Semi-Supervised Vision Transformers

Zejia Weng[1,2*], Xitong Yang[3*], Ang Li[4],
Zuxuan Wu[1,2†], Yu-Gang Jiang[1,2†]

[1] Shanghai Key Lab of Intell. Info. Processing, School of CS, Fudan University
[2] Shanghai Collaborative Innovation Center on Intelligent Visual Computing
[3] Meta AI [4] Baidu Apollo

Abstract. We study the training of Vision Transformers for semi-supervised image classification. Transformers have recently demonstrated impressive performance on a multitude of supervised learning tasks. Surprisingly, we show Vision Transformers perform significantly worse than Convolutional Neural Networks when only a small set of labeled data is available. Inspired by this observation, we introduce a joint semi-supervised learning framework, `Semiformer`, which contains a transformer stream, a convolutional stream and a carefully designed fusion module for knowledge sharing between these streams. The convolutional stream is trained on limited labeled data and further used to generate pseudo labels to supervise the training of the transformer stream on unlabeled data. Extensive experiments on ImageNet demonstrate that `Semiformer` achieves 75.5% top-1 accuracy, outperforming the state-of-the-art by a clear margin. In addition, we show, among other things, `Semiformer` is a general framework that is compatible with most modern transformer and convolutional neural architectures. Code is available at https://github.com/wengzejia1/Semiformer.

Keywords: Vision Transformers; CNNs; Semi-Supervised Learning

1 Introduction

Vision transformers (ViT) have achieved remarkable performance recently on a variety of supervised computer vision tasks [8, 15, 18]. Their success is largely fueled by high capacity models with self-attention layers trained on massive data. However, it is not always feasible to collect sufficient annotated data in many real world applications. When only a small number of labeled samples are provided, semi-supervised learning (SSL) [4, 41] is a powerful paradigm to achieve better performance by leveraging a huge amount of unlabeled data. Despite the success of Vision Transformers in fully supervised scenarios, the understanding of its effectiveness in SSL is still an empty space.

We perform a series of studies with Vision Transformers (ViT) [8] in the semi-supervised learning (SSL) setting on ImageNet. Surprisingly, the results show that simply training a ViT using a popular SSL approach, FixMatch [23], still

*Equal contributions. †Corresponding author.

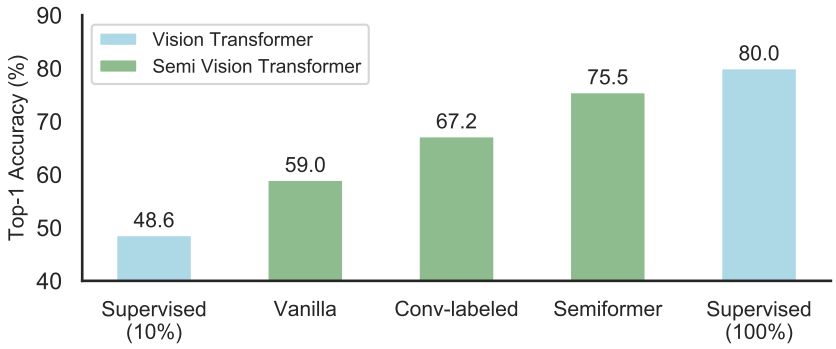

Fig. 1. Three semi-supervised vision transformers using 10% labeled and 90% unlabeled data (colored in green) vs. fully supervised vision transformers (colored in blue) using 10% and 100% labeled data. Our approach `Semiformer` achieves competitive performance, 75.5% top-1 accuracy.

leads to much worse performance than a CNN trained even without FixMatch. We believe this results from the fact that pseudo labels from CNNs are more accurate, possibly due to their encoded inductive bias.

To validate our hypothesis, we use CNNs to produce pseudo labels for the joint semi-supervised training of CNNs and transformers. By doing so, we are able to significantly improve the top-1 accuracy of the ViT by 8+% (c.f. `Conv-labeled` and `Vanilla` in Fig. 1). This highlights that labels derived from CNNs are also helpful for training transformers under the SSL setting. While pseudo labels from CNNs are effective, the final ViT is still slightly weaker than the "teacher" CNN. We posit that simply performing pseudo labeling (PL) with CNNs to derive supervisory signals for transformers is not sufficient. Instead, we hypothesize that a joint knowledge sharing mechanism at the architecture level is required to fully explore knowledge in CNNs.

In light of these, we introduce a novel semi-supervised learning framework for Vision Transformers, which we term as `Semiformer`. In particular, `Semiformer` composes of a convolutional stream and a transformer stream. It leverages labels produced by CNNs as supervisory signals to train the CNN and the transformers jointly using a popular SSL strategy. The two streams are further connected with a cross-stream feature interaction module, enabling streams to complement each other. Benefited from more accurate labels and the interaction design, `Semiformer` can be readily used for SSL.

We conduct extensive experiments to evaluate `Semiformer`. In particular, `Semiformer` achieves 75.5% top-1 accuracy on ImageNet and outperforms the state-of-the-art using 10% of labeled samples. We also show `Semiformer` outperforms alternative methods by clear margins under different labeling ratios. In addition, we empirically demonstrate `Semiformer` is a generic framework compatible with modern CNN and transformer architectures. We also pro-

vide qualitative evidence that `Semiformer` is better than ViTs in the SSL setting.

Contributions. Our contributions are three-folded:

1. We are the first to investigate the application of Vision Transformers for semi-supervised learning. We reveal that Vision Transformers perform poorly when labeled samples are limited, yet they can be improved by utilizing unlabeled data together with the help from Convolutional neural networks.
2. We propose a generic framework `Semiformer` for the semi-supervised learning of Vision Transformers, which not only explores predictions as supervisory signals but also feature-level clues from CNNs to improve the ViTs in the low-data learning regime.
3. We perform extensive experiments and studies to evaluate `Semiformer`. `Semiformer` achieves 75.5% top-1 accuracy on ImageNet and outperforms state-of-the-art methods in semi-supervised learning. Additional ablation studies are further conducted to understand its effectiveness.

2 Related work

Vision Transformers. A variety of Vision Transformers [8,15,17,25,28–30,37] have refreshed the state-of-the-art performance on ImageNet, demonstrating their powerful representation capability in solving vision tasks. Among them, the Vision Transformer (ViT) [8] is the first to prove that purely using the transformer structure can perform well on image classification tasks. It divides each image into a sequence of patches and then applies multiple transformer layers [27] to model their global relations. T2T-ViT [37] recursively aggregates neighboring tokens into one token for better modeling of local structures such as edges and lines among neighboring pixels, which outperforms ResNets [13] and also achieves comparable performance to light CNNs by directly training on ImageNet. Swin Transformer [18] creates a shifted windowing scheme cooperated with stacked local transformers for better information interaction among patches. With the continuous improvements of Vision Transformers, transformer based networks have achieved higher accuracy on medium-scale and large-scale datasets. Although transformers have been proven effective at solving visual tasks, it is known inferior to some CNNs when training from scratch on small-sized datasets mainly because ViTs lack image-specific inductive bias [8].

Touvron *et al*. [25] distill the knowledge of CNNs to ViTs, easing the training process of transformers to be more data efficient. The hard distillation idea is similar to the pseudo label approach in SSL. However, it differs from our work in that the teacher model in distillation is pre-trained in a fully supervised setting and frozen while we also use the pseudo labels to continuously updating the convolutional stream in our framework.

Semi-supervised learning. Effective supervised learning using deep neural networks usually requires annotating a large amount of data. However, creating such large datasets is costly and labor-intensive. A promising solution is

SSL, which leverages unlabeled data to improve model performance. Existing SSL methods are designed from the aspects of pseudo labeling where model predictions are converted to hard labels (*e.g.*, [16, 22, 36]), and consistency regularization where the model is constrained to have consistent outputs under different perturbations [1, 2, 21, 24, 34]. FixMatch [23] combines these two classic semi-supervised learning strategies. It predicts hard pseudo labels under weak perturbations and guides the model to learn on unlabeled data with strong perturbations. Our work is built upon FixMatch to explore the potential of semi-supervised Vision Transformers. The noisy student [35] extends the idea of self-training and distillation with larger student models and add noise to the student. [39] applies transformers to automated speech recognition using semi-supervised learning. Their superior performance is obtained by large scale pre-training and iterative self-training using the noisy student training approach.

As the advances of self-supervised learning approaches [3, 5], a new trend for semi-supervised learning becomes first utilizing the large scale unlabeled data for self-supervised pre-training and then use the labeled data for fine-tuning. Chen *et al*. [6] show that a big ResNet pre-trained using SimCLRv2 can achieve competitive semi-supervised performance after fine-tuning.

Joint modeling of CNNs and Transformers. CNNs and Transformers use two different ways to enforce geometric structure priors. A convolution operator is applied on patches of an image, which naturally results in a local geometric inductive bias. However, a Vision Transformer model utilizes the global self-attention to learn the relationships between global image elements [8]. From a complementary point of view, combining the advantages of CNNs in processing local visual structures and the advantages of transformer in processing global relationships is potentially a better approach for image modeling.

One research direction is to imitate the CNN operations into a Vision Transformer or vice versa [15, 30, 33, 37]. For example, Pooling-based Vision Transformer (PiT) [15] applies pooling operations to shrink the feature maps and gradually increases the channel dimension at the same time, similar to the practice of CNN. PyramidViT [30] and CvT [32] also adopt a similar hierarchical design. T2T-ViT [37] designs a progressive tokenization module to aggregate neighboring tokens. [33] replaces the ViT stem by a small number of stacked convolutions and observes it improves the stability of model training. They also keep the network deep and narrow, inspired by CNNs.

Probably the most relevant approaches are [9, 12, 19, 31] that aim to find ways to combine convolution and transformer into a single model. For example, the non-local network [31] adds self-attention layers to CNN backbones. Speech-Conformer [12] attempts to use convolution to enhance the capabilities of the transformer, while ConVit [9] introduces gated positional self-attention (GPSA) module which becomes equipped with a "soft" convolutional inductive bias. VisualConformer [19] decouples CNN and Vision Transformer streams and design a module for feature communication across streams. However, these studies are all focused on supervised learning while we propose a generic framework for training semi-supervised vision transformers. Another major difference lies in that, even

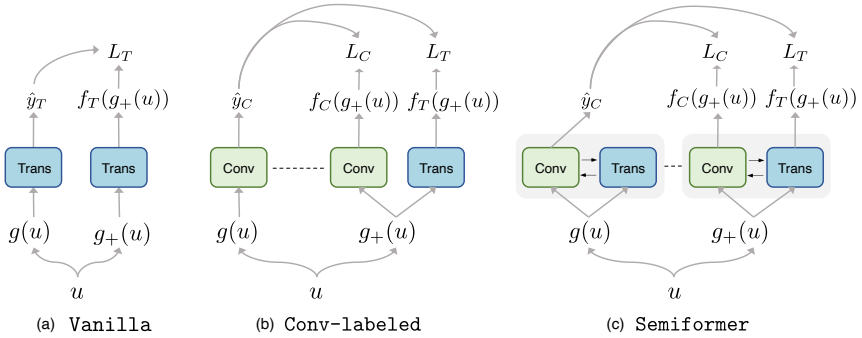

Fig. 2. We explore a variety of ways to apply vision transformer into semi-supervised learning task. Dotted line refers to weights sharing. u refers to the input image, g and g_+ refer to weak and strong data augmentation. $\hat{y}_T$ and $\hat{y}_C$ refer to pseudo labels produced by transformer and convolutional streams. $f_T(\cdot)$ and $f_C(\cdot)$ represent model predictions of transformer and convolutional streams respectively.

though we follow the same direction of fusing convolutions and transformers, our approach does not treat the combined architecture as an entirety, *e.g.*, the pseudo labels have to be generated by the convolutional stream only.

3 A Study with Vision Transformers for SSL

We start by presenting two frameworks that use pseudo labels for SSL. Although the two attempts are surprisingly unsatisfactory, their results reveal two important lessons which eventually inspire us to develop our framework. Below we provide the details of the two studies and our learned lessons.

Unlabeled data improves Vision Transformers. A natural approach to leverage unlabeled data is to do pseudo labeling through Vision Transformers. Our first hypothesis is that *a Vision Transformer can be improved when the total number of input-output training pairs increases (though many of them are pseudo labels)*. We verify this with a `Vanilla` framework, which uses the same architecture (*e.g.*, CNN or Transformer) and builds upon FixMatch [23] for SSL. In particular, FixMatch uses two types of augmentations, a strong one and a weak one. The pseudo label of the unlabeled data is obtained by applying the model on weakly augmented images. And the model is trained using the strongly augmented inputs with the pseudo labels.

Results in Tab. 1 show that after adding the other 90% images from the ImageNet as unlabeled training data, the transformer-based model can have an accuracy improvement by 10.4%, which is greater than the accuracy improvement of CNN's 8.3%. This validates our hypothesis, *i.e.*, large-scale data helps the Vision Transformer to learn better even when many of them are pseudo

Table 1. Results and comparisons with two different SSL frameworks, and comparisons with the supervised baselines.

Architecture	Method	Top-1 Acc (%)
CNN	`Sup.` only (10%)	60.2
	`Vanilla`	**68.5**
Transformer	`Sup.` only (10%)	48.6
	`Vanilla`	59.0
	`Conv-labeled`	67.2

labeled. However, despite the score increases, the performance of Vision Transformers in semi-supervised learning is still unsatisfactory, even inferior to the accuracy of fully supervised CNN training on only 10% of the labeled data.

Pseudo labels from CNNs are more accurate. We suspect that *the weak performance of* `Vanilla` *is due to the inaccurate pseudo labels generated by the transformer*. Vision Transformer contains less image-specific inductive bias, leading to poor performance on small-scale data and thus requires more data for representation learning. In contrast, CNNs are shown to possess strong image-specific inductive bias due to its convolution and pooling design. A natural question is: *what if we use a CNN to generate pseudo labels for Vision Transformer?*

We introduce a new SSL framework, `Conv-labeled`, which uses labels from CNNs for the SSL of CNN and transformers jointly, as illustrated in Fig. 2(b). As is seen in Tab. 1, the `Conv-labeled` approach results in 67.2% top-1 accuracy using the predictions from the ViT on ImageNet, improving the `Vanilla` approach by 8.2%, which suggests that CNNs provide better pseudo labels.

Conv-based pseudo labeling is not enough. Although the ViT's performance is boosted by a CNN pseudo-label generator, the final performance of the ViT (67.2%) is still worse than the CNN (68.5%), observed from Tab. 1. This suggests that the knowledge from the CNN is not yet fully utilized through the simple pseudo labeling approach. One major problem here is that the two models are mostly decoupled except for the unilateral supervision given by the CNN. On the one hand, knowledge from the CNN is not directly injected into the transformer model. On the other hand, the CNN does not gain any information from the ViT. This motivates us to consider jointly modeling both a convolution network and a transformer, which becomes the proposed `Semiformer` framework.

4 Our Approach: `Semiformer`

We introduce `Semiformer` (illustrated in Fig. 2(c) and Fig. 3), which jointly fuses a CNN architecture and a Transformer for semi-supervised learning.

Notation. We use $f(x;\theta)$ to represent the mapping function of our `Semiformer`, given the input x and the model parameter θ. $f_T(x)$ and $f_C(x)$

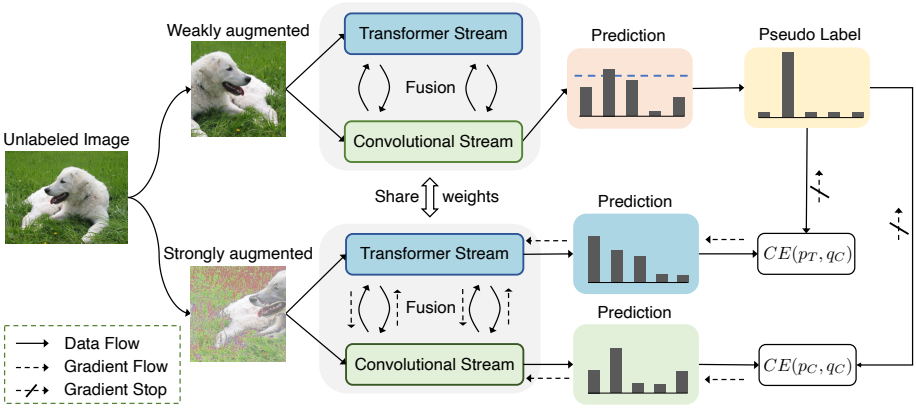

Fig. 3. Diagram of the `Semiformer` framework. For an unlabeled image, its weakly-augmented version (top) is fed into the model. The prediction of CNN is used for generating pseudo labels with a confidence threshold (dotted line). Then we compute the model's prediction for the strong augmented version of the same image (bottom). We expect both transformer and convolutional streams to match the pseudo label via cross-entropy losses. Streams complement each other with the feature-level modules.

are the vectorized output probability for each label from the transformer stream and the convolutional stream, respectively, and θ is omitted for simplicity. Additionally, a weak data augmentation function $g(\cdot)$ and a strong data augmentation function $g_+(\cdot)$ are used in our approach. We assume the semi-supervised dataset contains N_l labeled examples and N_u unlabeled examples. We use index i for labeled data, index j for unlabeled and index k for the label space.

Loss for labeled data. Formally, the total loss for labeled data is

$$\mathcal{L}_l = \sum_{i=1}^{N_l} \mathcal{L}_{xent}(y_i, f_T(g_+(x_i))) + \mathcal{L}_{xent}(y_i, f_C(g_+(x_i))) \tag{1}$$

where x_i is the i-th labeled example and y_i is its corresponding one-hot label vector. $\mathcal{L}_{xent}$ is the cross-entropy loss function, *i.e.*, $\mathcal{L}_{xent}(p, q) = \sum_k p_k \log q_k$.

Loss for unlabeled data. Given an unlabeled image u_j, we first perform strong data augmentation $g_+(\cdot)$ and weak data augmentation $g(\cdot)$, according to FixMatch [23], to obtain the two views of the same input image. However, only the prediction output of the convolutional stream $f_C(g(u_j))$ is used to generate the pseudo label. The probability p_j of an unlabeled input u_j becomes

$$p_j = f_C(g(u_j)) \ . \tag{2}$$

We define the pseudo labels as the class with maximum probability, *i.e.*, $\hat{p}_j = \arg\max_k p_{jk}$. We use $\hat{y}_j$ to represent the one-hot vector corresponding to pseudo label $\hat{p}_j$. These pseudo labels will in turn be used to calculate the cross entropy loss to back-propagate both the convolutional and the transformer streams with

the strongly augmented inputs $g_+(u_j)$. A filtering by threshold $\max_k p_{jk} \geq \tau$, equivalent to $\langle \hat{y}_j, p_j \rangle \geq \tau$, is applied to remove pseudo labels without sufficient certainty. The remaining pseudo labels are used to guide the semi-supervised learning. The total loss for unlabeled data becomes

$$\mathcal{L}_u = \sum_{j=1}^{N_u} \left(\mathcal{L}_{xent}(\hat{y}_j, f_T(g_+(u_j))) + \mathcal{L}_{xent}(\hat{y}_j, f_C(g_+(u_j))) \right) \delta[\langle \hat{y}_j, p_j \rangle \geq \tau] \ , \quad (3)$$

where $\delta[\cdot]$ is the delta function whose value is 1 when the condition is met and 0 otherwise.

Total loss. The total training loss is the sum of both labeled and unlabeled losses such that

$$\mathcal{L} = \mathcal{L}_l + \lambda \mathcal{L}_u \ , \quad (4)$$

where λ is a trade-off. A more detailed study of λ can be found in Sec. 5.4.

Stream fusion. Let M_T be the Vision Transformer feature map in a certain layer with the shape (d_T, h_T, w_T) representing depth, height and width, respectively. Let $M_{T,i}$ be the i-th patch feature according to M_T with the shape $(d_T, 1, 1)$. So, $M_{T,i}$ corresponds to a specific area of the original image and we denote the CNN sub-feature map who also corresponds to the same area as $M_{C,i}$ with the shape (d_C, h_C, w_C). Motivated by [19], we exchange information between patch features and its related CNN sub-feature map, described as Eq. (5) and Eq. (6):

$$M_{T,i} \mathrel{+}= \texttt{layernorm}(\texttt{pooling}(\texttt{align}(M_{C,i}))), \quad (5)$$

$$M_{C,i} \mathrel{+}= \texttt{batchnorm}(\texttt{upsample}(\texttt{align}(M_{T,i}))), \quad (6)$$

where the `align` operator refers to mapping features to the same dimensional space, `pooling` refers to downsampling, `upsample` refers to upsampling, `layernorm` refers to layer normalization and `BN` refers to batch normalization. Specifically, a `Conv1x1` layer is used for embedding dimension alignment (the `align` operator). The average pooling and spatial interpolation methods are used for spatial dimension alignment, *i.e.*, the `pooling` operator and the `upsample` operator, respectively.

To summarize, our framework consists of two parts, including carrying out a hard-way distillation manner by a convolutional stream to guide the transformer's learning from unlabeled data, and carrying out feature-level information interaction between the two streams so that the CNN's knowledge can be injected into the transformer and the convolutional stream can also be enhanced with a better global spatial information organization capability.

Inference. During training, we use the pseudo labels derived by the convolutional stream to train both the CNN and the Vision Transformer in a semi-supervised setting. For inference, we simply average combine predictions from the both streams as final scores, which is slightly better than using the transformer stream alone, as will be shown empirically.

5 Experiments

5.1 Experimental setup

Datasets and evaluation metrics. To evaluate the effectiveness of `Semiformer`, we mainly conduct experiments on ImageNet [7], which contains 1,000 classes and 1.3M images. In addition, we provide experimental results on Places205 [40]. Unlike ImageNet that contains generic categories, Places205 is a place-focused dataset, which contains 2.5M images annotated into 205 classes. We use top-1 accuracy as our evaluation metric. Through all experiments, following [23], we mainly select 10% labeled samples and leave the other 90% samples as unlabeled data, unless specified otherwise.

Models. The `Semiformer` framework emphasizes how to complement the characteristics of the CNN and the ViT to achieve improved results. For the convolutional stream, we use a ResNet-like model and a personalized ConvMixer [26], while within transformer stream, we experiment with both a slightly modified ViT-S [25] and the PiT-S [15] as backbone networks.

Implementation details. The initial learning rate is set to 10^{-3} and is decayed towards 10^{-5} following the cosine decay scheduler. We use 5 epochs to warm-up our models and another 25 epochs to train models on the labeled data before starting the semi-supervised learning process. In the training of ViT-ConvMixer model, the batch size of each GPU is 84, while in the training of ViT-Conv and PiT-Conv model, the batch size is 108 per GPU. We train models with 600 epochs using 32 NVIDIA V100 GPUs to produce our best top-1 accuracy by setting the number ratio of labeled and unlabeled images in each batch as 1:7. In order to avoid gains brought by data augmentation, we do not apply mixup, cutmix and repeat augmentation in our SSL process. We choose random augmentation, random erasing and color jitter as the strong data augmentation, and use random flipping and random cropping as the weak data augmentation. The value of λ which is the balance factor between loss terms is set as 4.0. In the semi-supervised learning with 5% ImageNet labeled samples, we reduce the number ratio of labeled and unlabeled images per batch to 1:9. All the experiments share the same G.T. data split.

For ablation studies and discussion, we train 300 epochs to speed up the experiments and we set the number ratio of labeled and unlabeled images in each batch as 1:5 and use the label smoothing trick on ground-truth labels.

5.2 Main Results

Comparisons with state-of-the-art. We first compare with state-of-the-art semi-supervised methods, such as UDA [34], FixMatch [23], S4L [38], MPL [20] and CowMix [10], as well as recent self-supevised methods. Experimental results in Tab. 2 show that our approach achieves better results by clear margins compared with alternative methods. For example, `Semiformer` is better than S4L [38] and CowMix [10] by 2.3% and 1.6% with only 11% and 67% of parameters of their models, respectively. In addition, while we follow the design the

Table 2. The results of `Semiformer` and comparisons with state-of-the-art methods. `Semiformer` achieves 75.5% top-1 accuracy and outperforms all Convolutional neural network based methods, while still keeping a reasonable parameter size. Here, the params does not include the final classifier.

Method	Architecture	Params	Top-1 Acc(%)
`Sup.` (10%)	ViT-S	23M	48.6
	Conv	13M	60.2
Self-supervised pretraining			
CPC [14]	ResNet-161	305M	71.5
SimCLR [5,6]	ResNet-50	24M	65.6
SimCLR [5,6]	ResNet-50 (2×)	94M	71.7
BYOL [11]	ResNet-50	24M	68.8
BYOL [11]	ResNet-50 (2×)	94M	73.5
DINO [3]	ViT-S	21M	72.2
Semi-supervised methods			
UDA [34]	ResNet-50	24M	68.8
FixMatch [23]	ResNet-50	24M	71.5
S4L [38]	ResNet-50 (4×)	375M	73.2
MPL [20]	ResNet-50	24M	73.9
CowMix [10]	ResNet-152	60M	73.9
`Semiformer`	ViT-S + Conv	40M	**75.5**

principle of FixMatch to generate pseudo labels, the knowledge sharing mechanism in `Semiformer` brings about 4% performance gain compared to FixMatch. Although MPL has a smaller model size, training MPL is computationally expensive as it requires meta updates. In addition, MPL uses complicated data augmentations, *i.e.*, AutoAugment, while we only use basic augmentations. Similarly, CowMix [10] introduces a new data augmentation strategy for SSL. We would like to point that `Semiformer` is a generic SSL framework that explores pseudo labels and knowledge in CNNs to promote the results of transformers. We believe it is in tandem with more advanced pseudo label generation strategies like MPL [20] and more complex augmentation methods [10]. In addition to SSL methods, we also compare with self-supervised learning results such as [6,11,14], which firstly learn representations with self-supervised methods and then perform finetuning on limited data. We see that `Semiformer` also performs favorably compared to this line of methods.

Effectiveness of `Semiformer` with different backbones. We evaluate the performance of `Semiformer` instantiated with different CNN and transformer backbones using 10% of labeled samples. We compare with the supervised training baseline (`Sup.`), the `Vanilla` method where the pseudo label generator share the same backbone used for SSL, and `Conv-labeled` that trains transformers with labels produced by CNNs. The results are summarized in Tab. 3. As the `Vanilla` results shown in the second block of Tab. 3, CNNs obviously achieve higher image classification accuracy than Vision Transformers under the SSL setting, verifying that labels from CNNs are more accurate. ConvMixer [26]

Table 3. Ablation Study: The effectiveness of `Semiformer` with various backbones and comparisons with alternative methods (*i.e.*, vanilla and conv-labeled). All models are trained with 300 epochs and without pseudo label smoothing. For `Conv-labeled` and `Semiformer`, A/B in the last column: A indicates scores from the transformer stream only and B indicates averaged predictions from CNNs and transformers.

Method	Backbone	Pseudo labels	Top-1 Acc(%)
`Sup.` (10%)	ViT-S	-	48.6
	PiT-S	-	50.0
`Vanilla`	Conv	Conv	68.5
	ConvMixer	ConvMixer	69.3
	ViT-S	ViT-S	59.0
	PiT-S	PiT-S	63.0
`Conv-labeled`	ViT-S + Conv	Conv	67.2 / 70.2
	PiT-S + Conv	Conv	67.8 / 70.5
	ViT-S + ConvMixer	ConvMixer	66.7 / 70.2
`Semiformer`	ViT-S + Conv	Conv	72.4 / 73.5
	PiT-S + Conv	Conv	70.8 / 71.6
	ViT-S + ConvMixer	ConvMixer	72.9 / **73.8**

achieves the best results among all `Vanilla` models, offering a top-1 accuracy of 69.3%. This possibly results from the fact ConvMixer integrates the architectural advantages of both transformers and CNNs. Results in the third block of Tab. 3 show that using CNNs instead of Vision Transformers to generate pseudo labels significantly improves the performance of the Vision Transformer, allowing PiT-S and ViT-S to reach an accuracy of 67.2% and 67.8% respectively, with the same CNN architecture. The improved accuracy is close to that of the `Vanilla` semi-supervised CNN, suggesting the quality of the pseudo labels makes a difference to the semi-supervised learning process of Vision Transformers.

Results in the last block of Tab. 3 show that `Semiformer` significantly improves the performance of Vision Transformers. This highlights the effectiveness of `Semiformer` in exploring the interactions of CNNs and transformers. Taking the combination of ViT-S and Conv as an example, after applying the feature-level interaction to accomplish the dual information exchange, the accuracy of ViT-S is improved by 5.2% from 67.2% to 72.4%, revealing the efficacy of our `Semiformer` framework. We also observe that `Semiformer` is a versatile framework compatible with modern CNN and transformer architectures. In addition, by further combining the predictions from both the convolutional and transformer streams, we observe consistent performance gains under all settings for `Conv-labeled` and `Semiformer`.

The ratio of labeled samples. we further experiment with 5% and 20% of labeled samples for SSL and compare with alternative methods. Except that we decrease the number of labeled and unlabeled images in each batch from 1:5 to 1:9 for 5% labeled samples, all the experimental settings are kept the

Table 4. SSL with different ratios of labeled samples on ImageNet.

Dataset	Ratio	ViT-S		Conv		ViT-S + Conv	
		Sup.	Vanilla	Sup.	Vanilla	Conv-labeled	Semiformer
ImageNet	5 %	28.6	45.7 (↑**17.1**)	44.2	61.3 (↑**17.1**)	62.0	66.3 (↑**4.3**)
	10 %	48.6	59.0 (↑**10.4**)	60.2	68.5 (↑ **8.3**)	70.2	73.5 (↑**3.3**)
	20 %	52.9	69.8 (↑**16.9**)	63.5	73.6 (↑**10.1**)	74.8	78.1 (↑**3.3**)

Table 5. Top-1 Accuracy of `Semiformer` on 5% labeled subset of Places205.

Dataset	Ratio	ViT-S		Conv		ViT-S + Conv	
		Sup.	Vanilla	Sup.	Vanilla	Conv-labeled	Semiformer
Places205	5 %	36.0	46.9 (↑**3.9**)	44.3	51.6 (↑**7.3**)	52.5	53.8 (↑**1.3**)

same as those of using 10% labeled samples. Tab. 4 presents the results. Vision transformer performs poorly when only 5% labels are available, with an accuracy of only 28.6%, which is 15.6% lower than the Conv accuracy of 44.2%. With the increase of the number of labeled samples, the performance gain of ViTs is more significant than that of CNNs. For example, with ViTs, the accuracy increases by 20% and 14.3% respectively, when the number labeled samples grows from 5% to 10% and from 10% to 20%, respectively, suggesting that the training of Vision Transformers is more sensitive to the number of labels. In addition, we see that pseudo labels from CNNs are more accurate and help ViT learn better.

Extension to Places205. We also conduct experiments on Places205 to further evaluate the effectiveness of `Semiformer`. As Places205 is roughly 2 times larger than ImageNet, we use 5% of labeled samples to assure the semi-supervised experiments on 5% Places205 and 10% ImageNet have approximately the same number of labeled samples. We see from Tab. 5 that `Semiformer` consistently produces the best results. For example, `Semiformer` is 1.3% and 6.9% better than `Conv-labeled` and `Vanilla`-ViT-S, respectively. Similar trends can be observed by comparing across Tab. 4 and Tab. 5, which further confirms the efficacy of `Semiformer`.

5.3 Qualitative Results

We visualize in Fig. 4 the attention maps of ViT and `Semiformer`. Thanks to the guidance of pseudo label generator CNN and its supplementary help of injecting the local information extraction ability, `Semiformer` can retain more local information of images and can correctly focus on the key local positions of the images. For example, when analyzing the Fig. 4(a) which corresponds to the class of *bow*, `Semiformer` is particularly more focused on the man's hand holding the bow, the man's head and the quiver carried by the person, and those attended areas are critical for identifying the bow category. In addition, `Semiformer` covers essential objects precisely. In Fig. 4(f), we can see the attention map of `Semiformer` not only covers the animal completely, but also covers

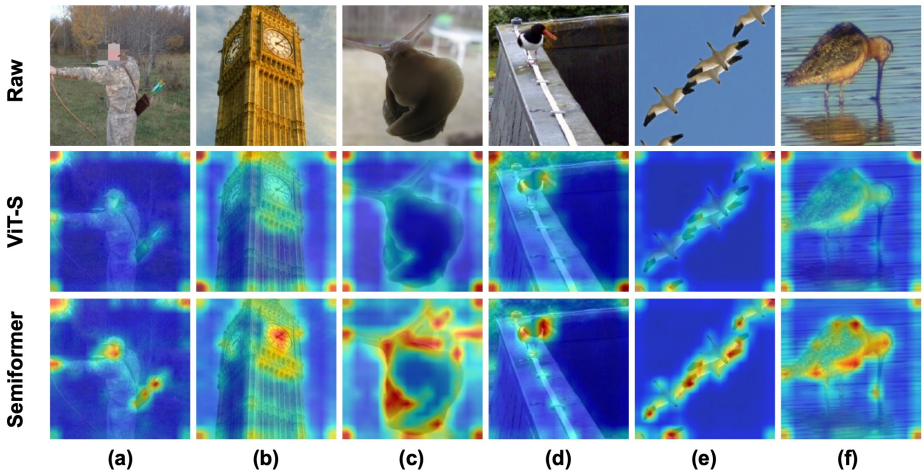

Fig. 4. Attention map of ViTs and `Semiformer` using samples from ImageNet. Compared to ViTs where the attention scores are scattered, `Semiformer` focuses more on critical objects.

the contours more tightly. And for images with many small objects, for instance, Fig. 4(e), `Semiformer` shows stronger ability to concentrate on key local areas and coverage the essential areas.

5.4 Discussion

What model should be used to produce pseudo labels? Although models in our `Semiformer` framework interact with each other, the CNN model still outperforms the vision transformer especially in the early training stage, making it important to retain the CNN hard-way distillation mode. To verify this, we replace the teacher stream which is responsible for generating pseudo labels. We use the following three strategies to produce pseudo labels: CNNs only, transformers only, and averaged predictions from CNNs and transformers. As shown in Tab. 6, using the CNNs as the teacher network brings the highest accuracy, *i.e.* 73.5%, while using the transformer stream to generate pseudo labels performs worst (*i.e.*, 67.4%) . As the quality of pseudo labels generated by vision transformers are limited, we do not get better results by simply averaging CNN and Vision Transformer outputs as pseudo labels under the same setting. This further confirms the effectiveness of our pseudo labeling strategy.

Does `Semiformer` performs well because of larger models? To clear up the confusion on the relationship between the number of parameters and accuracy, we ablate on the model architecture of `Semiformer` using different backbones. We experiment with different versions of ResNet [13] including ResNet-50 (R50), ResNet-101 (R101), ResNet-152 (R152). Results are presented in Tab. 7.

Table 6. Results by different pseudo labels.

PL Type	Acc@1(%)
CNN	**73.5**
Trans	67.4
Fusion	71.1

Table 7. Model size analysis. V and C refer to ViT-S and CNN, respectively. R represnets ResNet.

Architecture	R50	R101	R152	C	V	C+C	V+V	V+C
Params	24M	43M	58M	13M	23M	35M	47M	40M
Top-1 Acc(%)	68.3	70.8	71.8	68.5	59.0	66.9	59.6	**73.5**

We observe that by adding more layers to ResNet, the top-1 accuracy of semi-supervised learning does gradually increase. However, it is still lower than that of `Semiformer`. Even though the ResNet152 model contains 18M more parameters than `Semiformer`, its accuracy is still 1.7% worse than that of `Semiformer`, which proves the performance gain of `Semiformer` does not come from model sizes. We further instantiate the two streams of `Semiformer` with the same backbone, *i.e.* C+C and V+V respectively, and modify the stream connection correspondingly. Note that this is different from `Vanilla` as the two streams exchange information. Tab. 7 reveals that these combinations are significantly worse than `Semiformer`. For example, `Semiformer` outperforms V+V by 13.9% with 7M fewer parameters, which again shows the effectiveness of `Semiformer` is not due to extra parameters.

The impact of hyper-parameters. The default set of hyperparameters are: label and unlabeled data ratio is 1:5, confidence threshold is 0.7 and λ is set as 4. Based on the default setting, we control other variables unchanged and observe how the accuracy rate changes after independently changing the following three factors: different confidence threshold (0.65, 0.7, 0.75, 0.8); different λ value (1, 2, 3, 4); different proportion of the number of labeled and unlabeled data (1:3, 1:5, 1:7). `Semiformer` offers the best results with 0.7 confidence threshold, 1:7 labeled-unlabeled ratio, and $\lambda = 4$.

6 Conclusion

We presented `Semiformer`, the first framework to train Vision Transformers for semi-supervised learning. We found directly training a `Vanilla` transformer on semi-supervised data is ineffective. The proposed framework combines a CNN and a Vision Transformer using a cross fusion approach. The optimal semi-supervised learning performance is achieved by using only the convolutional stream to generate the pseudo labels. The final fused framework achieves 75.5% top-1 accuracy on ImageNet and outperforms the state-of-the-art in semi-supervised image classification.

Acknowledgement Y.-G. Jiang was sponsored in part by "Shuguang Program" supported by Shanghai Education Development Foundation and Shanghai Municipal Education Commission (No. 20SG01). Z. Wu was supported by NSFC under Grant No. 62102092.

References

1. Bachman, P., Alsharif, O., Precup, D.: Learning with pseudo-ensembles. In: NeurIPS (2014) 4
2. Berthelot, D., Carlini, N., Cubuk, E.D., Kurakin, A., Sohn, K., Zhang, H., Raffel, C.: Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. arXiv preprint arXiv:1911.09785 (2019) 4
3. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: ICCV (2021) 4, 10
4. Chapelle, O., Scholkopf, B., Zien, A.: Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. TNN (2009) 1
5. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML (2020) 4, 10
6. Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.E.: Big self-supervised models are strong semi-supervised learners. In: NeurIPS (2020) 4, 10
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009) 9
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) 1, 3, 4
9. d'Ascoli, S., Touvron, H., Leavitt, M.L., Morcos, A.S., Biroli, G., Sagun, L.: Convit: Improving vision transformers with soft convolutional inductive biases. In: ICML (2021) 4
10. French, G., Oliver, A., Salimans, T.: Milking cowmask for semi-supervised image classification. arXiv preprint arXiv:2003.12022 (2020) 9, 10
11. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. In: NeurIPS (2020) 10
12. Gulati, A., Qin, J., Chiu, C.C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., et al.: Conformer: Convolution-augmented transformer for speech recognition. arXiv preprint arXiv:2005.08100 (2020) 4
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016) 3, 13
14. Henaff, O.: Data-efficient image recognition with contrastive predictive coding. In: ICML (2020) 10
15. Heo, B., Yun, S., Han, D., Chun, S., Choe, J., Oh, S.J.: Rethinking spatial dimensions of vision transformers. In: ICCV (2021) 1, 3, 4, 9
16. Lee, D.H., et al.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: ICMLW (2013) 4
17. Li, Y., Yao, T., Pan, Y., Mei, T.: Contextual transformer networks for visual recognition. IEEE TPAMI (2022) 3
18. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV (2021) 1, 3
19. Peng, Z., Huang, W., Gu, S., Xie, L., Wang, Y., Jiao, J., Ye, Q.: Conformer: Local features coupling global representations for visual recognition. In: ICCV (2021) 4, 8
20. Pham, H., Dai, Z., Xie, Q., Le, Q.V.: Meta pseudo labels. In: CVPR (2021) 9, 10

21. Rasmus, A., Berglund, M., Honkala, M., Valpola, H., Raiko, T.: Semi-supervised learning with ladder networks. In: NeurIPS (2015) 4
22. Rosenberg, C., Hebert, M., Schneiderman, H.: Semi-supervised self-training of object detection models (2005) 4
23. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.L.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In: NeurIPS (2020) 1, 4, 5, 7, 9, 10
24. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: NeurIPS (2017) 4
25. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: ICML (2021) 3, 9
26. Trockman, A., Kolter, J.Z.: Patches are all you need? arXiv preprint arXiv:2201.09792 (2022) 9, 10
27. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017) 3
28. Wang, J., Yang, X., Li, H., Wu, Z., Jiang, Y.G.: Efficient video transformers with spatial-temporal token selection. In: ECCV (2022) 3
29. Wang, R., Chen, D., Wu, Z., Chen, Y., Dai, X., Liu, M., Jiang, Y.G., Zhou, L., Yuan, L.: Bevt: Bert pretraining of video transformers. In: CVPR (2022) 3
30. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: ICCV (2021) 3, 4
31. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: CVPR (2018) 4
32. Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., Zhang, L.: Cvt: Introducing convolutions to vision transformers. In: ICCV (2021) 4
33. Xiao, T., Dollar, P., Singh, M., Mintun, E., Darrell, T., Girshick, R.: Early convolutions help transformers see better. In: Advances in Neural Information Processing Systems (2021) 4
34. Xie, Q., Dai, Z., Hovy, E., Luong, T., Le, Q.: Unsupervised data augmentation for consistency training. In: NeurIPS (2020) 4, 9, 10
35. Xie, Q., Luong, M.T., Hovy, E., Le, Q.V.: Self-training with noisy student improves imagenet classification. In: CVPR (2020) 4
36. Yang, L., Wang, Y., Gao, M., Shrivastava, A., Weinberger, K.Q., Chao, W.L., Lim, S.N.: Deep co-training with task decomposition for semi-supervised domain adaptation. In: ICCV (2021) 4
37. Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z.H., Tay, F.E., Feng, J., Yan, S.: Tokens-to-token vit: Training vision transformers from scratch on imagenet. In: ICCV (2021) 3, 4
38. Zhai, X., Oliver, A., Kolesnikov, A., Beyer, L.: S4l: Self-supervised semi-supervised learning. In: ICCV (2019) 9, 10
39. Zhang, Y., Qin, J., Park, D.S., Han, W., Chiu, C.C., Pang, R., Le, Q.V., Wu, Y.: Pushing the limits of semi-supervised learning for automatic speech recognition. arXiv preprint arXiv:2010.10504 (2020) 4
40. Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., Oliva, A.: Learning deep features for scene recognition using places database. Advances in neural information processing systems **27** (2014) 9
41. Zhu, X.J.: Semi-supervised learning literature survey. Tech. rep., University of Wisconsin-Madison Department of Computer Sciences (2005) 1