Semi-Supervised Vision Transformers Supplementary Material

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

1 Model architecture

Table 1. Semiformer architecture with ViT-S and ConvMixer as the transformer branch and convolution branch, respectively.

	ViT-S				ConvMixer		
stages	**times**	**output size**	**operations**		**operations**	**output size**	**times**
stem	1×	112×112, 64 56×56, 64 14×14, 384	conv 7×7, 64, stride 2 3×3 max pooling, stride 2 conv 4×4, 384, stride 4		conv 8×8, 768, stride 8	28×28, 768	×1
Head	1×	14×14, 384	$\begin{bmatrix}\text{MHSA}-6,384\\1\times1,1536\\1\times1,384\end{bmatrix}$		dw_conv 9×9, 768, stride 1	28×28, 768	×1
Body	11×	14×14, 384	- $\begin{bmatrix}\text{MHSA}-6,384\\1\times1,1536\\1\times1,384\end{bmatrix}$	⇐ ⇒	N× $\begin{bmatrix}\text{d\_conv } 9\times9,\ 768, \text{s } 1\end{bmatrix}$ ⇓ 1× $\begin{bmatrix}-\\-\\\text{d\_conv } 9\times9,\ 768, \text{s } 1\end{bmatrix}$	28×28, 768	×11
Tail	1×	1×1000	class token 1×1, 1000		global pooling 1×1, 1000	1×1000	×1

To better understand the design pattern of the model, as shown in Tab. 1, we show the combination details of ViT-S and ConvMixer, which is one of the variant extensions of `Semiformer`.

2 Hyper-parameters study

Fig. 1 shows the results of tuning hyper-parameters in detail. We form `Semiformer` learning framework with ViT-S+Conv and do hyper-parameters sensitivity testing. Default hyper-parameters are set as: 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). Plots in Fig. 1 show that `Semiformer` offers the best results with 0.7 confidence threshold, 1:7 labeled-unlabeled ratio, and $\lambda = 4$. Also it can be seen that in the process of adjusting hyper-parameters, `Semiformer`, whose accuracy scores won't be greatly affected, maintains competitive.

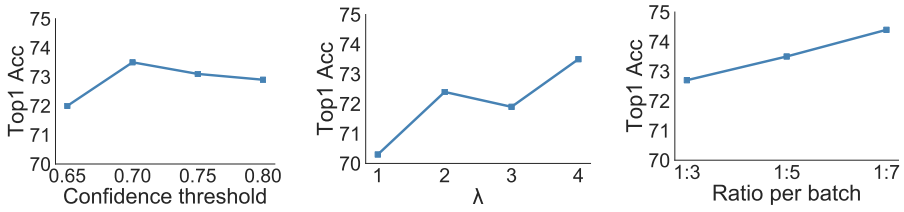

Fig. 1. Study of hyperparameters: (a) Varying the confidence threshold for pseudo-labels. (b) Varying the loss terms balance factor λ. (c) Varying the ratio of labeled and unlabeled images.

3 Training with fewer epochs.

For better verifing the competitive performance of `Semiformer`, we train models with fewer epochs and do comparisons among `Semiformer` and other baselines. As shown in Tab. 2, `Semiformer` consistently outperforms other baseline methods under various training epoch settings.

Table 2. Ablation Study. Comparisons among `Semiformer` and alternative methods (*i.e.*, vanilla and conv-labeled) with various epoch setting.

-	100E	150E	200E	300E
`Vanilla`	51.1	55.6	58.1	59.0
`Conv-labeled`	57.7	63.7	66.7	70.2
`Semiformer`	63.2	68.8	70.7	73.5

4 Fusion module overview.

We visualize **stream fusion** in Fig. 2 to make it easier for the reader to understand. In the fusion process, every two convolutional layers are paired with a

transformer layer, and the size of the CNN feature maps are always larger than or equal to that of the ViT patch maps. Therefore, for the final information interaction, we align each CNN grid sub-feature map with each ViT patch feature by down-sampling operation, and conversely align the ViT patch feature to the corresponding CNN grid sub-feature map by up-sampling operation. Unlike [1], our fusion is completely decoupled from the convolutional block, making it more flexible and able to be extended to different types of CNN networks such as the aforementioned ConvMixer.

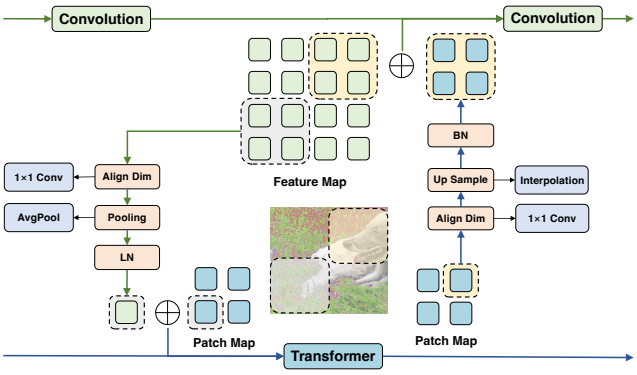

Fig. 2. Fusion diagram.

Since our stream fusion design is inspired by [1], we further replace `Semiformer` with Conformer and use it for SSL (all models are trained with 300 epochs). Results below demonstrate that `Semiformer` performs better in different scenarios, highlighting that not only the fusion structure but also the pseudo-labeling strategy are effective.

Table 3. Comparison between `Semiformer` the Conformer baseline. `Semiformer` achieves better performance than Conformer at different labeling ratios.

Dataset	Ratio	Conformer	`Semiformer`
ImageNet	5%	59.4	66.3 (↑**7.1**)
	10%	72.2	73.5 (↑**1.3**)
	20%	76.1	78.1 (↑**2.0**)

5 Extra experiment details.

The GPU device we use is "NVIDIA Tesla V100 32GB". We define "one epoch" as loading all labeled data once. When the number of epochs is fixed, we adjust

the ratio of (labeled: unlabeled) to make sure unlabeled data are fully seen during training. Thus, we use 1:9 ratio for the 5%-ImageNet scenario as there are more unlabeled images now compared to the 10%-ImageNet setting.

Our training process follows the end-to-end teacher-student frameworks. In particular, we use a small number of early epochs to warm up the model using labeled data only. This is generally referred to as the burn-in stage, a popular training strategy in SSL for better initialization so that more accurate pseudo labels can be obtained. After that, both labeled and unlabeled data are jointly used for model training.

References

1. 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) 3