSL for semantic segmentation [13]. In these methods, the 'Teacher-Student' structure is often explored by creating a teacher model and a student model either explicitly or implicitly [14], [15]. The core spirit is to impose consistency on the predictions between two models via an exponential moving average (EMA) of the student and force the unlabeled data to meet the smooth assumption in SSL.

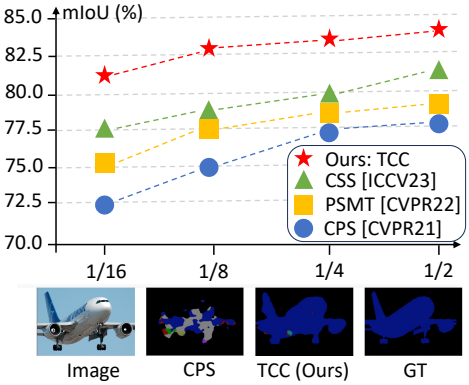


Fig. 1. **Increment over SoTA baseline** on the *PASCAL VOC* val set. Our TCC yields a noticeable improvement over previous methods, *e.g.*, CPS [6].

However, such a learning paradigm suffers from three key problems. First, using the isomorphic CNN-based models show limited learning capability of consistency regularization (See Tab. II). Previous works leveraged perturbations [16], different initialization [8] or different network structures [17] to impose model diversity. However, as the two feature extractors are inevitably coupled, it is difficult for them to extract complementary features in the later stage of training. Moreover, existing SSL methods merely leverage the pixel-wise predictions from CNNs, thus leading to a waste of rich inner knowledge in the feature space. Lastly, the input for SSL is always the entire image, making it difficult to learn both global and local (*i.e.*, long-range) semantic information although operated by strong data augmentation.

It has been shown that Vision Transformer (ViT) can achieve comparable or even superior performance on image recognition tasks at a large scale [18]. Differing from CNNs consisting of convolutions (Convs), ViT's basic computational paradigm is multi-head self-attentions (MHSAs). Park *et al.* [19] show that MHSAs and Convs exhibit opposite behaviors. That is, there exist surprisingly clear differences in the features and internal structures of ViT and CNN [20].

Motivation: Inspired by the success of ViT for visual recognition, in this paper, we explore the potential of ViT and CNNs to tackle the above-mentioned problems for semi-supervised segmentation. However, bringing the ViT to SSL is challenging because: a) the inner feature and output paradigm of ViT is heterogeneous from those of CNNs; b) the high-performance of ViT needs the pre-training with hundreds of millions of annotated images using a large infrastructure [21]; c) in SSL, how to make Convs and MSAs learn with each other for the *unlabeled* data from the feature and

†Corresponding author
[1]Xu zheng is with AI Thrust, HKUST(GZ), Guangzhou, China, Email: xzheng287@connect.hkust-gz.edu.cn
[4]Y. Luo is with the Department of Computer Science, Brown University, USA, Email: yluo73@cs.brown.edu
[3]C. Fu is with the Department of Computer Science and Engineering, Northeastern University, Shenyang, China, Email: fuchong@mail.neu.edu.cn
[1]K. Liu is with SMMG/ROAS Thrust, HKUST(GZ), Email: kangchengliu@hkust-gz.edu.cn
[1,2]L. Wang is with AI/CMA Thrust, HKUST(GZ) and Dept. of CSE, HKUST, China, Email: linwang@ust.hk

image level needs to be explored.

To this end, we propose a novel SSL method for semantic segmentation, called Transformer-CNN Cohort (**TCC**), by *subtly incorporating the multi-level distillation to add consistency on the pixel-wise predictions and the heterogeneous feature space via pseudo labeling for the unlabeled data*. Specifically, as the ViT and CNN students have different input and inner feature flow forms, we notice that the feature maps extracted encode crucial complementary class-wise statistics [22]. Therefore, we propose class-aware feature consistency distillation (**CFCD**) that first leverages the output of each student as the pseudo labels and generates pseudo prototype maps. Importantly, it is also the ***first*** time that we explore pseudo labeling in SSL to facilitate feature distillation for the unlabeled data. The class-aware feature (CF) maps are computed by averaging the features on all pixels having the same pseudo labels. Class-aware feature variation knowledge is transformed via the CF maps between the cohort. Moreover, as the ViT student has more uniform representations for all layers, we propose Consistency-aware Cross Distillation (**CCD**) to distill knowledge based on the pixel-wise predictions from the cohort. As such, we can reduce the large amount of training data required for ViT and accommodate ViT to SSL tasks with high performance. We conduct extensive experiments with various settings on two benchmarks: PASCAL VOC 2012 [23] and Cityscapes [24]. The experimental results show that our TCC framework surpasses the existing SoTA methods by 4.15% under 1/16 label ratio on POASCAL VOC 2012 dataset and 1.03% under 1/16 partition protocols on the CityScapes dataset.

Contributions: In summary, the contributions of our paper are four-fold. (I) We propose the first SSL framework, with the transformer-CNN cohort, that imposes multi-level consistency on the pixel-wise predictions and the heterogeneous feature space. (II) We propose CFCD to distill the complementary class-wise feature knowledge via pseudo labeling for the *unlabeled data*. Notably, we are also the *first* to explore pseudo labeling for feature distillation in semi-supervised segmentation. (III) We propose CCD to distill the pixel-wise prediction knowledge to impose consistency for the students in the cohort. (IV) Our TCC framework achieves *new* SoTA performance on both benchmarks.

II. Related work

Semi-supervised Semantic Segmentation. Consistency regularization is widely applied for semi-supervised segmentation [25], [26], [27]. The key insight of this branch of approaches is that the predictions or intermediate features should be consistent across different semantic-preserving transformations on input or model of the same data. The image-level perturbation methods, *e.g.*, [16], [28] randomly augment the input images while the feature-level perturbation methods, *e.g.*, [29] uses a multi-decoder strategy to augment the features. Moreover, CPS [6] enforces consistency by using the pseudo segmentation maps with additional benefits like expanding the training data. Concurrently, [17] proposes SSL-based method for medical imaging using ViT's and CNN's predictions. Differently, we propose the first SSL method for semantic segmentation by exploring the potential of ViT and CNN and subtly incorporating the multi-level consistency distillation on the pixel-wise predictions and the heterogeneous feature space via pseudo labeling for the unlabeled data.

Vision Transformer. Transformer was proposed by Vaswani et al. [30] to solve the machine translation tasks. Several works have applied ViT to high-level vision tasks, *e.g.*, object detection [31], [32], [33], [34] and semantic segmentation [35], [36], [37], [38]. Recently, PVT [39] introduces the pyramid structure into ViT to generate multi-scale features for dense prediction tasks. ViT has been continually improved and achieved better performance on large-scale datasets [21]. However, ViT does not generalize well in case of insufficient data [18], [40], [41]. Therefore, pre-training on a curated data is required for training an efficient ViT model. To tackle this issue, Bao et al. [42] introduce a masked image modeling approach to the pre-trained ViT while Touvron et al. [21] explore a hard distillation method. We explore the potential of ViT and incorporate it with CNNs as a cohort for semi-supervised semantic segmentation. Our TCC framework subtly imposes the multi-level consistency distillation on the pixel-wise predictions and the heterogeneous feature space via pseudo labeling for the unlabeled data.

Knowledge Distillation (KD) aims to build a smaller (student) model with the softmax labels of a larger (teacher) model [43], [41]. There are several paradigms in KD, including soft distillation [43], hard-label distillation, and label smoothing [44]. Some works [45], [46], [47], [48], [49] explore the structural information within the feature space to learn more generic representation. DeiT [21] first introduces a KD method specific to ViT aiming to distill the token. It shows that using the CNNs as teachers achieves better performance than using the ViT models mainly because of the inductive bias brought by convolution (Convs). Recently, Raghu *et al.*. [20] analyzes the representation structure of ViT and CNN on the visual recognition tasks and finds striking differences between the two models. That is, Convs and MSAs are two ways of extracting features, making CNNs and ViT sensitive to different features. Applying KD to CNNs and ViT cohort in semi-supervised segmentation is challenging as the inputs and learning capability for both students are different; we thus propose CFCD that leverages the feature maps via pseudo prototype from each student and transfers the complementary class-wise information.

III. The Proposed Approach

An overview of the proposed TCC framework is shown in Fig. 2 (a). Given a labeled set D_l of N labeled images and a set D_u of M unlabeled images, we propose the first yet novel SSL method for semantic segmentation, called Transformer-CNN cohort (TCC), by subtly incorporating the multi-level distillation to add consistency on the pixel-wise predictions and the heterogeneous feature space via pseudo labeling for the unlabeled data. Our TCC framework consists of two students: the ViT student $f(X;\theta_{ViT})$ and the CNN student $f(X;\theta_{CNN})$. That is, given the input images X, we aim to attain the segmentation confidence maps (P_{CNN} and P_{ViT}), high-level features (F_{CNN} and F_{ViT}) and pseudo labels(L_{CNN} and L_{ViT}) from the cohort $f(X;\theta_{ViT})$ and $f(X;\theta_{CNN})$, which can be formulated as:

$$L_{CNN}, P_{CNN}, F_{CNN} = f(X;\theta_{CNN}); \quad (1)$$

$$L_{ViT}, P_{ViT}, F_{ViT} = f(X;\theta_{ViT}). \quad (2)$$

Our key ideas are three folds. Firstly, as $f(X;\theta_{ViT})$ and $f(X;\theta_{CNN})$ have different inputs and inner feature forms, the extracted feature maps encode crucial complementary class-wise statistics. Therefore, we propose class-aware feature consistency distillation (CFCD) that first leverages the output of each student as the pseudo labels and generates feature prototype maps.Note that it is the *first time we explore pseudo labelling in SSL to facilitate feature distillation for the unlabeled data*. The class-aware features maps are computed by averaging the features on all pixels having the same pseudo labels. Feature variation knowledge is transformed via the cohort's class-aware feature (CF) maps.

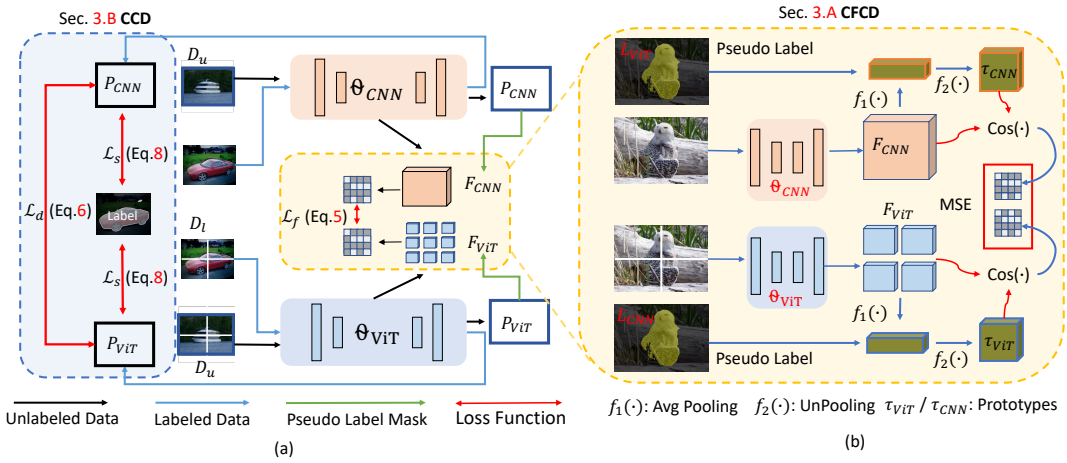


Fig. 2. (a) The proposed TCC framework comprises two students: the ViT $f(X;\theta_{ViT})$ and the CNN student $f(X;\theta_{CNN})$. The dashed lines indicate the fully supervised learning of the cohort with limited labeled data. TCC incorporates multi-level consistency distillation for the pixel-wise predictions (CCD) and class-aware feature consistency distillation (CFCD) for the unlabeled data. (b) The proposed CFCD involves two networks with different class-wise feature variations, characterized by the similarity between the feature on each pixel and its corresponding class-wise prototype (dashed lines). The higher similarity between the prototypes results in lower variation. The two networks, $f(\theta_{ViT})$ and $f(\theta_{CNN})$, learn from each other accordingly.

Secondly, as $f(X;\theta_{ViT})$ possesses more uniform representations for all layers [20], we propose Consistency-aware Cross Distillation (CCD) that distills the pixel-wise prediction information bidirectionally based on the heterogeneous students. Lastly, similar to other SSL methods, *e.g.*, [25], supervised training is applied to both $(X;\theta_{ViT})$ and $f(X;\theta_{CNN})$ for the limited labeled data. We now describe these components in detail.

A. Class-aware Feature Consistency Distillation

Pseudo Prototype Theoretically, high-dimensional feature representations obtained by two different models should be distinct explicitly but share implicit commonalities [22]. For all pixels of the same class in the corresponding class-wise label maps, their mapping center in feature space is referred to as the class-wise prototype. *This prototype is based on the condition that all pixels belonging to the same class should coincide in feature mapping space*. However, the feature maps of all pixels of the same class may not fall on the prototype completely; therefore, we estimate the class-aware feature variation to measure the similarity between the mapping of each pixel and the prototype. The class-aware feature variation for pixels can be obtained from the predictions and the inner features of $f(X;\theta_{ViT})$ and $f(X;\theta_{CNN})$. As features reflects how students in the cohort understand the input, it is crucial that reliable prototype calculation is guaranteed. However, for the unlabeled data, there are no prior labels for computing the prototype in SSL; thus, we leverage the pseudo labels predicted from one student as the source prototype for the other. That is, the CF map obtained from $f(X;\theta_{ViT})$ is taken as the standard prototype for student $f(X;\theta_{CNN})$ in the feature space, and vice versa. As such, $f(X;\theta_{ViT})$ and $f(X;\theta_{CNN})$ can be better correlated to each other despite of the difference of computing paradigms (MHSAs for $f(X;\theta_{ViT})$ and Convs for $f(X;\theta_{CNN})$).

As shown in Fig. 2 (b), to impose class-aware feature consistency to both students in the cohort, pseudo labels are down-sampled with nearest neighbour interpolation to match the spatial size of the high-dimensional features. Then, average pooling is operated on the masked features, corresponding to pixels with the same label for each class to get the class-wise pseudo prototype. Finally, we perform average pooling on the masked region of each pseudo prototype to ensure each position stores the corresponding high-dimensional feature of the class-wise prototype. Overall, the prototype can be formulated as:

$$\mathcal{T}_{ViT} = f_2(f_1(Mask(F_{ViT}, L_{CNN}))); \tag{3}$$

$$\mathcal{T}_{CNN} = f_2(f_1(Mask(F_{CNN}, L_{ViT}))), \tag{4}$$

where $\mathcal{T}_{ViT}$ and $\mathcal{T}_{CNN}$ denote the pseudo prototypes for the students $f(X;\theta_{ViT})$ and $f(X;\theta_{CNN})$, respectively; $f_1(\cdot)$ is the average pooling; $f_2(\cdot)$ is the unpooling operation; F_{ViT} and F_{CNN} are the high-dimensional features masked by pseudo label maps L_{CNN} and L_{ViT}, respectively. As such, we can obtain CF maps via calculating the similarity, *e.g.*, cosine similarity, between $\mathcal{T}_{CNN}$ and $\mathcal{T}_{CNN}$ for each student, as shown in Fig. 2 (b). More details of CF map calculation will be described in the following section.

Feature Distillation via Pseudo Prototype The CNN kernel of $f(X;\theta_{CNN})$ inspects adjacent pixels and gradually expands its receptive field to a more significant portion of an image, producing features with high locality. By contrast, $f(X;\theta_{ViT})$ manipulates patch-level image at every stage, having a global vision even at the beginning. Though the two models achieve similar performance after the training, their learning process is distinctive. Moreover, the feature maps extracted from them encode vital complementary class-wise statistics. Intuitively, we find that transferring the feature-level knowledge can complement the drawbacks of each model and thus yield better results. Inspired by [22], we apply consistency regularization on the pseudo feature variation between $f(X;\theta_{ViT})$ and $f(X;\theta_{CNN})$, as depicted in Fig. 2 (b). Especially, we extract the high-dimensional features output from the last stage of $f(X;\theta_{ViT})$ and $f(X;\theta_{CNN})$ to calculate their corresponding class-wise prototype. The prototype $\mathcal{T}$ at pixel p of class c is computed by averaging the features on all pixels having class label c, given by Eq. 5:

$$\mathcal{T}(i) = \frac{1}{|S_c|} \sum_{i \in S_c} f(i), \tag{5}$$

$$M(i) = Cos(f(i), \mathcal{T}(i)); \tag{6}$$

where $f(i)$ denotes the feature on pixel i, S_c is the set of pixels having the label c, $|S_c|$ stands for the size of the set S_c, $\mathcal{T}(i)$ is the class-wise feature variation map from its cohort, and $M(i)$ denotes the value of class-wise feature variation map at pixel i. In Eq. 6, due to the intrinsic difference (*e.g.*, magnitude, deviation)

in the feature maps between $f(X;\theta_{ViT})$ and $f(X;\theta_{CNN})$, we adopt the cosine similarity $Cos(\cdot)$ to better formulate the relative distribution for each class. Finally, the CFCD loss $\mathcal{L}_f$ minimizes the distance between feature variation maps of the students $f(X;\theta_{ViT})$ and $f(X;\theta_{CNN})$. Specifically, we employ the Mean Squared Error (MSE) loss as follows:

$$\mathcal{L}_f = \frac{1}{N}\sum_{p\in\Omega}(M_{CNN}(i) - M_{ViT}(i))^2, \tag{7}$$

where N is the total numbers of pixels, Ω denotes the image, $M_{CNN}(p)$ and $M_{ViT}(p)$ represent the corresponding pseudo feature variation map of the CNN and ViT students.

B. Consistency-aware Cross Distillation

Though $f(\theta_{ViT})$ and $f(\theta_{CNN})$ share different learning capacities for the unlabeled data D_u, their predictions should be consistent according to the *smoothness assumption* where samples in the same cluster are expected to have the same labels. In particular, instead of using the Exponential Moving Average [7] to update the predictions, we find that measuring the cross-model discrepancy between $f(\theta_{ViT})$ and $f(\theta_{CNN})$ helps improve each student's representations. Accordingly, we propose Consistency-aware Cross Distillation (CCD) to enforce consistency between the outputs of the cohort to extract additional information for the unlabeled data D_u, as shown in Fig. 1 (a). CCD is bidirectional: one is from $f(\theta_{CNN})$ to $f(\theta_{ViT})$ and the other one is from $f(\theta_{ViT})$ to $f(\theta_{CNN})$. That is, we use the logits output P_{CNN} from the CNN student $f(\theta_{CNN})$ to supervise the logits output P_{ViT} of the ViT student $f(\theta_{ViT})$, and vice versa. The CCD loss on the unlabeled data D_u is:

$$\mathcal{L}_d^u = \frac{1}{|D_l|}(\mathcal{L}_{kl}(P_{CNN}, P_{ViT}) + \mathcal{L}_{kl}(P_{ViT}, P_{CNN})), \tag{8}$$

where the $\mathcal{L}_{kl}$ is the standard KL divergence. The CCD loss $\mathcal{L}_d^l$ on the labeled data D_l can be defined in the same manner. The total CCD loss is the combination of losses on both the labeled D_l and unlabeled data D_u: $\mathcal{L}_d = \mathcal{L}_d^l + \mathcal{L}_d^u$.

C. Optimization

The training objective contains three losses as follows:

$$\mathcal{L} = \mathcal{L}_s + g(t)\cdot(\mathcal{L}_d + \lambda\mathcal{L}_f), \tag{9}$$

where the $\mathcal{L}_s$ is supervised loss, the $\mathcal{L}_d$ refers to the prediction-level CCD loss and the $\mathcal{L}_f$ is the CFCD loss that measures the class-aware feature variation consistency between $f(\theta_{ViT})$ and $f(\theta_{CNN})$. The $g(t)$ is a consistency ramp-up function following [50], and λ is a fixed constant. The supervision loss $\mathcal{L}_s$ is formulated as:

$$\mathcal{L}_s = \frac{1}{|D_l|}\sum_{X\in D_l}(\mathcal{L}_{dice}(P_{CNN}, Y_{CNN}) + \mathcal{L}_{dice}(P_{ViT}, Y_{ViT})), \tag{10}$$

where $\mathcal{L}_{dice}$ indicates the Dice coefficient loss function and Y are the ground truth (GT) labels.

IV. Experiments and Evaluation

Datasets. **Pascal VOC** contains 20 foreground object classes plus an extra background class. The standard training, validation, and test sets consist of 1464, 1449, and 1456 images, respectively. We adopt the augmented set from [57] which contains 10582 images as our full training set. **Cityscapes** contains a diverse set of video sequences recorded in street scenes from 50 cities, with high-quality pixel-level annotations. The official split has 2975 images for training, 500 for validation, and 1525 for testing. Each image is finely annotated with pixel-level labels of 19 classes. We divide the whole training set into two groups by randomly sampling 1/2, 1/4, 1/8, and 1/16 of the entire set as labeled images and the rest as unlabeled images. For a fair comparison, images in each set are the same as CPS [6].

TABLE I

Comparison with state-of-the-arts on the PASCAL VOC.

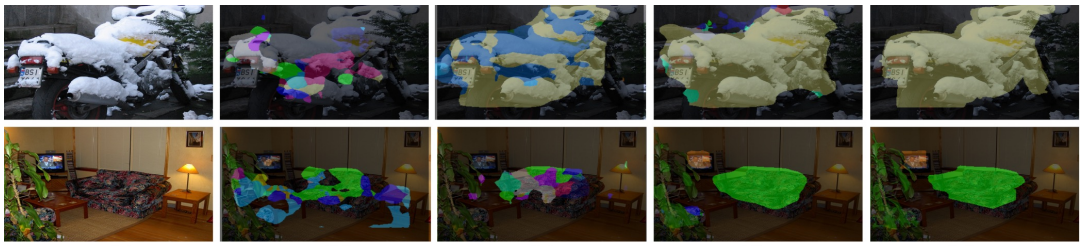
Method	Backbone	Label Rate			
		1/16	1/8	1/4	1/2
CCT [12]	ResNet-50	65.22	70.87	73.43	74.75
	ResNet-101	67.94	73.00	76.17	77.56
CPS [6]	ResNet-50	68.21	73.20	74.24	75.91
	ResNet-101	72.18	75.83	77.55	78.64
n-CPS [9]	ResNet-50	68.36	73.45	75.75	77.00
	ResNet-101	73.51	76.46	78.59	79.90
AEL [51]	ResNet-101	77.20	77.57	78.06	80.29
ST++ [52]	ResNet-50	73.20	75.50	76.00	-
	ResNet-101	74.70	77.90	77.90	-
U^2PL [53]	ResNet-101	77.21	79.01	79.30	80.50
ELN [54]	ResNet-50	70.52	73.20	74.63	-
	ResNet-101	72.52	75.10	76.58	-
PAMT [55]	ResNet-101	75.50	78.20	78.72	79.76
CSS [56]	ResNet-101	78.73	79.54	80.82	81.06
Ours	TCC-S	79.16	80.28	82.32	82.52
	w/ Cutmix	**81.35**	**83.05**	**83.55**	**84.04**
	TCC-B	80.17	81.17	82.42	82.80
	w/ Cutmix	**81.36**	**83.42**	**84.27**	**84.29**

TABLE II

Comparison with the state-of-the-arts on the Cityscapes val set under different partition protocols. All the methods are based on DeepLabv3+. (TCC-B: TCC-Base; w/ Cutmix: with cutmix augmentation)

Method	Backbone	Label Rate			
		1/16	1/8	1/4	1/2
MT [7]	ResNet-101	68.08	73.71	76.53	78.59
CCT [12]	ResNet-101	69.64	74.48	76.35	78.29
CPS [6]	ResNet-101	70.50	75.71	77.41	80.08
3-CPS [9]	ResNet-101	75.86	77.99	78.95	80.26
AEL [51]	ResNet-101	75.83	77.90	79.01	80.28
U^2PL [53]	ResNet-101	70.30	74.37	76.47	79.05
Ours	TCC-B	75.79	77.53	78.47	80.83
	w/ CutMix	**76.89**	**78.52**	**80.04**	**80.93**

Evaluation and comparison. We leverage the mean Intersection-over-Union (mIOU) as an evaluation metric. Our trained models are evaluated on PASCAL VOC 2012 validation set (1456 images) and the *Cityscapes* validation set (500 images) via testing at a single scale, respectively. We report the mIOU of the ViT-based model in the cohort. For comparison, in all tables, ***ResNet50, PVT-M, and ConvNext-S mean that two students in our TCC are based on the same backbone. Also, all the compared methods are implemented with dual students***. **Implementation details**. We implement our TCC framework using Pytorch. We initialize the weights of two backbones, *i.e.*, the students $f(\theta_{ViT})$ and $f(\theta_{CNN})$, with the weights pre-trained on ImageNet $1K$ and the weights of two segmentation heads (of DeepLabv3+ [58]) randomly. We use the AdamW optimizer and train the TCC framework for 30000 iterations with a total batch size of 16 for the PASCAL VOC dataset and 80000 iterations with a total batch size of 4 for Cityscapes. We employ the poly learning rate policy where the initial learning rate α is multiplied by $(1 - \frac{iteration}{maxiterations})^{0.9}$, and λ is simply set to 1 for both datasets. For fair comparisons, two types of settings in our TCC are settled to be compared with ResNet-50 and ResNet-101, *i.e.*, Cohort-Small (TCC-S) and Cohort-Base (TCC-



(a) (b) (c) (d) (e)

Fig. 3. **Example results from PASCAL VOC 2012**. (a) input, (b) CCT [12], (c) CPS [6], (d) ours (TCC), and (e) ground truth. All the approaches are based on DeepLabv3+.

B). Specifically, *ResNet-50 and PVT-S are the two backbones in TCC-S, while ConvNeXt-S and PVT-M in TCC-B*. We also report the results of ResNet, PVT, and ConvNext-based implementations of TCC in Tab. VI.

A. Comparison with the SoTA methods

We compare our method with the SoTA semi-supervised methods including Mean-Teacher(MT) [7], Cross Consistency Training(CCT) [12], Cross Pseudo Supervision(CPS) [6], AEL [51], ST++ [52], U^2PL [53] and ELN [54] under different label ratios.

Our method even outperforms n-CPS[9], a 3-model based method, by 7.85%, 6.96%, 5.68%, and 4.90%, respectively, under the same labeled ratios on *Pascal VOC* dataset. The greatest improvement at the 1/16 label ratio indicates that the combination of CNN and ViT students can facilitate each other when learning unlabeled data and generalize well on labeled data, which is consistent with our assumption.

Detailed results by datasets. *PASCAL VOC 2012*: Tab. I shows the comparison results. On all label ratios (1/2, 1/4, 1/8, and 1/16), our TCC approach (w/o CutMix) consistently outperforms the other methods. And our TCC approach (w/ CutMix) achieves the best performance and sets new SoTA under all label ratios. With lower labeled ratios, our approach (w/ CutMix) outperforms the U^2PL by 4.15% and 4.41%, respectively, under 1/16 and 1/8 label ratios. ***This confirms that using two heterogeneous models, i.e., ViT and CNN, in the consistency regularization approaches achieves better performance, especially with less annotated data***. Fig. 3 shows the qualitative results. CCT struggles to catch the main objects from inputs and wrongly classifies many regions of interest into the background (black). It renders a colorful mask over a single object and is especially devastating if some natural camouflage exists, *e.g.*, the snow-covered motorbike. CPS performs relatively better than CCT. It can roughly outline the boundary of objects of interest but the inner pixel-wise segmentation results are still heterogeneous for a single object. In contrast, our method achieves much neater and cleaner segmentation results (4th column), which is much closer to the GT label maps (5th column).

Cityscapes val set: Tab. II shows the quantitative results where our TCC approach consistently outperforms the SoTA methods. The improvements of mIOU of our method (w/o CutMix augmentation) over the 2-model baseline method (AEL) are 1.06%, 0.62%, 1.03%, 0.65% under label ratio of 1/16, 1/8, 1/4, and 1/2, respectively. The qualitative results are shown in Fig. 4.

B. Ablation study and Analysis

Improving Few-Supervision Since our TCC framework outperforms the SoTA methods with less labeled data, as mentioned above, we study the performance of TCC on the PASCAL VOC 2012 with few labels by following the same partition in PseudoSeg [61] which randomly samples $1/2$, $1/4$, $1/8$, and $1/16$ of images in the standard training set (1464 images) to construct the labeled set. The remaining images are used as an unlabeled set. We report the results of our approach (w/ and w/o CutMix augmentation). The results are listed in Tab. V. The improvements of mIOU of our model (w/ CutMix augmentation) over CPS [6] are 6.92%, 5.21%, 7.19% and 9.32%, respectively, under 1/16, 1/8, 1/4, and 1/2 label ratios. Our method achieves the best results and is superior to CPS again on the few labeled case. This validates the fact that our TCC subtly incorporates the multi-level distillation to add consistency on the pixel-wise predictions and the heterogeneous feature space via pseudo labeling for the unlabeled data.

TABLE III
DIFFERENT LOSS COMBINATIONS ON PASCAL VOC.

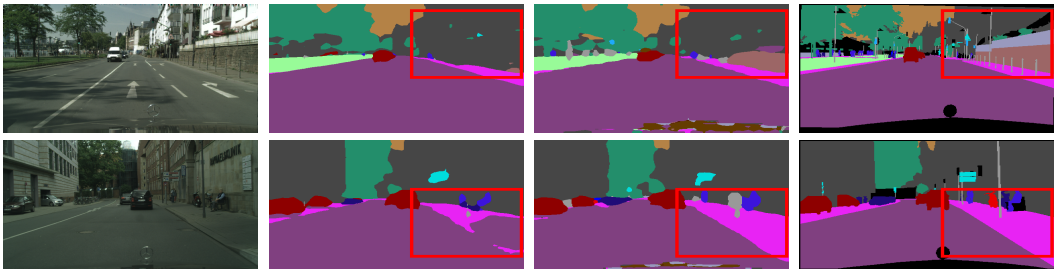
Losses			PASCAL VOC 2012			
$\mathcal{L}_s$	$\mathcal{L}_d$	$\mathcal{L}_f$	1/16	1/8	1/4	1/2
✓			74.65	76.22	76.71	77.01
✓	✓		77.90	80.46	81.60	81.95
✓	✓	✓	**80.17**	**81.17**	**82.42**	**82.80**

TABLE IV
COMPARISON OF OUR TCC FRAMEWORK TRAINED FORM SCRATCH (WITHOUT PRE-TRAINED MODELS) AND PREVIOUS SOTA SEMI-SUPERVISED SEGMENTATION METHODS.

Method	1/16	1/8	1/4	1/2
MT[7]	66.77	70.78	73.22	75.41
CCT[12]	65.22	70.87	73.43	74.75
CPS[6]	68.21	73.20	74.24	75.91
CPS*[6]	74.18	74.19	74.76	78.37
3-CPS-mc[9]	68.36	73.45	75.75	77.00
3-CPS-mc[9]	72.03	74.18	75.85	76.65
Ours(w/ pre-train)	**80.17**	**81.17**	**82.42**	**82.80**
Ours(w/o pre-train)	**78.61**	**79.43**	**80.09**	**80.14**

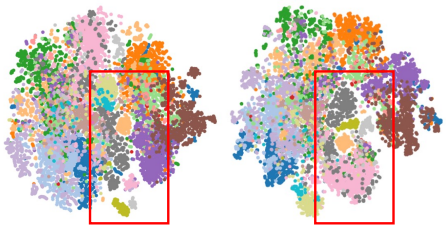
TABLE V
COMPARISON FOR FEW-SUPERVISION ON PASCAL VOC DATASET.

Method	labeled samples			
	1/2	1/4	1/8	1/16
AdvSemSeg[59]	65.27	59.97	47.58	39.69
CCT[12]	62.10	58.80	47.60	33.10
GCT[29]	70.67	64.71	54.98	46.04
VAT[60]	63.34	56.88	49.35	36.92
CutMix-Seg[16]	69.84	68.36	63.20	55.58
PseudoSeg[61]	72.41	69.14	65.50	57.60
CPS+CutMix[6]	75.88	71.71	67.42	64.07
Ours	77.92	74.25	72.99	72.01
Ours w/ CutMix	**82.80**	**76.92**	**74.61**	**73.39**



(a) (b) (c) (d)

Fig. 4. **Example results from Cityscapes**. (a) input, (b) CPS [6], (c) ours, and (d) GT. All methods are based on DeepLabv3+.



(a) (b)

Fig. 5. TSNE visualization of (a) w/o and (b) w/ CFCD.

Train from scratch *We also train our TCC framework without pre-trained models.* Tab. IV demonstrates that, even without ImageNet 1K pre-trained models, our TCC framework outperforms the previous SoTA method with pre-trained models by a large margin.

Loss functions. We conduct ablation experiments on the PASCAL VOC 2012 to analyze the impact of the CCD loss $\mathcal{L}_d$ and CFCD loss $\mathcal{L}_f$ in our TCC framework. In Tab. III, different combinations of losses are applied. We can see that TCC framework leverages the unlabeled data well with an average improvement of 5.49% over fully supervised by labeled data. Both $\mathcal{L}_d$ and $\mathcal{L}_f$ contribute positively to the validation mIOU in every label rate. We notice that the increases brought by $\mathcal{L}_f$ decline when the number of unlabeled data decreases. It is also intuitive since CF maps in $\mathcal{L}_f$ is inferred by unlabeled data. Reduction of the hampers models fitting the actual class-wise feature variance distribution, thus lowering the improvement downward.

Difference with CPS [6] Our TCC shares a quite different spirit with CPS in that **1)** ours imposes *multi-level* consistency on the pixel-wise predictions and heterogeneous features, but CPS merely utilizes the prediction-level cross-pseudo supervision which neglects the feature-level knowledge; **2)** we are the ***first*** to explore pseudo labeling for feature distillation; **3)** we are the ***first*** to study how heterogeneous models (CNNs and ViTs) benefit SSL performance. Note that CPS only utilizes the same backbone.

Visualization of class-aware feature distillation. In Fig. 5 below, we provide the TSNE visualization of features with (b) and without (a) our proposed CFCD. Obviously, our class-wise feature distillation makes the models in the cohort achieve better class-wise distinguishing abilities.

Varying backbone models. Since there are two models in our TCC framework, we compare the CPS[6] following the two-model structure. We conduct the ablation studies on the PASCAL VOC 2012 with 1/8 label rate to analyze the impact of the changing backbones, including ResNet, PVT, ConvNext and our TCC. **Tab. VI shows that our cohort can better promote the segmentation performance under the dual-student framework in SSL.** We fully explore the potential of heterogeneous computing paradigms (MSAs and Convs) via FVCD in the feature space and CCD on the prediction. This way, we successfully introduce ViT into the semi-supervised semantic segmentation.

TABLE VI

TCC WITH DIFFERENT BACKBONES ON PASCAL VOC 2012.

Method	Backbones	Label Ratio			
		1/16	1/8	1/4	1/2
U^2PL	ResNet-101	77.21	79.01	79.30	80.50
Ours	ResNet-50	77.79	78.95	80.51	80.98
	PVT-M	71.12	73.84	74.44	76.72
	ConvNext-S	78.92	80.15	80.81	80.95
	TCC-S	79.16	80.28	82.32	82.52
	TCC-B	**80.17**	**81.17**	**82.42**	**82.80**

TABLE VII

EVALUATION OF CNN AND VIT IN OUR TCC-S.

Method	Backbone	PASCAL VOC			
		1/16	1/8	1/4	1/2
TCC-S	ResNet-50	80.98	81.34	81.78	82.67
	PVT-S	81.35	83.05	83.55	84.04

Inference with dual students. During inference, our proposed CFCD makes the two students in the cohort learn from each other, and the ViT-based student achieves better results than the CNN-based one thanks to the better learning ability of MHSAs. Thus, we report the performance of ViT in the cohort in all tables. We also report dual students' performance in Tab. VII. Obviously, the ViT-based PVT-S in the dual student achieves better results.

V. CONCLUSION

In this paper, we proposed TCC, a novel framework for semi-supervised semantic segmentation by exploring the best of both students. Our method subtly incorporates the multi-level consistency regularization on both the predictions and the heterogeneous feature space via pseudo labeling for the unlabeled data. First, the feature variation knowledge is transformed via CF maps between the cohort. Second, we also proposed to distill knowledge from the pixel-wise predictions based on the heterogeneous students. The proposed TCC framework significantly outperformed the SoTA semi-supervised methods by a large margin.

Limitation and future work: The proposed TCC demonstrates that using heterogeneous models in consistency regularization gives high-performance gain in semi-supervised learning. For future work, we plan to explore the dual-student framework with more heterogeneous models for semi-supervised semantic segmentation.

Acknowledgement: This paper is supported by the National Natural Science Foundation of China (NSF) under Grant No. NSFC22FYT45 and the Guangzhou City, University and Enterprise Joint Fund under Grant No.SL2022A03J01278.

REFERENCES

[1] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[2] J. Chen, D. Deguchi, C. Zhang, X. Zheng, and H. Murase, "Frozen is better than learning: A new design of prototype-based classifier for semantic segmentation," *Available at SSRN 4617170*.

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[4] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3349–3364, 2020.

[5] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "Denseaspp for semantic segmentation in street scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3684–3692.

[6] X. Chen, Y. Yuan, G. Zeng, and J. Wang, "Semi-supervised semantic segmentation with cross pseudo supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2613–2622.

[7] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in neural information processing systems*, vol. 30, 2017.

[8] Z. Ke, D. Wang, Q. Yan, J. Ren, and R. W. Lau, "Dual student: Breaking the limits of the teacher in semi-supervised learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6728–6736.

[9] D. Filipiak, P. Tempczyk, and M. Cygan, "n-cps: Generalising cross pseudo supervision to n networks for semi-supervised semantic segmentation," *arXiv preprint arXiv:2112.07528*, 2021.

[10] Y. Grandvalet, Y. Bengio, *et al.*, "Semi-supervised learning by entropy minimization." *CAP*, vol. 367, pp. 281–296, 2005.

[11] F. Grezl and M. Karafiát, "Semi-supervised bootstrapping approach for neural network feature extractor training," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE, 2013, pp. 470–475.

[12] Y. Ouali, C. Hudelot, and M. Tami, "Semi-supervised semantic segmentation with cross-consistency training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 674–12 684.

[13] S. Mittal, M. Tatarchenko, and T. Brox, "Semi-supervised semantic segmentation with high-and low-level consistency," *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[14] B. Zhang, Y. Wang, W. Hou, H. Wu, J. Wang, M. Okumura, and T. Shinozaki, "Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[15] L. Wang and K.-J. Yoon, "Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[16] G. French, S. Laine, T. Aila, M. Mackiewicz, and G. Finlayson, "Semi-supervised semantic segmentation needs strong, varied perturbations," *arXiv preprint arXiv:1906.01916*, 2019.

[17] X. Luo, M. Hu, T. Song, G. Wang, and S. Zhang, "Semi-supervised medical image segmentation via cross teaching between cnn and transformer," *arXiv preprint arXiv:2112.04894*, 2021.

[18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[19] N. Park and S. Kim, "How do vision transformers work?" in *International Conference on Learning Representations*, 2021.

[20] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?" *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[21] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International Conference on Machine Learning*. PMLR, 2021, pp. 10 347–10 357.

[22] Y. Wang, W. Zhou, T. Jiang, X. Bai, and Y. Xu, "Intra-class feature variation distillation for semantic segmentation," in *European Conference on Computer Vision*. Springer, 2020, pp. 346–362.

[23] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results," http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[24] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[25] K. Sohn, D. Berthelot, C.-L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, and C. Raffel, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *arXiv preprint arXiv:2001.07685*, 2020.

[26] X. Zheng, C. Fu, H. Xie, J. Chen, X. Wang, and C.-W. Sham, "Uncertainty-aware deep co-training for semi-supervised medical image segmentation," *Computers in Biology and Medicine*, vol. 149, p. 106051, 2022.

[27] J. Chen, C. Fu, H. Xie, X. Zheng, R. Geng, and C.-W. Sham, "Uncertainty teacher with dense focal loss for semi-supervised medical image segmentation," *Computers in Biology and Medicine*, vol. 149, p. 106034, 2022.

[28] H. Xie, C. Fu, X. Zheng, Y. Zheng, C.-W. Sham, and X. Wang, "Adversarial co-training for semantic segmentation over medical images," *Computers in biology and medicine*, vol. 157, p. 106736, 2023.

[29] Z. Ke, D. Qiu, K. Li, Q. Yan, and R. W. Lau, "Guided collaborative training for pixel-wise semi-supervised learning," in *European conference on computer vision*. Springer, 2020, pp. 429–445.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[31] J. Beal, E. Kim, E. Tzeng, D. H. Park, A. Zhai, and D. Kislyuk, "Toward transformer-based object detection," *arXiv preprint arXiv:2012.09958*, 2020.

[32] Z. Dai, B. Cai, Y. Lin, and J. Chen, "Up-detr: Unsupervised pre-training for object detection with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1601–1610.

[33] Z. Sun, S. Cao, Y. Yang, and K. M. Kitani, "Rethinking transformer-based set prediction for object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3611–3620.

[34] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.

[35] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr, *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6881–6890.

[36] B. Wu, C. Xu, X. Dai, A. Wan, P. Zhang, Z. Yan, M. Tomizuka, J. Gonzalez, K. Keutzer, and P. Vajda, "Visual transformers: Token-based image representation and processing for computer vision," *arXiv preprint arXiv:2006.03677*, 2020.

[37] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, and W. Gao, "Pre-trained image processing transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 299–12 310.

[38] Y. Wang, Z. Xu, X. Wang, C. Shen, B. Cheng, H. Shen, and H. Xia, "End-to-end video instance segmentation with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8741–8750.

[39] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," *arXiv preprint arXiv:2102.12122*, 2021.

[40] J. Zhu, Y. Luo, X. Zheng, H. Wang, and L. Wang, "A good student is cooperative and reliable: Cnn-transformer collaborative learning for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 11 720–11 730.

[41] X. Zheng, Y. Luo, P. Zhou, and L. Wang, "Distilling efficient vision transformers from cnns for semantic segmentation," *arXiv preprint arXiv:2310.07265*, 2023.

[42] H. Bao, L. Dong, and F. Wei, "Beit: Bert pre-training of image transformers," *arXiv preprint arXiv:2106.08254*, 2021.
[43] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.
[44] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.
[45] N. Passalis and A. Tefas, "Learning deep representations with probabilistic knowledge transfer," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 268–284.
[46] B. Peng, X. Jin, J. Liu, D. Li, Y. Wu, Y. Liu, S. Zhou, and Z. Zhang, "Correlation congruence for knowledge distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5007–5016.
[47] Y. Tian, D. Krishnan, and P. Isola, "Contrastive representation distillation," *arXiv preprint arXiv:1910.10699*, 2019.
[48] X. Zheng, J. Zhu, Y. Liu, Z. Cao, C. Fu, and L. Wang, "Both style and distortion matter: Dual-path unsupervised domain adaptation for panoramic semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1285–1295.
[49] X. Zheng, T. Pan, Y. Luo, and L. Wang, "Look at the neighbor: Distortion-aware unsupervised domain adaptation for panoramic semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 18 687–18 698.
[50] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," *arXiv preprint arXiv:1610.02242*, 2016.
[51] H. Hu, F. Wei, H. Hu, Q. Ye, J. Cui, and L. Wang, "Semi-supervised semantic segmentation via adaptive equalization learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 22 106–22 118, 2021.
[52] L. Yang, W. Zhuo, L. Qi, Y. Shi, and Y. 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*, 2022, pp. 4268–4277.
[53] Y. Wang, H. Wang, Y. Shen, J. Fei, W. Li, G. Jin, L. Wu, R. Zhao, and X. Le, "Semi-supervised semantic segmentation using unreliable pseudo-labels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4248–4257.
[54] D. Kwon and S. Kwak, "Semi-supervised semantic segmentation with error localization network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9957–9967.
[55] Y. Liu, Y. Tian, Y. Chen, F. Liu, V. Belagiannis, and G. Carneiro, "Perturbed and strict mean teachers for semi-supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4258–4267.
[56] C. Wang, H. Xie, Y. Yuan, C. Fu, and X. Yue, "Space engage: Collaborative space supervision for contrastive-based semi-supervised semantic segmentation," *arXiv preprint arXiv:2307.09755*, 2023.
[57] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik, "Semantic contours from inverse detectors," in *2011 international conference on computer vision*. IEEE, 2011, pp. 991–998.
[58] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.
[59] W.-C. Hung, Y.-H. Tsai, Y.-T. Liou, Y.-Y. Lin, and M.-H. Yang, "Adversarial learning for semi-supervised semantic segmentation," *arXiv preprint arXiv:1802.07934*, 2018.
[60] T. Miyato, S.-i. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: a regularization method for supervised and semi-supervised learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 1979–1993, 2018.
[61] Y. Zou, Z. Zhang, H. Zhang, C.-L. Li, X. Bian, J.-B. Huang, and T. Pfister, "Pseudoseg: Designing pseudo labels for semantic segmentation," *arXiv preprint arXiv:2010.09713*, 2020.