A Good Student is Cooperative and Reliable: CNN-Transformer Collaborative Learning for Semantic Segmentation -Supplementary-

Jinjing Zhu[1] Yunhao Luo[3] Xu Zheng[1] Hao Wang [4] Lin Wang[1,2*]

[1] AI Thrust, HKUST(GZ) [2] Dept. of CSE, HKUST [3] Brown University [4] Alibaba Cloud, Alibaba Group

zhujinjing.hkust@gmail.com, devinluo27@gmail.com, zhengxu128@gmail.com, cashenry@126.com, linwang@ust.hk

Project Page: https://vlislab22.github.io/CTCL/

Abstract

Due to the lack of space in the main paper, we provide more details of the proposed method and experimental results in the supplementary material. Sec.1 introduces the details of the proposed method. Sec.2 provides the details of the encoders used in this work. Lastly, Sec.3 provides pseudo algorithm of the proposed method. Sec.4 shows some discussions about our proposed method.

1. Details of the Proposed Method

Tab. 1 shows the architecture of MobileNetV2, ResNet-50, MiT-B1, and MiT-B2, respectively. We take the collaborative learning between MobileNetV2 and MiT-B1 as an example and present the details of our proposed method.

1.1. Heterogeneous Feature Distillation (HFD)

The first-layer feature F_1^V size of MobileNetV2 is 24×128×128 and the first-stage feature F_1^V size of MiT-B1 is 64×128×128. To match the sizes of features, we utilize the linear transformations Γ_1^C and Γ_1^V to reshape the sizes of F_1^C and F_1^V as 64×128×128 and 24×128×128, respectively. Then, we can use the transformed feature to calculate the HFD loss as follow:

$$\begin{aligned} \mathcal{L}_{\text{HFD}}^C &= cos(\text{Attn}((F_1^{\hat{C}})), F_2^V), \\ \mathcal{L}_{\text{HFD}}^V &= cos(\text{MLP}(F_1^{\hat{V}}), F_2^C), \end{aligned} \tag{1}$$

where $F_1^{\hat{C}}$ and $F_1^{\hat{V}}$ is the transformed feature, the shapes of which are 64×128×128 and 24×128×128, respectively.

1.2. Region-wise Bidirectional Selective Distillation

The last-layer feature F_l^C size of MobileNetV2 is 96×64×64 and the last-stage feature F_l^V is 512× 16×16. To match the sizes of features, we exploit the linear transformations Γ_l^C and Γ_l^V to reshape the sizes of F_1^C and F_1^V as 96×16×16 and 96×16×16, respectively. The transformed features are donated as $F_l^{\hat{C}}$ and $F_l^{\hat{V}}$, separately. It is worth noting that the shapes of the predictions are 512×512. To match the sizes of transformed features $F_l^{\hat{C}}$ (or $F_l^{\hat{V}}$)and predictions P^C (or P^V), we divide the prediction map into 16×16 size. A $\frac{512}{16} \times \frac{512}{16}$ sized prediction map P^C (or P^V) at the same location corresponds to one region in $F_l^{\hat{C}}$ (or $F_l^{\hat{V}}$). Then we use the sum of cross entropy loss of $\frac{512}{16} \times \frac{512}{16}$ sized prediction map to decide the transferred direction between two students' regions with the same location. Finally, the region-wise BSD loss is defined as

$$\begin{aligned} \mathcal{L}_R^C &= \frac{1}{16 \times 16 - \hat{M}} \sum_{\hat{h}=1}^{16} \sum_{\hat{w}=1}^{16} (1 - \hat{m}_{(\hat{h},\hat{w})}) S_{(\hat{h},\hat{w})}, \\ \mathcal{L}_R^V &= \frac{1}{\hat{M}} \sum_{\hat{h}=1}^{16} \sum_{\hat{w}=1}^{16} m_{(\hat{h},\hat{w})} S_{(\hat{h},\hat{w})}, \end{aligned} \tag{2}$$

where $\hat{M}$ decides the direction of KD for each region and calculate the cross-student region-wise similarity matrix $S_{(\hat{h},\hat{w})}$ is the similarity matrix (as introduced in main paper).

2. Parameters of Encoder

Tab. 2 shows the parameters of encoder for different methods. For CNN-based students, we adopt the famous segmentation architecture DeepLabV3+ with encoders of MobileNetV2 and ResNet-50; for ViT-based students, we utilize the lightweight SegFormer with encoders of MiT-B1 and MiT-B2, which have comparable or smaller parameters with their CNN counterparts, respectively.

*Corresponding author.

3. Algorithm

The pseudo algorithm of the proposed method is shown in Algorithm. 1.

Algorithm 1 The Proposed framework

Input: $\{X, Y\}$; max iterations: T
model: $f(X, \theta^C)$, $f(X, \theta^V)$;
Initialization: Set θ^C and θ^V;
for for t $\leftarrow$ 1 to T **do**
  Attain the segmentation prediction maps and feature representations for each student, respectively: $\left(P^{\mathrm{C}}, F^{\mathrm{C}}\right) = f\left(X; \theta^{\mathrm{C}}\right), \quad \left(P^{\mathrm{V}}, F^{\mathrm{V}}\right) = f\left(X; \theta^{\mathrm{V}}\right)$;
  Compute the pixel-wise segmentation loss for each student:
  $\mathcal{L}_{CE}^{C} = \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} CE\left(\sigma\left({P^{C}}_{(h,w)}\right), y_{(h,w)}\right)$,
  $\mathcal{L}_{CE}^{V} = \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} CE\left(\sigma\left({P^{V}}_{(h,w)}\right), y_{(h,w)}\right)$;
  Compute the HFD loss for each student:
  $\mathcal{L}_{\text{HFD}}^{C} = cos(Attn((F_1^{\hat{C}})), F_2^{V})$,
  $\mathcal{L}_{\text{HFD}}^{V} = cos(MLP(F_1^{\hat{V}}), F_2^{C})$;
  Compute the region-wise BSD loss for each student:
  $\mathcal{L}_{R}^{C} = \frac{1}{\hat{H} \times \hat{W} - \hat{M}} \sum_{\hat{h}=1}^{\hat{H}} \sum_{\hat{w}=1}^{\hat{W}} (1 - \hat{m}_{(\hat{h},\hat{w})}) S_{(\hat{h},\hat{w})}$,
  $\mathcal{L}_{R}^{V} = \frac{1}{\hat{M}} \sum_{\hat{h}=1}^{\hat{H}} \sum_{\hat{w}=1}^{\hat{W}} m_{(\hat{h},\hat{w})} S_{(\hat{h},\hat{w})}$;
  Compute the pixel-wise BSD loss for each student:
  $\mathcal{L}_{\text{BSD}}^{C} = \mathcal{L}_{R}^{C} + \alpha \mathcal{L}_{P}^{C}$,
  $\mathcal{L}_{\text{BSD}}^{V} = \mathcal{L}_{R}^{V} + \alpha \mathcal{L}_{P}^{V}$;
  Compute the total objective for each student:
  $\mathcal{L}^{C} = \mathcal{L}_{CE}^{C} + \beta \mathcal{L}_{\text{HFD}}^{C} + \gamma \mathcal{L}_{\text{BSD}}^{C}$,
  $\mathcal{L}^{V} = \mathcal{L}_{CE}^{C} + \beta \mathcal{L}_{\text{HFD}}^{V} + \gamma \mathcal{L}_{\text{BSD}}^{V}$.
  Back propagation for $\mathcal{L}^{C}$ and $\mathcal{L}^{V}$;
  Update the students θ^C and θ^V with $\mathcal{L}^{C}$ and $\mathcal{L}^{V}$, respectively.
end for
return θ^C and θ^V.
End.

4. Discussion

4.1. Intuition of BSD

The design of BSD is one of the critical contributions of this paper as it facilitates the two students to collaboratively learn reliable knowledge from each other and the knowledge is transferred bidirectionally. Due to the different performance at different regions between the ViT and CNN students, we intend to dynamically select reliable knowledge between the two students in the feature space, so as to benefit each other. However, there is a challenging problem: 'how to decide the directions of transferring knowledge fro different regions during training?' To this end, we propose to manage the directions of KD via combining the predictions and GT labels, where we regard the directions of KD for different regions as a sequential decision making problem. Consequently, we propose a directional selective distillation (BSD) for enabling students to learn collaboratively. As the principle of collaborative learning requires bidirectional knowledge transfer, BSD should be "bidirectional" to enable CNNs to learn from ViT while ViT learns from CNNs. Our key idea is "selective" due to the considerable model size gap and learning capacity gap between CNNs and ViT. The reasons causing the gaps are 1): The discrepancies in features and predictions between CNNs and ViT caused by the distinct computing paradigms make it challenging to do online KD. 2): These methods only transfer the knowledge in logit space; however, there is more reliable and efficient knowledge in the features extracted by both models. 3) There are considerable model size gap and learning capacity gap between CNNs and ViT.

4.2. Intuition of HFD

We make students learn the heterogeneous features from each other in the first-layer feature space and align these features in the second layer. That is, we input the transformed features into the second layer and then align the outputs instead of directly aligning features of the first layer. This way, it can make both students learn the global and local features in the first-layer space.

4.3. Selection of Layers

We use the first-layer features as low-layer features of CNNs and ViT are less distinct and heterogeneous, making CNNs and ViT learn from each other more effectively. Moreover, due to the different computing paradigms and learning capacities of CNNs and ViT, aligning high-layer features is less approachable and practical. Lastly, aligning multiple low-layer features lead to an increase in the computation cost. Tab. 3 in the paper shows the effectiveness of our proposed method between heterogeneous students with different performance abilities.

4.4. About MLP or Attn in HFD Module

MLP consisting of convolutional layers extracts the local semantic features, and Attn consisting of a self-attention module extracts the global semantic features. Therefore, after inputting the local features into Attn or inputting the global features into MLP, these output features are comparable. As such, we use cosine similarity to measure the similarity of these features and enable students to learn from each other in the low-layer space.

4.5. About Operations in Eq.2

Attn updates the first-layer features of CNNs, while MLP updates the first-layer features of ViT. However, if we apply Attn operation in 'CNNs to ViT" and MLP in "ViT to CNNs', Attn operation can optimize the first two layers of

Layer of MobileNetV2	First-layer F_1^C	Second-layer F_2^C	Third-layer	Last-layer F_l^C
Output Size	24×128×128	32×64×64	64×64×64	96×64×64
Layer of ResNet-50	First-layer F_1^C	Second-layer F_2^C	Third-layer	Last-layer F_l^C
Output Size	256×128×128	512×64×64	1024×32×32	2048×32×32
Stage of MiT-B1	First-stage F_1^V	Second-stage F_2^V	Third-stage	Last-stage F_l^V
Output Size	64×128×128	128×64×64	320×32×32	512× 16×16
Stage of MiT-B2	First-stage F_1^V	Second-stage F_2^V	Third-stage	Last-stage F_l^V
Output Size	64×128×128	128×64×64	320×32×32	512× 16×16

Table 1: Output size of each layer (stage) of different encoders.

$Method$	Encoder	Parameters(M)
DeepLabV3+	MobileNetV2	15.4
SegFormer	MiT-B1	13.7
DeepLabV3+	ResNet-50	43.7
SegFormer	MiT-B2	27.5

Table 2: The Parameters of methods with different encoder.

ViT while MLP operation can optimize the first two layers of CNNs. Both approaches can facilitate collaborative learning between CNNs and ViT but optimizing the first two layers increases computation cost.

4.6. About ViT-ViT setting

ViT is not absolutely better while CNN still matters; therefore, we explore to take full advantage of CNN and ViT while compensating for their limitations. Moreover, in Tab.4, our method demonstrates superior performance compared to previous studies in ViT-ViT setting.

4.7. Results on ADE-20K:

The effectiveness of our method is further demonstrated by the results obtained on the more challenging ADE-20K dataset, as shown in Tab. 4. The results will be included in the final version.

4.8. Distillation on hybrid network:

We explore the potential of our framework between the CNN-based (ViT-based) student and hybrid network-based student MaxViT [1], to further demonstrate its effectiveness in Tab. 5. The significant improvements **+7.59%** and **+5.45%** underscore the effectiveness and practicality of employing our proposed methodology within hybrid network architectures.

4.9. About the motivation

As ViT is notoriously impeded by limitations, such as the lack of certain inductive biases and poor performance on small-scale datasets; while CNN excels at capturing local features although CNN may underperform ViT on large-scale datasets. Therefore, ViT is not absolutely better while CNN still matters, and it is promising to take full advantage of CNN and ViT while compensating for their limitations. From this new perspective, prior arts [1,2] adopting the CNN for an auxiliary purpose, are less optimal and intuitive. So, our motivation is reasonable and novel. Our key idea is to simultaneously learn compact yet effective CNN-based and ViT-based models by selecting and exchanging reliable knowledge between them for semantic segmentation. Although 'ViT is shown to have higher upper bounds than CNN', we observe in Figs. 1(b) and 4 that ViTs may exhibit less accurate segmentation results in certain regions compared to CNNs within the same image. To address this, we introduce BSD to compensate for students' weaknesses in region-wise and pixel-wise levels. We further demonstrate the effectiveness of our proposed method in collaborative learning between CNN-based (or ViT-based) and hybrid network-based students by conducting experiments as shown in Tab. 5.

4.10. About reliable knowledge in BSD

Here, 'reliable' does not indicate 'regions', but *indicates better predictions with relatively higher segmentation accuracy* (See Fig. 1). Predictions in region R_1^V (R_2^C) of ViT (CNN) is more reliable compared with predictions in region R_1^C (R_2^V) of CNN (ViT). Then we utilize BSD to enable R_1^C (R_2^V) to learn from R_1^V (R_2^C). Finally, we obtain more accurate region predictions $\hat{R}_1^C$ ($\hat{R}_2^V$). *BSD enables students to learn collaboratively and guarantees the correctness and consistency of soft label*. Qualitative results are in Tabs. 4, 5, 7, and 9 (in main paper), and visualized results in Fig. 4 specifically highlight the effectiveness of BSD.

References

[1] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. In *European conference on computer vision*, pages 459–479. Springer, 2022.

Method	MobileNetV2	MiT-B2	Δ	ResNet-50	MiT-B1	Δ
Vanilla	67.54	82.03	0.00	76.05	78.48	0.00
Ours	**69.21**$_{+1.67}$	**82.27**$_{+0.24}$	**+1.91**	**77.59**$_{+1.54}$	**79.56**$_{+1.08}$	**+2.62**

Table 3: Comparison with the Vanilla methods on the **PASCAL VOC 2012** dataset for our CNN-based (MobileNetV2 and ResNet-50) and ViT-based (MiT-B1 and MiT-B2) students.

Method		MiT-B1	MiT-B2	Δ		MobileNet	MiT-B1	Δ
Vanilla	**CamVid**	76.26	77.76	0.00	**ADE-20K**	22.53	40.07	0.00
DML		75.84	77.40	-0.78		22.02	40.12	-0.46
KDCL		76.61	77.55	+0.14		22.16	41.62	+1.18
IFVD		76.43	77.45	-0.14		21.42	40.64	-0.54
Ours		**77.89**	**78.01**	**+1.88**		**26.47**	**42.28**	**+6.15**

Table 4: Comparison on the **CamVid** for MiT-B2 and MiT-B2 students, and **ADE-20K** for MobileNetV2 and MiT-B1 students.

Method	ResNet-50	MaxViT	Δ	MiT-B2	MaxViT	Δ
Vanilla	58.12	61.89	0.00	77.76	61.89	0.00
DML	59.07	63.80	+2.86	77.09	60.61	-1.95
KDCL	58.64	61.61	+0.24	77.49	63.26	+1.10
IFVD	59.69	62.01	+1.69	77.08	63.10	+0.53
Ours	**62.13**	**65.47**	**+7.59**	**77.96**	**67.14**	**+5.45**

Table 5: Comparison on the **CamVid** for ResNet-50 (MiT-B2) and MaxViT students.

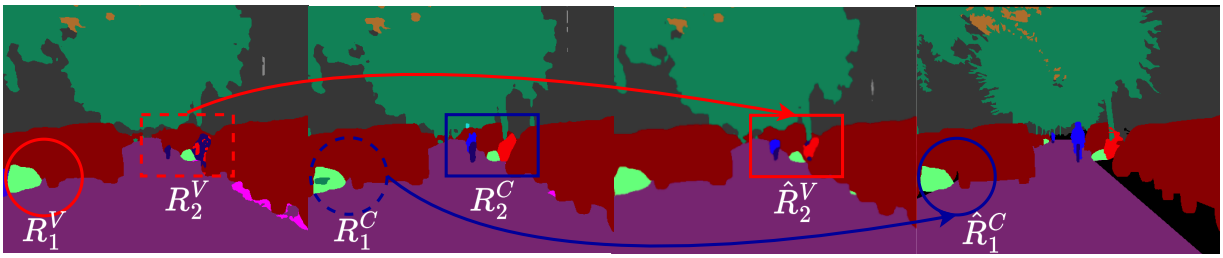

Figure 1: CNN and ViT learns collaboratively by exchanging reliable knowledge.