Training Vision Transformers for Semi-Supervised Semantic Segmentation

Xinting Hu xhu@mpi-inf.mpg.de
Li Jiang lijiangcse@gmail.com
Bernt Schiele schiele@mpi-inf.mpg.de
Max Planck Institute for Informatics, Saarland Informatics Campus, Germany

Abstract

We present S^{4}Former, a novel approach to training Vision Transformers for Semi-Supervised Semantic Segmentation (S^4). At its core, S^{4}Former employs a Vision Transformer within a classic teacher-student framework, and then leverages three novel technical ingredients: PatchShuffle as a parameter-free perturbation technique, Patch-Adaptive Self-Attention (PASA) as a fine-grained feature modulation method, and the innovative Negative Class Ranking (NCR) regularization loss. Based on these regularization modules aligned with Transformer-specific characteristics across the image input, feature, and output dimensions, S^{4}Former exploits the Transformer's ability to capture and differentiate consistent global contextual information in unlabeled images. Overall, S^{4}Former not only defines a new state of the art in S^4 but also maintains a streamlined and scalable architecture. Being readily compatible with existing frameworks, S^{4}Former achieves strong improvements (up to 4.9%) on benchmarks like Pascal VOC 2012, COCO, and Cityscapes, with varying numbers of labeled data. The code is at https://github.com/JoyHuYY1412/S4Former.

1. Introduction

Semi-supervised semantic segmentation (S^4) aims to relieve the heavy dependence on extensive pixel-level annotations by leveraging unlabeled images. Prevailing S^4 works [19, 45, 53, 61] employ the teacher-student mechanism [27, 42]: pseudo-labels from a weakly augmented unlabeled image guide the training of the strongly augmented counterpart. Despite significant progress, recent works that rely on ConvNet-based segmenters [53, 55, 56] seem to reach a performance plateau, as shown in Fig. 1. Vision Transformers, having demonstrated their efficacy in supervised and other semi-supervised tasks [4, 5, 50], offer a promising yet under-explored path for advancing S^4.

We introduce S^{4}Former, a novel approach that integrates Vision Transformers into the teacher-student paradigm for the S^4 task for the first time (to the best of our knowledge). Beyond simply substituting backbones from ConvNets to Vision Transformers, we emphasize the need for tailored training mechanisms to fully exploit Vision Transformers in S^4. Our empirical findings reveal the critical importance of regularization strategies for Vision Transformers. As illustrated in Figure 2, with suitable regularization techniques, training with a limited set of unlabeled data can outperform naïve training with a larger unlabeled dataset. To better exploit Vision Transformers, our primary motivation is to harness its capabilities in comprehending global and long-range contexts. This capability ensures a more consistent understanding of semantic content across various perturbed instances of unlabeled images. Stemming from this insight, we propose a suite of regularization techniques tailored for Vision Transformers, including PatchShuffle, Patch-Adaptive Self-Attention (PASA), and Negative Class Ranking (NCR) loss, addressing image, feature, and output regularization, respectively.

Our proposed components in the S^{4}Former framework capitalize on the Vision Transformer's architectural attributes, such as its innate patch-based architecture and self-attention mechanism. Consequently, we can regularize S^{4}Former with consistent global and long-range contextual cues inherent in images. In particular, PatchShuffle chal-

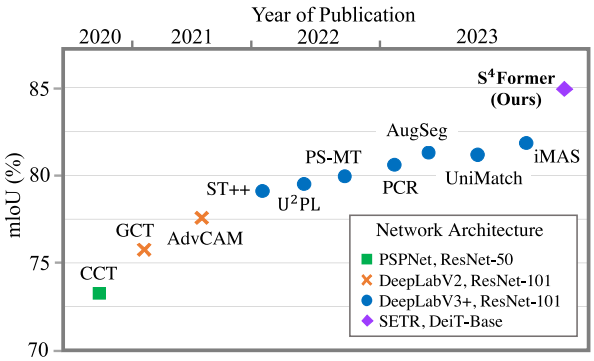

Figure 1. **Advancements of semi-supervised semantic segmentation (S^4) work over time.** Our proposed S^{4}Former achieves significant improvements over previous methods with a limited number of labeled images. Results are reported on the Pascal VOC [14] dataset with 1,464 images labeled and 9,118 images from SBD [18] unlabeled.

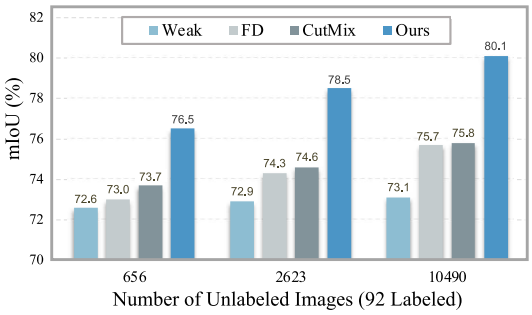

Figure 2. With diverse training strategies based on the Transformer backbone, the performance gain exhibits considerable variation when obtaining more unlabeled images. Weak: weak image augmentations that are the same as those used for the teacher model. FD: feature dropout. CutMix: cut and mix regions over images [54].

lenges the Transformer to reconstruct semantic understanding from a shuffled spatial structure, thereby reinforcing its reliance on global contextual awareness, while the PASA module adjusts the self-attention process to prioritize complex and ambiguous areas that are typically challenging for segmentation. The NCR loss, while not exclusive to Transformers, extends consistency constraints beyond the standard "positive" pseudo-label and encompasses "negative" classes to enhance regularization.

Armed with the proposed mechanisms, S^{4}Former not only pioneers the training of Vision Transformers for semi-supervised semantic segmentation but does so by overcoming the specific challenges of spatial and class confusion that have limited the effectiveness of the existing ConvNet-based methods from a new perspective. Notably, the overall proposed method is straightforward, yet effective, with no additional training parameters or memory bank introduced. We anticipate that our streamlined S^{4}Former approach will not only establish a new performance benchmark but also serve as a timely reference point for future research into semi-supervised semantic segmentation using Transformer architectures, potentially inspiring innovative methodologies in this evolving field.

Our contributions are summarized as follows:

- We pioneer the study of training Vision Transformers for S^4, establishing a new effective benchmark in this domain.
- Without introducing extra trainable parameters, our proposed S^{4}Former enhances the teacher-student framework with innovative perturbation and regularization techniques, addressing the consistency regularization unique to the Transformer architecture.
- Our extensive evaluations confirm that S^{4}Former sets a new state-of-the-art on Pascal VOC 2012, COCO, and Cityscapes, laying solid groundwork for future research with Transformers in S^4.

2. Related Works

Semi-Supervised Learning (SSL). Semi-supervised learning aims to utilize unlabeled data to improve the model learned on labeled data. Pseudo-labeling methods [26, 38, 49] extend the training dataset by predicting pseudo-labels for unlabeled data. Notably, Mean Teacher [42] optimizes the generation of robust pseudo-labels by employing an exponential moving average (EMA) of the student model. Consistency regularization approaches [3, 25, 40] enforce the model to yield consistent predictions for different perturbed versions of the same unlabeled image. Prevailing SSL methods [2, 41] combine the two existing techniques and predict improved pseudo-labels. Specifically, FixMatch [41], proposes to inject strong perturbations to unlabeled images and supervise the training process with predictions from weakly perturbed ones. This paper focuses on how different types of perturbations can be systematically exploited in semi-supervised vision tasks, going beyond the established weak-to-strong paradigm.

Semi-Supervised Semantic Segmentation (S^4). Based on the teacher-student architecture [27, 52, 53, 61], some works aim to improve the quality of pseudo-labels by stabilizing the teacher model with multi-head [15, 21, 31] or cross-head supervision [9, 24, 46], or adopting contrastive learning [1, 44, 59], uncertainty-based thresholding [23, 45], and class-imbalance learning [16, 19]. Other works focus on injecting strong perturbations into the student to avoid confirmation bias [28]. In practice, strong data augmentation [17, 47, 52, 56]/feature perturbations [31, 35, 53] are applied to the input images/features of the student. Recent state-of-the-art S^4 methods [19, 31, 37, 45] combine the two strengths by imposing stronger augmentation and generating better pseudo-labels. Our work extends these advances by incorporating Vision Transformers and introducing a novel image augmentation strategy, attention adjustment mechanism, and regularization loss function, thereby establishing a new direction for S^4 methods.

Vision Transformers. Vision Transformers have recently been embraced in various computer vision domains, including image classification [8, 13], object detection [7, 60], semantic segmentation [10, 48, 57], and other high-level vision tasks [29, 36, 51], thanks to their ability to model long-range dependencies among elements [22, 32, 33]. In this work, we use ImageNet [39] pre-trained Vision Transformers for semantic segmentation, *i.e.*, DeiT [57] and MiT [48] as our backbone models. Recently, some works [20, 58] in semi-supervised semantic segmentation have attempted to introduce the Vision Transformer as a parallel branch to the ConvNet model. Our paper is the first, to the best of our knowledge, to explore the training of Vision Transformers instead of ConvNets for semi-supervised semantic segmentation.

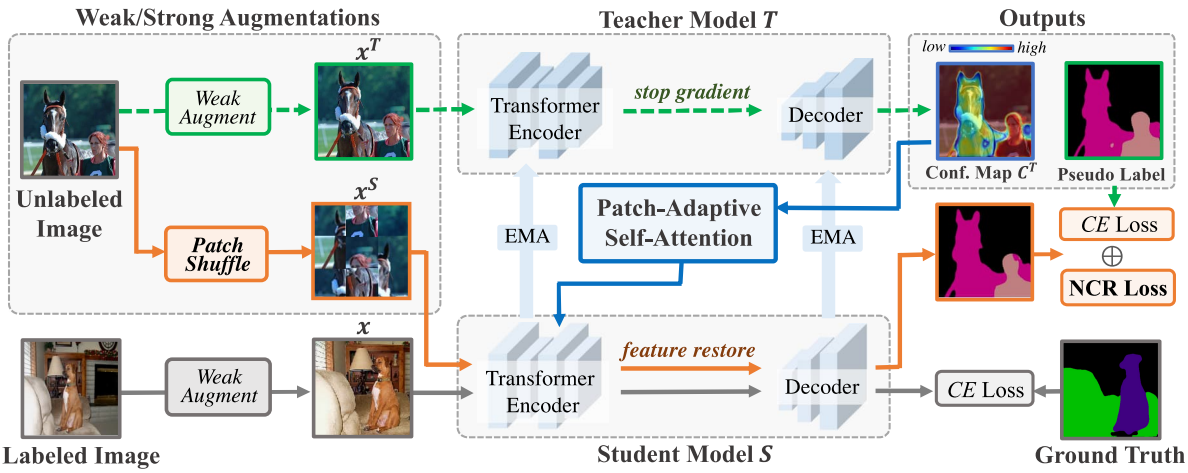

Figure 3. **Overview of S^{4}Former.** The S^{4}Former incorporates both student and teacher models, each equipped with a Transformer encoder and an image decoder. The weights of the teacher model are the exponential moving average of the updating student. For labeled images, we train them with the student model by standard supervised loss. For unlabeled images: 1) they undergo both weak and strong augmentations, with the strong version utilizing the Patch-Shuffle technique (Section 3.2); 2) the teacher model processes the weakly augmented version to derive pseudo-labels for the strongly augmented images; 3) the student model is trained using these pseudo-labels, incorporating Patch-Adaptive Self-Attention (Section 3.3) for feature perturbation; 4) The outputs are aligned with the pseudo-labels using a consistency loss enriched by Negative Class Ranking Loss (Section 3.4).

3. S^{4}Former

3.1. Overall Architecture

In the realm of semi-supervised semantic segmentation, the training dataset D, consisting of $|D|$ images of size $H \times W$, is divided into two disjoint subsets: a labeled dataset D^L with $|D^L|$ images, and an unlabeled set D^U with the remaining images. Each labeled image $x \in D^L$ is paired with the dense one-hot label $y \in \mathbb{R}^{H \times W \times K}$, where K denotes the number of classes. Typically, $|D^L| \ll |D|$ highlights the challenge of scarcity in labeled data.

As shown in Figure 3, our S^{4}Former aligns with the well-established teacher-student [41, 42] framework. Within this framework, S^{4}Former integrates a student model S and teacher model T. Both models S and T adopt a Transformer-based encoder and a decoder to generate segmentation outputs. θ^S, the parameters of S, are learned with all images in the dataset. θ^T, the parameters of T, are updated via Exponential Moving Average (EMA) based on θ^S:

$$\theta_t^T = \mu \cdot \theta_{t-1}^T + (1-\mu) \cdot \theta_t^S, \tag{1}$$

Here, t is the iteration index, and μ is a momentum decay factor indicating the update rate. $\mu = 0$ implies direct copying of parameters from the student model.

During training, we optimize a composite loss function $\mathcal{L} = \mathcal{L}^l + \mathcal{L}^u$ with both labeled and unlabeled images. For labeled images, $\mathcal{L}^l$ is the standard pixel-wise cross-entropy (CE) loss:

$$\mathcal{L}^l = \texttt{CE}(y, S(x; \theta^S)). \tag{2}$$

For unlabeled images, we use different augmented versions x^S and x^T as the student and teacher input. $\mathcal{L}^u$ aligns the teacher and student predictions as:

$$\mathcal{L}^u = \mathbb{I}(C^T > \beta) \cdot \texttt{Consistency}(T(x^T, \theta^T), S(x^S, \theta^S)). \tag{3}$$

In practice, teacher input image x^T receives simple transformations, such as cropping and resizing, while student input image x^S uses stronger augmentations such as CutMix [54]. The teacher model computes the predictions $T(x^T, \theta^T)$ and a confidence map C^T: $C^T = \texttt{max}\left(T(x^T, \theta^T)\right)$. C^T shows the maximum prediction value across classes for each pixel. To ensure that only reliable pseudo-labels guide the student's learning, `Consistency` loss is applied between student and teacher predictions only for those pixels whose C^T exceeds a predefined threshold β.

Our S^{4}Former introduces PatchShuffle for x^S to challenge the student model with shuffled patch sequences (see Section 3.2). Additionally, Patch-Adaptive Self-Attention (PASA) injects feature perturbation to student model S within the self-attention mechanism (see Section 3.3). Furthermore, we modify `Consistency` with Negative Class Ranking (NCR) loss to establish a broader regularization spectrum across all classes (see Section 3.4).

3.2. PatchShuffle

We propose the PatchShuffle augmentation to generate the strong augmented version x^S for an unlabeled image. Vi-

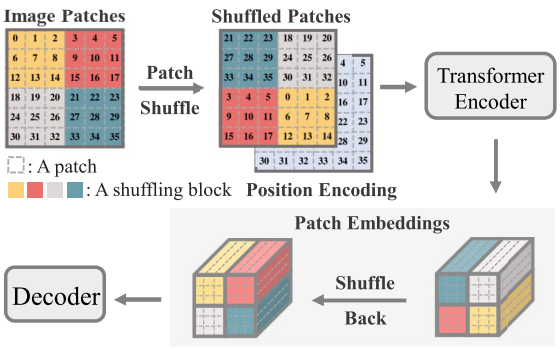

Figure 4. **Illustration of our PatchShuffle process.** An image is divided into an $L \times L$ patch grid (*e.g.*, $L = 6$). Neighboring patches are grouped into $M \times M$ blocks (*e.g.*, $M = 2$) and shuffled before Transformer encoding. The patch embeddings are then shuffled back to their original layout before the decoder, ensuring positional correspondence and edge coherence for the consistency loss computation against the pseudo-label.

sion Transformers take image patches as the input. Beginning with an $L \times L$ grid of image patches, PatchShuffle disrupts the standard spatial arrangement by randomly rearranging these patches. This is performed at a block level: neighboring patches are grouped to get $M \times M$ blocks, each of which is shuffled as a whole, preserving local content while significantly modifying the global structure.

As shown in Figure 4, here we use $L = 6$ and $M = 2$, the Transformer encoder first processes the shuffled patches. Subsequently, the feature embeddings are shuffled back to their original sequence before the decoder. The shuffling back of features restores positional correspondence and edge coherence during the feature up-sampling within the decoder, which is necessary for calculating the consistency loss against pseudo-labels calculated on unshuffled image x^T. Notably, for Transformers with explicit position encodings, the encodings are not shuffled to maintain the model's sensitivity to patch placement.

PatchShuffle's effect is further boosted when combined with additional augmentation strategies such as CutMix [54]. This combination introduces a greater diversity of transformations, thereby enhancing the model's capacity for generalization with more challenging visual scenarios. Please find more discussion in Section 4 and Appendix.

Discussion with "Jigsaw" transformation. The "Jigsaw" transformation [34] also shuffles image splits but the objective is different. While self-supervised learning models with "Jigsaw" transformation aim to predict the original image's layout, our model with PatchShuffle aims to maintain consistent outputs despite input perturbations, in line with the semi-supervised learning framework. Besides, we discuss shuffling patches at the block level based on Vision Transformers, distinguishing from "Jigsaw".

Discussion with position encoding shuffling. For Transformers with explicit position encodings, PatchShuffle can be interpreted as a permutation of position encodings. For the SegFormer [48] architecture which omits explicit position encodings, PatchShuffle still improves. In Section 4.3, we discuss various adjustments made to the position encodings for unlabeled images and show that our implementation of PatchShuffle is particularly effective.

3.3. Patch-Adaptive Self-Attention (PASA)

The self-attention mechanism helps Transformers model global dependencies by attending to all patches of an input image. However, this global view can sometimes lead to the dilution of important local features, especially when handling images with varying confidence levels across patches. PASA dynamically adjusts self-attention weights by incorporating patch-wise confidence, focusing on areas of the image where the model is less certain.

With an $L \times L$ grid of image patches, a standard self-attention layer operates on a sequence of $N = L^2$ input embeddings $\mathbf{e} = [e_1, e_2, \dots e_N]^\mathbf{T}$, $e_i \in \mathbb{R}^d$ as follows:

$$\widetilde{\mathbf{e}} = \underbrace{\text{softmax}\left((\mathbf{e}\mathbf{W}^Q)(\mathbf{e}\mathbf{W}^K)^\mathbf{T}\right)}_{\text{self-attention weight matrix}}(\mathbf{e}\mathbf{W}^V). \tag{4}$$

Here, $\mathbf{W}^Q$, $\mathbf{W}^K$, and $\mathbf{W}^V \in \mathbb{R}^{d \times d'}$ are parameter matrices for queries, keys, and values, respectively. For brevity, scaling, residual connections, and multi-head computations are omitted here; please refer to [43] for details.

As shown in Figure 5, our PASA modulates this attention mechanism by adjusting the attention matrix via:

$$\widetilde{\mathbf{e}} = \text{softmax}\left(\mathcal{M} + (\mathbf{e}\mathbf{W}^Q)(\mathbf{e}\mathbf{W}^K)^\mathbf{T}\right)(\mathbf{e}\mathbf{W}^V). \tag{5}$$

The attention mask matrix $\mathcal{M} \in \mathbb{R}^{N \times N}$ at location (i, j) is influenced by the confidence measure $\bar{C}^T$ of patch i and j:

$$\bar{C}^T = \texttt{PatchAvg}\left(\mathbb{I}(C^T > \beta)\right), \tag{6}$$

where $\mathbb{I}(C^T > \beta) \in \{0, 1\}^{H \times W}$, same as in Equation 3, creates a binary mask of size $H \times W$, with '1' indicating the pixel confidence exceeds the threshold β (β commonly set to 0.95 following prior works [41, 53]).

In our experiments, we set $\mathcal{M}_{ij}$ based on $\bar{C}^T$ empirically as:

$$\mathcal{M}_{ij} = \begin{cases} 0 & \text{if } \bar{C}_i^T > \texttt{Median}(\bar{C}^T), \\ \alpha \cdot (1 - \bar{C}_j^T) & \text{otherwise.} \end{cases} \tag{7}$$

Here, $\alpha > 0$ is a hyper-parameter, and $\texttt{Median}()$ means getting the median value. For patches with high $\bar{C}^T$ (above the median threshold), their attention weights remain unchanged (*i.e.*, add 0). In contrast, patch i with lower confidence $\bar{C}_i^T$ has its attention weights adjusted: $\mathcal{M}_i$ increase its self-attention and mutual attention toward other low-confidence patches. See Section 4 and Appendix for other implementations of $\mathcal{M}_{ij}$.

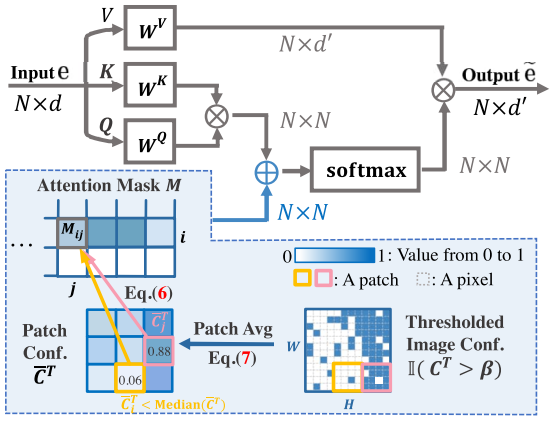

Figure 5. **Illustration of our Patch-Adaptive Self-Attention (PASA).** PASA generates an attention mask $\mathcal{M}$, added to the original self-attention weights, and emphasizes attention among unconfident regions within the image. $\mathcal{M}$, as depicted in blue part, is generated based on the confidence map C_T predicted by the teacher model with Equation 6 and Equation 7.

3.4. Negative Class Ranking (NCR) Loss

Conventionally, the consistency loss is the cross-entropy (CE) loss to encourage the student model's predicted logits p^S to match its pseudo-label $\tilde{y}$. For a reliable pixel, its pseudo-label $\tilde{y}$ is the class that receives the highest logit from the teacher's output $p^T \in \mathbb{R}^K$ (K is the total number of classes) and the CE loss is written as:

$$\mathcal{L}_{CE} = -\log\left(\frac{\exp(p^S_{\tilde{y}})}{\sum_{k=1}^{K}\exp(p^S_k)}\right) \tag{8}$$

The CE loss operates on a contrastive principle, boosting the logit of the positive class (the pseudo-label) while collectively dampening the logits of all other classes (negative). However, this can be sub-optimal as it does not account for the relative distribution of the negative class probabilities.

To improve this, we introduce the Negative Class Ranking (NCR) loss which also accounts for the distribution of negative class probabilities. We first apply a softmax function over the logits of all negative classes to obtain normalized distributions for the teacher and student predictions. The NCR loss is calculated as the L2 distance (unless specified) between these normalized negative class probabilities as:

$$\begin{aligned}\mathcal{L}_{NCR} &= \sum_{k\neq\tilde{y},k\in K}\left(\hat{p}^T_k - \hat{p}^S_k\right)^2, \quad \text{where} \\ \hat{p}^{T\text{ or }S}_k &= \frac{\exp(p^{T\text{ or }S}_k)}{\sum_{k\neq\tilde{y},k\in K}\exp(p^{T\text{ or }S}_k)}\end{aligned} \tag{9}$$

By incorporating the NCR loss, we compel the student model to not only predict the correct class but also to maintain the relative distribution of probabilities for the negative classes. This strategy enhances the model's generalization capabilities by considering a more complete representation of the data distribution.

4. Experiments

To evaluate the effectiveness of our S^{4}Former, we conduct extensive experiments on benchmark datasets and settings. The following subsections describe the details of the experiments, results, and analyses.

4.1. Settings

Datasets. We use three benchmark datasets: Pascal VOC 2012 [14], COCO [30], and Cityscapes [12]. Pascal VOC 2012 includes 1,464 training images and 1,449 validation images across 21 object classes. Following standard augmentation procedures [19, 45], we enrich the training set with an additional 9,118 images from the SBD [18]. The COCO dataset is more challenging, including 118k training and 5k validation images spanning 81 classes. Cityscapes focuses on urban scenes, comprising 2,975 training and 500 validation images of 19 classes.

Experimental Protocol. In the semi-supervised setting, we randomly sample a small fraction of the training images as labeled data, while the remaining images are used as unlabeled data. For Pascal VOC 2012, we adopt two distinct scenarios as defined in prior work [45, 52, 53]: the *classic* regime, which utilizes the original high-quality training set for sampling labeled data, and the *blend* regime, which extends the labeled set with additional samples from the augmented dataset. The labeled/unlabeled split is conducted using a consistent random seed as employed in [9, 53] to maintain fair comparison. Evaluations are conducted using the mean intersection-over-union (mIoU) metric on the validation sets, with center-cropping for Pascal VOC 2012 / COCO and a sliding window technique for Cityscapes, all at single-scale inference.

Implementation Details. We use MMSegmentation [11] for all implementations. All Transformer encoders are ImageNet-1k [39] pre-trained. Batch sizes are set to 16 across datasets, with an even split between labeled and unlabeled images (*i.e*, 8 labeled and 8 unlabeled). Hyperparameters such as learning rate, optimizer, and weight decay follow the framework's defaults. Comprehensive hyperparameter settings are detailed in the Appendix.

4.2. Comparison with State-of-The-Art Methods

S^{4}Former-Base: A Strong Baseline. Introducing Transformers as image encoders, we set a strong baseline with our S^{4}Former-Base. This model harnesses Transformers' power in feature extraction, enriched with robust augmentation strategies like color-jittering and CutMix. As demonstrated in Tables 1, 2, and 3, our baseline improves significantly over the supervised-only ("Sup-Only") results, and

Dataset Split		Pascal VOC *classic*					Pascal VOC *blend*		
Backbone	**Methods**	**1/16 (92)**	**1/8 (183)**	**1/4 (366)**	**1/2 (732)**	**Full (1464)**	**1/16 (662)**	**1/8 (1323)**	**1/4 (2646)**
DeepLabV3+ (ResNet101)	Sup-Only	50.7	59.1	65.0	70.6	74.1	67.5	71.1	74.2
	U^2PL [45]	68.0	69.2	73.7	76.2	79.5	74.4	77.6	78.7
	PS-MT [31]	65.8	69.6	76.6	78.4	–	75.5	78.2	78.7
	iMAS [55]	70.0	75.3	79.1	80.2	82.0	77.2	78.4	79.3
	AugSeg [56]	71.1	75.5	78.8	80.3	81.4	77.0	77.3	78.8
	UniMatch [53]	75.2	77.2	78.8	79.9	81.2	78.1	78.4	79.2
CVT (Dual Backbones)	Sup-Only	47.0	59.7	68.4	73.7	74.6	72.4	74.3	77.9
	SemiCVT [20]	68.6	71.3	75.0	78.5	80.3	78.2	80.0	80.2
SETR (DeiT-Base)	Sup-Only	67.7	72.8	77.4	80.7	82.5	76.6	77.8	79.9
	S^4Former-Base	75.8	77.4	80.0	81.7	83.9	79.0	80.1	80.7
	+ Ours	**80.1**	**81.3**	**82.4**	**83.2**	**85.0**	**79.9**	**80.5**	**81.3**

Table 1. Comparison of mIoU (%) with *state-of-the-art* methods on the Pascal VOC 2012 dataset. Results are presented for two dataset splits following previous works [53, 56]: *classic*, with labeled samples drawn from the original dataset, and *blend*, with labeled samples drawn from the augmented dataset inclusive of SBD. The fractions (*e.g.*, 1/16) and numbers (*e.g.*, 92) denote the proportion and number of labeled images. Best performances for DeepLabV3+ and our architecture are highlighted with underline and **bold**, respectively. "Dual Backbones" used in CVT [20] include both the DeepLabV3+ (ResNet101) and SwinUNet [6].

Dataset Split		1/512	1/256	1/128
Backbone	Methods	(232)	(463)	(925)
DeepLabV3+ (ResNet101)	Sup-Only	22.9	28.0	33.6
	PseudoSeg [61]	29.8	37.1	39.1
	PC^2Seg [59]	29.9	37.5	40.1
	UniMatch [53]	31.9	38.9	44.4
SETR (DeiT-Base)	Sup-Only	31.5	38.3	43.1
	S^4Former-Base	34.5	41.6	46.3
	+ Ours	**35.2**	**43.1**	**46.9**

Table 2. Comparison with *state-of-the-art* methods on the COCO dataset. Best performances for DeepLabV3+ and our architecture are highlighted with underline and **bold**, respectively.

Dataset Split		1/16	1/8	1/4	1/2
Backbone	Methods	(186)	(372)	(744)	(1488)
DeepLabV3+ (ResNet101)	Sup-Only	66.3	72.8	75.0	78.0
	U^2PL [45]	74.9	76.5	78.5	79.1
	PS-MT [31]	–	76.9	77.6	79.1
	AugSeg [56]	75.2	77.8	79.6	80.4
	UniMatch [53]	76.6	77.9	79.2	79.5
CVT (Dual Back.)	Sup-Only	67.2	73.1	75.1	78.6
	SemiCVT [20]	72.2	75.4	77.2	79.6
SegFormer (MiT-B4)	Sup-Only	73.3	77.0	79.1	80.4
	S^4Former-Base	76.7	79.2	80.1	80.6
	+ Ours	**78.5**	**79.9**	**80.6**	**80.9**

Table 3. Comparison with *state-of-the-art* methods on Cityscapes dataset. Best performances for DeepLabV3+ and our architecture are highlighted with underline and **bold**, respectively. "Dual Back." used in CVT [20] include both the DeepLabV3+ (ResNet101) and SwinUNet [6].

matches or even outperforms state-of-the-art methods, highlighting the Transformer's capability to utilize contextual information from unlabeled images effectively.

Results for Pascal VOC 2012. Table 1 showcases that our methods consistently boost model performance. The notable improvements over S^4Former-Base when augmented with our proposed components ("+ Ours"), underscore our effective utilization of unlabeled data to enhance semi-supervised learning. For the most challenging scenario with only 92 labeled images, our model achieves 80.1%, surpassing the previous best model by 4.9%. Compared to SemiCVT [20] which combines ConvNet and Vision Transformer architecture, our improvements are impressive. In contrast to SemiCVT, our Vision Transformer is pre-trained on ImageNet-1k and trained with our specific training mechanisms. Furthermore, we have applied our components to state-of-the-art methods, *i.e.*, UniMatch [53] and Augseg [56], transplanting them to a Transformer backbone. As illustrated in Figure 6, our approach consistently surpasses the prior best models, yielding an average increase of 1.7% on UniMatch and 1.2% on AugSeg. These gains further underscore the robustness and adaptability of our methods across different architectures and settings.

Results for COCO. With its extensive variety of 81 classes, the COCO dataset presents a more challenging setting for semi-supervised learning. Despite this complexity, as indicated in Table 2, our S^4Former-Base with additional components ("+ Ours") demonstrates a notable improvement over the supervised-only ("Sup-Only") results, validating the effectiveness of our approach in harnessing unlabeled data even in diverse and challenging datasets. Applied to UniMatch and Augseg methods, our adaptations to a Trans-

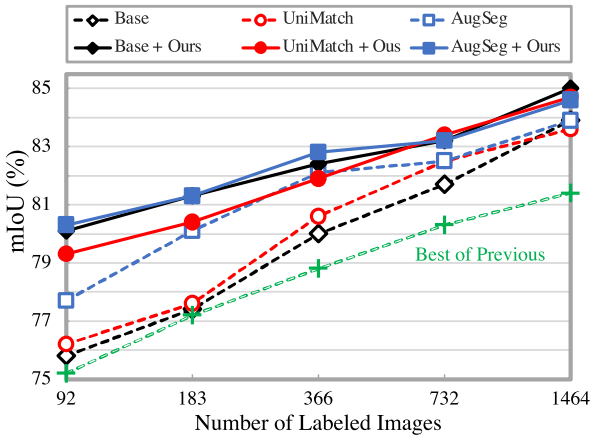

Figure 6. Comparison over S^{4}Former-Base, Unimatch [53] and Augseg [56] with our proposed components ("+ Ours") on Pascal VOC 2012 *classic* settings.

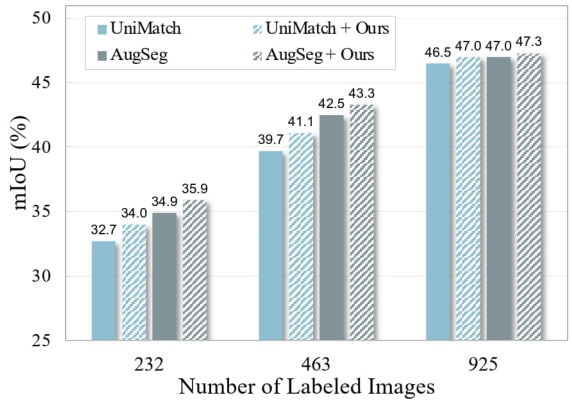

Figure 7. Comparison over UniMatch [53] and Augseg [56] with our proposed components ("+ Ours") on the COCO dataset.

former backbone lead to an average mIoU increase of 1.1% and 0.7%, respectively, as shown in Figure 7.

Results for Cityscapes. We employed the efficient SegFormer (MiT) [48] architecture due to the high computational demands of SETR (DeiT) [57] on Cityscapes. Table 3 shows that our S^{4}Former-Base, together with the "+ Ours" enhancements achieves consistent performance over other state-of-the-art methods. This underlines the advantages of our approach in a real-world urban scene understanding.

4.3. Additional Results

Effectiveness of Each Component. Table 4 presents the separate and combined effects of our proposed components, PatchShuffle, Patch-Adaptive Self-Attention (PASA), and Negative Class Ranking (NCR) loss, across different partitions of labeled data. One can see that each component individually enhances the performance over the baseline, with the cumulative effect of combining all components yielding even greater gains. This suggests that our components work in concert to improve the learning process.

Methods	**Split = 1/16**	**1/8**	**1/4**
SupOnly	67.7	72.8	77.4
Base	75.8	77.4	80.0
+PatchShuffle	78.4 (+2.6)	80.2 (+2.8)	82.1 (+2.1)
+PASA	77.9 (+2.1)	79.4 (+2.0)	81.3 (+1.3)
+NCR	77.5 (+1.7)	78.7 (+1.3)	81.9 (+1.9)
+All	80.1 (+4.3)	81.3 (+3.9)	82.4 (+2.4)

Table 4. The individual improvements over mIoU of our proposed components. Results for the Pascal VOC 2012 *classic* settings.

Split	Base	woPE	avgPE	dupPE	PS
1/16	75.8	76.2	77.3	77.0	**78.4**
1 / 8	77.4	77.0	78.6	76.2	**80.2**

Table 5. Comparison of mIoU using different operations of position embeddings against our PatchShuffle (PS) approach. Given explicit position encodings, *PS* is equivalent to a block-level permutation of these position encodings. Results for the Pascal VOC 2012 *classic* settings.

Different Position Encoding Adjustments. Table 5 evaluates different position encoding adjustments for unlabeled student images x^S, as discussed in Section 3.2. We assess the effects of discarding all position encodings (*woPE*), averaging them within shuffling blocks and applying the average to all patches within each block (*avgPE*), and duplicating them across blocks (*dupPE*). Our PatchShuffle, equivalent to a block-level permutation of position encodings, yields superior performance improvements, demonstrating its effectiveness over the alternate encoding strategies.

Different Attention-Mask Adjustments. The efficacy of our Patch-Adaptive Self-Attention (PASA) is underscored in Table 6, compared with different attention mask adjustment approaches and feature dropout (FD) in the previous method [53]. The *Rand* strategy uses random scaled numbers to adjust the mask, working as the naïve attention perturbation to the student model. In contrast to PASA, *Reverse1* and *Reverse2* reversely increase focus on confident regions. *Reverse1* emphasizes the self-attention of confident regions within high-confidence areas, and *Reverse2* encourages less confident regions to focus more on high-confidence counterparts (See Appendix for details). By comparing *Rand*, *Reverse1*, *Reverse2*, and *FD* with ours, we underline the significance of directing attention to less confident regions.

Different Regularization Loss. Table 7 compares different methods for the regularizing loss over unlabeled images. *Base* employs the standard cross entropy (CE) loss, and *Soft* applies the cross entropy loss with soft labels as in [50] (See Appendix for detail). *All-CR* denotes the approach of apply-

Split	Base	Rand	Reverse1	Reverse2	FD	Ours
1/16	75.8	76.2	76.1	76.2	76.7	**77.9**
1 / 8	77.2	78.2	78.4	77.8	77.8	**79.4**

Table 6. Comparison of mIoU across different attention-mask adjustment strategies and feature dropout (FD). Results for the Pascal VOC 2012 *classic* settings.

Split	Base	Soft	All-CR	NCR (L2)	NCR (KL)
1/16	75.8	74.8	74.9	77.5	**78.0**
1 / 8	77.0	78.0	78.1	**78.7**	78.6
1 / 4	80.0	80.7	80.4	**81.9**	81.4

Table 7. Comparison of mIoU with different regularization losses. Results for the Pascal VOC 2012 *classic* settings.

	Methods	Split = 1/16	1/8
SegFormer (MiT-B4)	Sup-Only	68.2	67.4
	S^4Former-Base	75.7	78.8
	+ Ours	**77.1** (+1.4)	**80.4** (+1.6)
	UniMatch	76.8	78.9
	+ Ours	**77.6** (+0.8)	**79.5** (+0.6)
	AugSeg	77.4	80.2
	+ Ours	**77.8** (+0.4)	**80.6** (+0.4)

Table 8. Effectiveness of integrating our proposed methods with the SegFormer [48] backbone. "+ Ours" denotes the integration of our components into existing methods. Results for the Pascal VOC 2012 *classic* settings.

ing L2 distance loss across all classes, including both "positive" and "negative" classes. Our *NCR*, focusing on "negative" classes, excels over these alternatives. We also try KL loss to regularize the distribution among negative classes in Equation 9, which reaches comparable performance as the L2 loss we used in default.

Backbone Compatibility. Our methodology's compatibility with different Transformer backbones is confirmed with SegFormer [48] in Table 8. The consistent performance improvements with the proposed components validate the robustness and versatility of our approach.

Ablation on Hyperparameters. Table 9 and Table 10 present the ablation on the number of shuffling blocks M in Section 3.2 and the coefficient weight α in Equation 7, respectively. Table 9 shows PatchShuffle's robustness across varying numbers of shuffling blocks M, consistently improving over the baseline (M=1). Our method achieves improvements across various settings, and in practice, we use $M = 4$ and $\alpha = 5$ to obtain the best performance.

4.4. Qualitative Results

We provide qualitative results in Figure 8. With our training components, S^4Former correctly separates cow from horse in the top row, detects and segments more chairs in the bottom row. We also show the attention weights on the bottle-neck image feature of the interested patch (the yellow box). We show that our components effectively improve the network's ability to attend to the relevant areas across the whole image and generate more accurate predictions.

Split	M=1	2	4	8	16	32
1/16	75.8	77.1	**78.4**	78.3	78.0	77.5
1 / 8	77.4	80.0	**80.2**	79.7	79.5	79.6

Table 9. Comparison of mIoU with different values of shuffling blocks M. $M = 1$ means no shuffling (*i.e.*, S^4Former-Base). Results for the Pascal VOC 2012 *classic* settings.

Split	$\alpha = 0$	2	5	10	20
1/16	75.8	76.4	**77.9**	77.6	77.0
1 / 8	77.0	78.7	79.4	**80.0**	79.4

Table 10. Comparison of mIoU with different values of α. $\alpha = 0$ means no attention adjustment is applied (*i.e.*, S^4Former-Base). Results for the Pascal VOC 2012 *classic* settings.

Figure 8. Comparative visual results on Pascal VOC 2012 *classic* setting with a limited set of 92 labeled images. We show segmentation predictions as well as the corresponding attention-weight heatmaps for a selected patch (denoted by the yellow box in the ground truth). The red region corresponds to a high contribution. More illustrations are shown in the Appendix.

5. Conclusion

Despite the remarkable success in a broad range of vision tasks, Vision Transformers have not yet been explored in semi-supervised semantic segmentation (S^4). In this paper, we present S^4Former, a simple yet strong framework that combines the power of Vision Transformers with the conventional teacher-student paradigm. Based on S^4Former, we introduce PatchShuffle, Patch-Adaptive Self-Attention (PASA), and Negative Class Ranking (NCR) loss, tailor the training regularization from image, feature, and output ends, respectively. Armed with those training strategies, S^4Former demonstrates the potential of utilizing Vision Transformers for S^4 with state-of-the-art results. We hope our S^4Former will not only serve as a robust foundation for subsequent research in S^4 but will also inspire innovative approaches to leverage Vision Transformers in similar contexts with limited labeled data.

References

[1] Iñigo Alonso, Alberto Sabater, David Ferstl, Luis Montesano, and Ana C Murillo. Semi-Supervised Semantic Segmentation With Pixel-Level Contrastive Learning From a Class-Wise Memory Bank. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[2] David Berthelot, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. ReMixMatch: Semi-Supervised Learning with Distribution Alignment and Augmentation Anchoring. In *International Conference on Learning Representations (ICLR)*, 2019. 2

[3] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 2

[4] Quentin Bouniot, Angélique Loesch, Amaury Habrard, and Romaric Audigier. Towards Few-Annotation Learning for Object Detection: Are Transformer-based Models More Efficient? In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023. 1

[5] Zhaowei Cai, Avinash Ravichandran, Paolo Favaro, Manchen Wang, Davide Modolo, Rahul Bhotika, Zhuowen Tu, and Stefano Soatto. Semi-supervised Vision Transformers at Scale. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1

[6] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation. In *Proceedings of the European Conference on Computer Vision Workshops(ECCVW)*, 2022. 6

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-End Object Detection with Transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[8] Chun-Fu (Richard) Chen, Quanfu Fan, and Rameswar Panda. CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[9] 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 (CVPR)*, 2021. 2, 5

[10] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-Pixel Classification is Not All You Need for Semantic Segmentation. 2021. 2

[11] MMSegmentation Contributors. MMSegmentation: OpenMMLab Semantic Segmentation Toolbox and Benchmark. https://github.com/open-mmlab/mmsegmentation, 2020. 5

[12] 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. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. 2

[14] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes Challenge: A Retrospective. *International Journal of Computer Vision (IJCV)*, 2015. 1, 5

[15] Jiashuo Fan, Bin Gao, Huan Jin, and Lihui Jiang. UCC: Uncertainty Guided Cross-Head Co-Training for Semi-Supervised Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[16] Zhengyang Feng, Qianyu Zhou, Guangliang Cheng, Xin Tan, Jianping Shi, and Lizhuang Ma. Semi-Supervised Semantic Segmentation via Dynamic Self-Training and Class-Balanced Curriculum. *ArXiv*, abs/2004.08514, 2020. 2

[17] Geoff French, Samuli Laine, Timo Aila, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, varied perturbations. In *British Machine Vision Conference (BMVC)*. BMVA Press, 2020. 2

[18] Bharath Hariharan, Pablo Arbel'aez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic Contours from Inverse Detectors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2011. 1, 5

[19] Hanzhe Hu, Fangyun Wei, Han Hu, Qiwei Ye, Jinshi Cui, and Liwei Wang. Semi-Supervised Semantic Segmentation via Adaptive Equalization Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 1, 2, 5

[20] Huimin Huang, Shiao Xie, Lanfen Lin, Ruofeng Tong, Yen-Wei Chen, Yuexiang Li, Hong Wang, Yawen Huang, and Yefeng Zheng. SemiCVT: Semi-Supervised Convolutional Vision Transformer for Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 6

[21] Ying Jin, Jiaqi Wang, and Dahua Lin. Semi-Supervised Semantic Segmentation via Gentle Teaching Assistant. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2

[22] Salman H. Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *CoRR*, 2021. 2

[23] Donghyeon Kwon and Suha Kwak. Semi-Supervised Semantic Segmentation With Error Localization Network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[24] Xin Lai, Zhuotao Tian, Li Jiang, Shu Liu, Hengshuang Zhao, Liwei Wang, and Jiaya Jia. Semi-Supervised Semantic Segmentation with Directional Context-Aware Consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[25] Samuli Laine and Timo Aila. Temporal Ensembling for Semi-Supervised Learning. In *International Conference on Learning Representations (ICLR)*, 2016. 2

[26] Dong-Hyun Lee. Pseudo-Label: The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks. *ICML 2013 Workshop: Challenges in Representation Learning (WREPL)*, 2013. 2

[27] Dong-Hyun Lee, Joonseok Kim, and Junmo Kang. Pseudo-Label: The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks. In *International Conference on Machine Learning (ICML) Workshop*, 2013. 1, 2

[28] Kwonjoon Lee, Junsoo Lee, and Jinwoo Shin. Pseudo-Labeling and Confirmation Bias in Deep Semi-Supervised Learning. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, 2020. 2

[29] Zhengyu Liang, Yingqian Wang, Longguang Wang, Jungang Yang, and Shilin Zhou. Light field image super-resolution with transformers. *IEEE Signal Processing Letters*, 2022. 2

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. *European Conference on Computer vision (ECCV)*, 2014. 5

[31] Yuyuan Liu, Yu Tian, Yuanhong Chen, Fengbei Liu, Vasileios Belagiannis, and Gustavo Carneiro. Perturbed and Strict Mean Teachers for Semi-Supervised Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 6

[32] 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 (ICCV)*, 2021. 2

[33] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing Properties of Vision Transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2

[34] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 4

[35] 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 (CVPR)*, pages 7847–7856, 2020. 2

[36] Mengyang Pu, Yaping Huang, Yuming Liu, Qingji Guan, and Haibin Ling. EDTER: Edge Detection With Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[37] Pengchong Qiao, Zhidan Wei, Yu Wang, Zhennan Wang, Guoli Song, Fan Xu, Xiangyang Ji, Chang Liu, and Jie Chen. Fuzzy Positive Learning for Semi-Supervised Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2

[38] Nayeem Rizve, Kevin Duarte, Yogesh Rawat, and Mubarak Shah. In Defense of Pseudo-Labeling: An Uncertainty-Aware Pseudo-label Selection Framework for Semi-Supervised Learning. In *International Conference on Learning Representations (ICLR)*, 2021. 2

[39] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 2015. 2, 5

[40] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. 2

[41] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2, 3, 4

[42] 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 (NeurIPS)*, 2017. 1, 2, 3

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems (NeurIPs)*, 2017. 4

[44] Xiaoyang Wang, Bingfeng Zhang, Limin Yu, and Jimin Xiao. Hunting Sparsity: Density-Guided Contrastive Learning for Semi-Supervised Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[45] 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 International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 5, 6

[46] Zicheng Wang, Zhen Zhao, Xiaoxia Xing, Dong Xu, Xiangyu Kong, and Luping Zhou. Conflict-Based Cross-View Consistency for Semi-Supervised Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[47] Linshan Wu, Leyuan Fang, Xingxin He, Min He, Jiayi Ma, and Zhun Zhong. Querying Labeled for Unlabeled: Cross-Image Semantic Consistency Guided Semi-Supervised Semantic Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[48] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2, 4, 7, 8

[49] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-Training With Noisy Student Improves ImageNet Classification. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition (CVPR), 2020. 2

[50] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-End Semi-Supervised Object Detection with Soft Teacher. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1, 7

[51] Bin Yan, Yi Jiang, Peize Sun, Dong Wang, Zehuan Yuan, Ping Luo, and Huchuan Lu. Towards Grand Unification of Object Tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2

[52] 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 (CVPR)*, 2022. 2, 5

[53] Lihe Yang, Lei Qi, Litong Feng, Wayne Zhang, and Yinghuan Shi. Revisiting Weak-to-Strong Consistency in Semi-Supervised Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2, 4, 5, 6, 7

[54] 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 (ICCV)*, 2019. 2, 3, 4

[55] Zhen Zhao, Sifan Long, Jimin Pi, Jingdong Wang, and Luping Zhou. Instance-specific and Model-adaptive Supervision for Semi-supervised Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 6

[56] Zhen Zhao, Lihe Yang, Sifan Long, Jimin Pi, Luping Zhou, and Jingdong Wang. Augmentation Matters: A Simple-yet-Effective Approach to Semi-supervised Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2, 6, 7

[57] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H.S. Torr, and Li Zhang. Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 7

[58] Xu Zheng, Yunhao Luo, Hao Wang, Chong Fu, and Lin Wang. Transformer-CNN Cohort: Semi-supervised Semantic Segmentation by the Best of Both Students, 2022. 2

[59] 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 (ICCV)*, 2021. 2, 6

[60] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable Transformers for End-to-End Object Detection. *arXiv preprint arXiv:2010.04159*, 2020. 2

[61] Yuliang Zou, Zizhao Zhang, Han Zhang, Chun-Liang Li, Xiao Bian, Jia-Bin Huang, and Tomas Pfister. PseudoSeg: Designing Pseudo Labels for Semantic Segmentation. In *International Conference on Learning Representations (ICLR)*, 2021. 1, 2, 6