Pattern Recognition Letters 164 (2022) 239–245

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

TRL: Transformer based refinement learning for hybrid-supervised semantic segmentation

Lin Cheng[a], Pengfei Fang[b], Yan Yan[a], Yang Lu[a], Hanzi Wang[a,*]

[a] Fujian Key Laboratory of Sensing and Computing for Smart City, School of Informatics, Xiamen University, Xiamen 361005, China
[b] College of Engineering and Computer Science, Australian National University, Canberra, ACT 2601, Australia

ARTICLE INFO

Article history:
Received 20 August 2022
Revised 19 October 2022
Accepted 12 November 2022
Available online 15 November 2022

Edited by Jiwen Lu

Keywords:
Hybrid-supervised semantic segmentation
Simi-supervised semantic segmentation
Weakly-supervised semantic segmentation
Refinement learning
Heat map
Pseudo label

ABSTRACT

This paper studies a new yet practical setting of semi-supervised semantic segmentation, i.e., hybrid-supervised semantic segmentation, where a small number of pixel-level (strong) annotations and a large number of image-level (weak) annotations are provided. It is a common practice to utilize pseudo labels to mitigate the issue of lacking strong annotations. However, most of the existing works focus on improving the model representation with unlabeled data, while ignoring the quality of pseudo labels, leading to poor segmentation performance. It is difficult to directly learn a model with limited images to produce high-quality pseudo labels. To address this problem, we propose a novel learning method, i.e., Transformer based Refinement Learning (TRL), which explores a learning process under the assistance of weak annotations and the supervision of strong annotations. TRL progressively refines heat maps from the poor qualities to the better ones to obtain satisfactory pseudo labels. Specifically, we propose a Dual-Cross Transformer Network (DCTN) to perform the refinement learning. DCTN extracts the features from both images and heat maps by a dual-stream network. The cross attentions inside DCTN hierarchically fuse the dual-stream features.

The experiments on the PASCAL VOC and COCO datasets show that TRL outperforms the state-of-the-art methods for hybrid-supervised semantic segmentation.

1. Introduction

Semantic segmentation [1–4] is a fundamental task in computer vision and it plays an important role in real-world applications, such as autonomous driving [5], medical diagnosis [6], and digital makeup [7]. Training a segmentation network usually requires a large number of densely annotated images, which are very expensive to collect. For example, annotating an image with 2048 × 1024 pixel labels usually costs more than 1.5 h [8], which is around 270 times slower than annotating an image with class labels [9].

To alleviate the problem of expensive annotations, some weakly-supervised methods [10–12] exploit much cheaper annotations (e.g., class labels) to extract heat maps (e.g., the 2nd column in Fig. 1) to generate pseudo labels for the segmentation training. However, the class labels only provide the limited supervisory information, which restricts the quality of the generated pseudo labels, leading to unsatisfactory segmentation. On the other hand, a set of semi-supervised methods [13–15] explore using a small number of images with pixel-level labels and a large number of unlabeled images to train segmentation models. Nevertheless, insufficient quantity of labeled data also hinder the performance improvement. Therefore, we study a more effective and efficient learning paradigm, i.e., a special semi-supervised semantic segmentation, where a small number of pixel-level annotations and a large number of image-level annotations are provided. We formally define it as hybrid-supervised semantic segmentation. This paradigm allows the model to obtain a higher performance while keeping low annotation costs, and thus it is more practical and feasible in real-world applications [16].

Few methods have been proposed specially for the task of hybrid-supervised semantic segmentation. Recently, Luo and Yang [16] propose a strong-weak dual-branch network (SWDN) to tackle the problem of hybrid supervisions, where they adopt the weakly-supervised method DSRG [17] to produce pseudo labels. However, weakly-supervised methods often rely on some heuristic priors and manually fine-tuned thresholds, to improve the quality of pseudo labels, while the learned pseudo labels are still unsatisfactory compared to human-annotated labels. In the paradigm

* Corresponding author.
E-mail addresses: lincheng@stu.xmu.edu.cn (L. Cheng), Pengfei.Fang@anu.edu.au (P. Fang), yanyan@xmu.edu.cn (Y. Yan), luyang@xmu.edu.cn (Y. Lu), hanzi.wang@xmu.edu.cn (H. Wang).

https://doi.org/10.1016/j.patrec.2022.11.015
0167-8655/

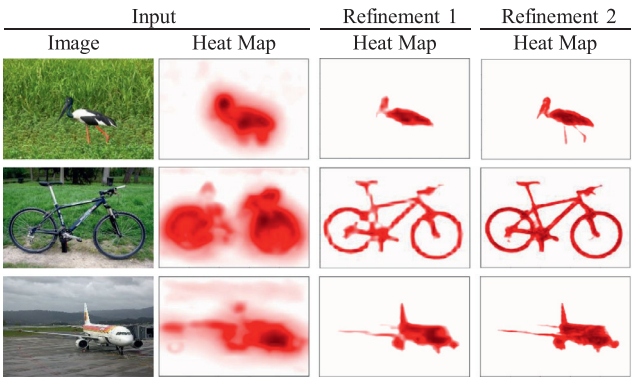

Fig. 1. Examples of the refinement learning of the heat maps. The original heat maps of input are of low quality. The heat maps are gradually improved during the refining process.

of hybrid-supervised learning, most existing methods neglect using strong annotations to learn high-quality pseudo labels to improve the segmentation performance. Intuitively, the higher quality of pseudo labels will result in better segmentation performance. However, it is difficult to directly learn a model from limited images to produce desirable pseudo labels, which is considered a challenge.

To address this problem, we propose a learning method, i.e., Transformer based Refinement Learning (TRL), to progressively refine the heat maps (e.g., Fig. 1) generated from weak annotations to produce desirable pseudo labels, supervised by the strong annotations. TRL takes heat maps and natural images as two-stream inputs and it outputs prediction maps, which can be understood as refined heat maps. Then, the refined heat maps and images are fed into the next round of refinement learning. Specifically, a Dual-Cross Transformer Network (DCTN) is devised to perform the refinement learning. DCTN deeply extracts the features from the two-stream inputs. The designed cross attention layers inside DCTN hierarchically fuse the two-stream features. Ultimately, high-quality pseudo labels are generated from DCTN and they are adopted to train the segmentation model. Our method does not require manual thresholds or CRF [18] as the post-processing step. Extensive experiments on the PASCAL VOC [19] and COCO [20] datasets show the superiority of our method against several state-of-the-art methods for hybrid-supervised semantic segmentation.

The main **contributions** of this paper are three-fold: (1) We propose a novel learning method, i.e., Transformer based Refinement Learning (TRL), which learns a refining process of heat maps to generate high-quality pseudo labels, to address the hybrid-supervised semantic segmentation. (2) We design a Dual-Cross Transformer Network (DCTN) to implement the refinement learning, where the hierarchical cross attentions inside DCTN benefit the interaction of two-stream features. (3) We evaluate the proposed TRL by extensive experiments, which show that TRL achieves the state-of-the-art performance, even surpassing the fully-supervised learning method.

2. Related work

Semi-supervised semantic segmentation. There are two main groups of methods to deal with the semi-supervised semantic segmentation task: consistency learning and self-training.

Consistency learning aims to make use of unlabeled data to help the training model learn better representations. Consistency learning based methods [13,15,21–23] enforce the features or predictions to keep consistency when the input images or intermediate features are perturbed. For example, CCT [21] uses different types of perturbations to adjust the features from the encoders and uses a consistency loss to constrain the outputs of a main decoder and auxiliary decoders, aiming to boost the representation ability of the main decoder. DCC [23] randomly crops two patches with overlapping regions from the original image to train the segmentation model, which makes the learned representations robust to different contexts. However, these methods do not improve the quality of pseudo labels, limiting the segmentation performance.

Several self-training based methods [24–26] are developed to tackle semi-supervised semantic segmentation. Those methods first use the labeled data to train a segmentation model, and then use the trained model to generate pseudo labels on the unlabeled data. Finally, these generated labels are used to retrain the model. The whole progress can be iteratively performed several times. Different from self-training, TRL does not directly use original images to train the model, by which it is very hard to produce high-quality labels. Instead, we propose an effective two-stage refinement strategy to generate high-quality heat maps.

Weakly-supervised semantic segmentation. Weakly-supervised semantic segmentation takes cheaper annotations as supervisions, including image classes, points, scribbles, and bounding boxes. Image class labels, as the most frequently used weak annotations, can be utilized to train a classification model, which is then visualized by CAM [10] to output heat maps. Furthermore, some post-processing techniques [12,18] are appended on heat maps to produce pseudo labels. Due to the low quality of these pseudo labels, several weakly-supervised learning methods, such as DSRG [17], FickleNet [27], IRNet [28] and AdvCAM [29], are developed to improve the quality of the generated pseudo labels. However, these improved labels are not sufficient to train a satisfactory segmentation model.

Hybrid-supervised semantic segmentation. This task adopts pixel-level annotations and other weaker annotations, e.g., image class labels, as supervisions. This paradigm can reach a better balance between the annotation costs and learning performance. Although several semi-supervised learning methods, e.g., CCT [21] and DCC [23], can also deal with the hybrid-supervised semantic segmentation task, they only append an extra loss with pseudo labels to other losses and neglect the relation between weak annotations and strong annotations. SWDN [16] deals with the issue of hybrid supervision with different qualities and uses the dual-branch supervisions to handle the hybrid-supervised task.

Unlike these methods, we develop a Transformer based Refinement Learning (TRL) framework to address the hybrid-supervised semantic segmentation. TRL progressively refines heat maps to generate high-quality pseudo labels for training a segmentation model.

3. Method

3.1. Problem definition

Hybrid-supervised semantic segmentation aims to learn a segmentation model under different kinds of supervisions. Strong supervisions use pixel-level annotated data, denoted as $\mathbf{X}^s = \left\{(\mathbf{x}_i^s, \mathbf{y}_i^s)\right\}_{i=1}^{N_s}$, where $\mathbf{x}_i^s$ and $\mathbf{y}_i^s$ are the ith strongly annotated image and label. Weak supervisions use weakly annotated data, denoted as $\mathbf{X}^w = \left\{(\mathbf{x}_i^w, \mathbf{y}_i^w)\right\}_{i=1}^{N_w}$, where $\mathbf{x}_i^w$ and $\mathbf{y}_i^w$ are the ith weakly annotated image and label. The subscripts in the following are omitted for simplicity. The number of strong annotations N_s is usually much less than that of weak annotations N_w. This paper adopts the widely-used image classes (i.e., image-level labels) as weak annotations.

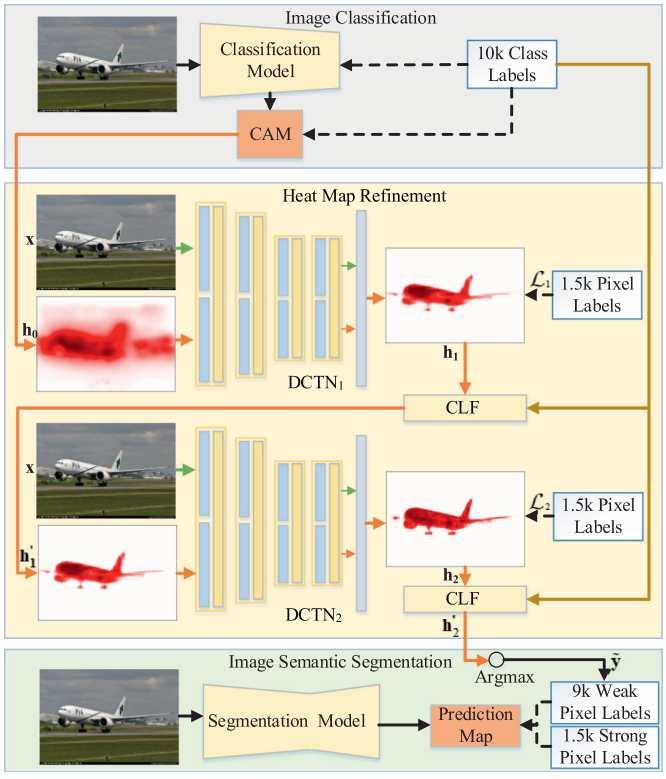

Fig. 2. Overview of our learning framework. CAM denotes class activation map [10]. DCTN_1 and DCTN_2 denote the first and the second dual-cross Transformer networks. CLF denotes the class label filtering module. We take the VOC dataset as an example.

3.2. Refinement learning process

With a large amount of image-level labeled data $\mathbf{X}^w$, we can train a multi-label classification model. Then, given an image $\mathbf{x}^w$ and its class label $\mathbf{y}^w$, an initial heat map $\mathbf{h}_0 \in \mathbb{R}^{C\times H\times W}$ can be generated from the trained classification model by the CNN visualization techniques [10,30,31], such as CAM [10], where H, W, C are the height, width and number of classes. However, these heat maps cannot be directly adopted to produce high-quality pseudo labels $\tilde{\mathbf{y}} \in \mathbb{R}^{H\times W}$ for the segmentation training. Therefore, we propose a refinement learning method (i.e., TRL), which progressively refines the heat maps under the supervision of the small amount of pixel-level labeled data $\mathbf{X}^s$ to generate high-quality pseudo labels (Fig. 2).

The initial heat map $\mathbf{h}_0$ and the strongly annotated image $\mathbf{x}^s$ are fed into the first dual-cross Transformer network, i.e., DCTN_1 (detailed latter) and are trained with the label $\mathbf{y}^s$. We use the standard cross-entropy loss l_{ce} as the training loss $\mathcal{L}$:

$$\mathcal{L} = l_{ce}(\text{DCTN}_1(\mathbf{x}^s, \mathbf{h}_0; \theta_1), \mathbf{y}^s), \quad (\mathbf{x}^s, \mathbf{y}^s) \in \mathbf{X}^s, \tag{1}$$

where θ_1 denotes the network parameters. In the inference phase, given the weakly annotated image $\mathbf{x}^w$, DCTN_1 outputs a prediction map $\mathbf{h}_1$, i.e., the higher-quality heat map.

Following DCTN_1, a class label filtering module (i.e., CLF) is added to boost the quality of the heat map, by setting the channels of non-existent classes in the prediction map to zero according to the provided class labels $\mathbf{y}^w$. Thus, a filtered prediction map $\mathbf{h}_1'$ is obtained as $\mathbf{h}_1' = \text{CLF}(\mathbf{h}_1, \mathbf{y}^w)$.

Then, $\mathbf{h}_1'$ is treated as a new input heat map and it is fed with the image into the dual-cross Transformer network to train a second network with the same loss in Eq. (1). At the second round, we add FPN [32,33] to the second dual-cross Transformer network (DCTN_2) to further improve the capacity of refinement. In the inference phase, an output heat map $\mathbf{h}_2'$ is obtained by DCTN_2 and another CLF, as

$$\mathbf{h}_2' = \text{CLF}(\mathbf{h}_2, \mathbf{y}^w) = \text{CLF}(\text{DCTN}_2(\mathbf{x}^w, \mathbf{h}_1'; \theta_2), \mathbf{y}^w), \tag{2}$$

where θ_2 denotes the parameters of DCTN_2.

Finally, a large amount of image-level labeled data $\mathbf{X}^w$ are fed into the refining process to produce high-quality pseudo labels $\tilde{\mathbf{y}}$ (where $\tilde{\mathbf{y}} = \text{argmax}(\mathbf{h}_2')$), without using any manually selected threshold or post-processing technique (e.g., CRF [18]). Afterwards, the generated pseudo (weak) labels and the small number of original (strong) labels are fed into a segmentation model to train the final semantic segmentation.

How can refinement learning outperform the direct learning for generating pseudo labels? Instead of directly learning a model from the original images, the proposed refinement learning decomposes the whole procedure into the progressive processes. Moreover, with the assistance of heat maps as a part of the inputs, the learning process will be much easier, as heat maps can be deemed as high-level features.

3.3. Dual-Cross Transformer network

In our learning framework, the input source contains two different types of data, i.e., natural images and heat maps. Between these two data, there exists a large distribution gap, which increases the learning difficulty. To address this problem, we propose a dual-stream Transformer with cross attentions, namely dual-cross Transformer network (DCTN), to perform the heat map refinement.

As shown in Fig. 3, the proposed model contains a warm-up module, low-level feature extractors, four-stage Transformer blocks, and prediction heads. The warm-up module consists of two batch normalizations with a point-wise convolution in between. This module is used to normalize the distribution of heat maps. Each of the low-level feature extractors is composed of two layers of convolutional blocks, which are used to encode and preserve the low-level features. The Transformer blocks contain two parallel branches, each of which handles one source stream. The parallel architecture allows the model to learn different features individually. Nevertheless, these two-stream Transformers are not entirely independent. We devise a cross Transformer block for the last module at each stage, which allows the information from the two-stream branches to be fused and compensated for each other.

Cross Transformer block. As illustrated in the right part of Fig. 3, we propose to construct a cross Transformer block, which contains dual inputs and dual outputs corresponding to the image stream and the heat map stream. The queries $\mathbf{q}_1$ and keys $\mathbf{k}_1$ for the image stream are computed from the image stream input $\mathbf{x}_{in}$. The queries $\mathbf{q}_2$ and keys $\mathbf{k}_2$ for the heat map stream are computed from the heat map stream input $\mathbf{h}_{in}$. The formulations are written as

$$\begin{aligned} \mathbf{q}_1 &= \text{Conv}_{q1}(\text{LN}(\mathbf{x}_{in})), \quad \mathbf{k}_1 = \text{Conv}_{k1}(\text{LN}(\mathbf{x}_{in})), \\ \mathbf{q}_2 &= \text{Conv}_{q2}(\text{LN}(\mathbf{h}_{in})), \quad \mathbf{k}_2 = \text{Conv}_{k2}(\text{LN}(\mathbf{h}_{in})), \end{aligned} \tag{3}$$

where Conv_ρ denotes the 1×1 convolution for the corresponding outputs ρ. LN denotes Layer Normalization [34]. Specially, the values $\mathbf{v}_1$ for the image stream and the values $\mathbf{v}_2$ for the heat map are derived from the cross stream inputs (i.e., $\mathbf{h}_{in}$ and $\mathbf{x}_{in}$) as cross connections:

$$\mathbf{v}_1 = \text{Conv}_{v1}(\text{LN}(\mathbf{h}_{in})), \quad \mathbf{v}_2 = \text{Conv}_{v2}(\text{LN}(\mathbf{x}_{in})). \tag{4}$$

Then, cross attentions $\mathbf{a}_1$ and $\mathbf{a}_2$ are obtained by the scaled dot product [35]:

$$\begin{aligned} \mathbf{a}_1 &= \text{Conv}_{a1}(\text{softmax}(s\mathbf{q}_1\mathbf{k}_1^T)\mathbf{v}_1), \\ \mathbf{a}_2 &= \text{Conv}_{a2}(\text{softmax}(s\mathbf{q}_2\mathbf{k}_2^T)\mathbf{v}_2), \end{aligned} \tag{5}$$

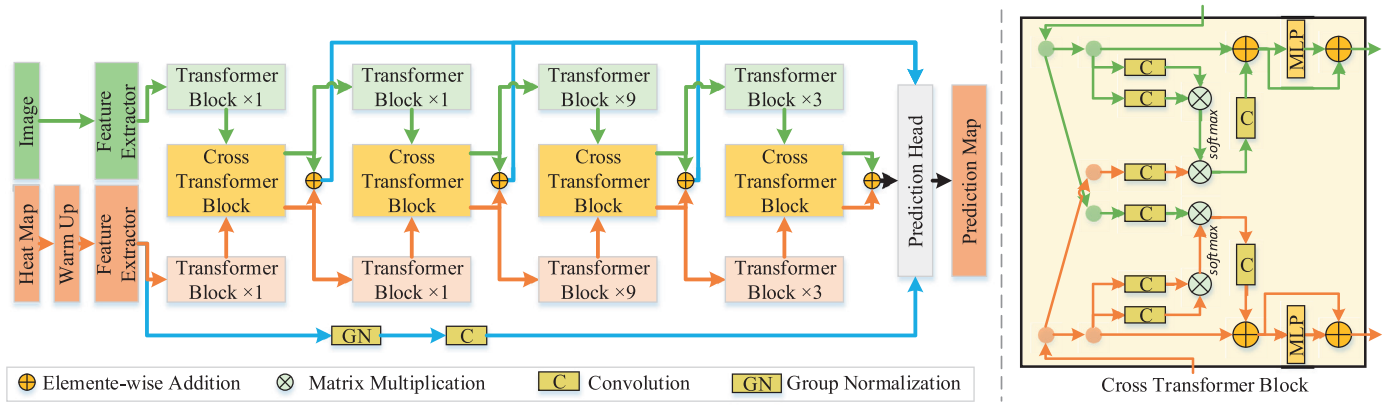

Fig. 3. The pipeline of the proposed dual-cross Transformer network (DCTN). The left part is the proposed network. The right part is the cross Transformer block as the last module at each stage. The patch embedding, the position encoding generator and the layer normalization are omitted for convenience.

where s denotes the scaling factor. Note that the cross attentions can be combined with a multi-head mechanism, which is omitted here for simplicity. Next, we obtain the residual cross attentions ($\mathbf{r}_1 = \mathbf{x}_{in} + \mathbf{a}_1, \mathbf{r}_2 = \mathbf{h}_{in} + \mathbf{a}_2$). Further, the outputs of the cross Transformer block $\mathbf{x}_{out}$ and $\mathbf{h}_{out}$ are achieved by passing $\mathbf{r}_1$ and $\mathbf{r}_2$ through the MLP module, which can be formulated as $\mathbf{x}_{out} = \mathbf{r}_1 + \mathrm{MLP}(\mathrm{LN}(\mathbf{r}_1))$, $\mathbf{h}_{out} = \mathbf{r}_2 + \mathrm{MLP}(\mathrm{LN}(\mathbf{r}_2))$. Cross connections for the values $\mathbf{v}_1$ and $\mathbf{v}_2$ lead to better feature fusion and information interaction.

Prediction head. In DCTN_1, we sum the two-stream outputs $\mathbf{h}_{out(4)}$ and $\mathbf{x}_{out(4)}$ of the fourth-stage block and then feed the summation into a prediction head. The head for DCTN_1 consists of a layer normalization LN and a 1×1 convolution Conv_{s1}. Formally, the first prediction head is formulated as

$$\mathrm{Head}_1 = \mathrm{Conv}_{s1}(\mathrm{LN}(\mathbf{h}_{out(4)} + \mathbf{x}_{out(4)})). \tag{6}$$

In DCTN_2, the two-stream outputs of the four-stage blocks are collected and fed into FPN [32,33]. Then, they are fused with the low-level feature maps of the heat map stream $\mathbf{h}_{low}$. Thus, the second prediction head is formulated as

$$\begin{aligned}\mathrm{Head}_2 = \mathrm{FPN}\big(&\mathrm{LN}(\mathbf{h}_{out(1)} + \mathbf{x}_{out(1)}),\\ &\mathrm{LN}(\mathbf{h}_{out(2)} + \mathbf{x}_{out(2)}), \mathrm{LN}(\mathbf{h}_{out(3)} + \mathbf{x}_{out(3)}),\\ &\mathrm{LN}(\mathbf{h}_{out(4)} + \mathbf{x}_{out(4)})\big) + \mathrm{GN}(\mathrm{Conv}_{s2}(\mathbf{h}_{low})),\end{aligned} \tag{7}$$

where GN is Group Normalization [36].

Differences between DCTN and segmentation Transformers. (1) Function: Segmentation Transformers process natural images, and output the class mask. DCTN performs the heat map refinement, where it processes both heat maps and images, and outputs refined heat maps. (2) Framework: The segmentation Transformers, e.g., Twins [37] and Swin [38], usually contain single-stream Transformer blocks. DCTN comprises two parallel Transformers with the cross Transformer blocks at the end of each stage, which can hierarchically fuse the two modal features. (3) Attention: Compared with the common self-attention, the proposed cross attention (Eq. (5)) computes the correlation within one stream (e.g., both $\mathbf{q}_1$ and $\mathbf{k}_1$ from image features) to recompose the other stream (e.g., $\mathbf{v}_1$ from heat map features). As shown in Fig. 4, DCTN can successfully attend to the target, while the dual Transformer network without cross attention fails.

4. Experiments

4.1. Implementation details

Network architecture. During the refinement learning, we use DCTN based on the backbone of Twins-SVT [37] to generate pseudo

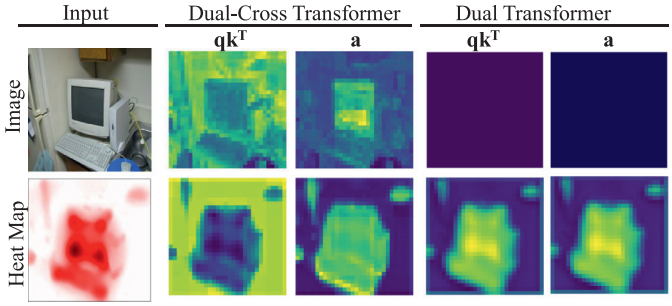

Fig. 4. Comparison of attention maps in dual-cross Transformer network (DCTN) and dual Transformer network (DTN). DCTN can probe the spatial similarity (of tv, keyboard, and background) by $\mathbf{qk}^T$, and attend to the target (of tv) by $\mathbf{a}$. However, DTN without cross attention fails.

Table 1
Hyper-parameters of the proposed DCTN.

	Layer	Patch size	Channel number	Sub-window size	Output size
Warm up	Conv0	-	64	-	$H \times W$
Feature extractor	Conv1	-	64	-	$H \times W$
	Conv2	-	64	-	$H \times W$
Transformer	Stage 1	4	64	5	$H/4 \times W/4$
	Stage 2	2	128	5	$H/8 \times W/8$
	Stage 3	2	256	5	$H/16 \times W/16$
	Stage 4	1	512	5	$H/16 \times W/16$

labels. The hyper-parameters of DCTN are listed in Table 1. More details of Transformer can be found in [37]. During the segmentation learning, following the baseline SWDN [16], we use the DeepLabV1 model [39] with ResNet101 as the backbone model, which is trained under the strong-weak dual-branch supervisions. Otherwise, we will state specific backbones.

Datasets. The experiments are conducted on the PASCAL VOC 2012 [19] and COCO 2017 [20] datasets. Following [16], for VOC, 1.5k pixel-level labels are used as the strong annotations. 9k image-level labels are used as the weak annotations, which are adopted to generate pseudo labels. Similarly, for COCO, the proportion of strong/weak annotations is 20k/98k.

Experimental settings. We apply horizontal flip, random rescaling, and cropping (320×320 for the refinement learning and 328×328 for the segmentation learning) for data augmentations. We adopt the Adam algorithm [40] with an initial learning rate of 1e-4 and a decay rate of 0.1 for optimization. During the refinement learning, the models are trained with the batch size of 8 for

Table 2
Ablation study for TRL framework. mIoU is reported on the VOC train-aug set [41]. "STN" denotes the original single-stream Transformer network [37]. "CLF_1" and "CLF_2" denote the first and second CLF modules.

Image	CAM	STN	$DCTN_1$	CLF_1	$DCTN_2$	CLF_2	mIoU
	✓						47.3
✓		✓					53.8
	✓	✓					57.5
✓	✓	✓					69.4
✓	✓		✓				72.7
✓	✓		✓	✓			74.4
✓	✓		✓	✓	✓		75.6
✓	✓		✓	✓	✓	✓	**75.7**

Table 3
Ablation study for DCTN model, focusing on the first procedure. DTN means the dual Transformer network without cross attention.

Method	Cross attention	Warm up	FPN	mIoU
DTN		✓		62.2
$DCTN_1$ w/o warm up	✓			68.5
$DCTN_1$ w/ FPN	✓	✓	✓	68.3
$DCTN_1$	✓	✓		**72.7**

150 epochs on VOC and 30 epochs on COCO. During the segmentation learning, the models are trained with the batch size of 4 for 32 epochs on VOC and 10 epochs on COCO. We adopt mIoU as the evaluation metric for all experiments.

Computational complexity. In the testing phase of segmentation, we only perform segmentation by using the DeepLab model without running DCTN, which consumes the same FLOPs (on 328 × 328 image) as our baseline SWDN [16] (i.e. 83 G).

4.2. Ablation studies

The ablation studies are conducted on the VOC dataset by evaluating the quality of generated pseudo labels.

Analysis of the TRL framework. As shown in Table 2, the initial quality of CAM (47.3% mIoU) is reported as reference. When only using original 1.5k images to train a single-stream Transformer network, the mIoU of the generated pseudo labels is 53.8%. On the other hand, when only using the original 1.5k heat maps from CAM to train a single-stream network, the mIoU of generated pseudo labels is 57.5%. We also test a single Transformer with images and heat maps concatenated together as input, resulting in 69.4% mIoU. All of these three results are not satisfactory. In contrast, we adopt images and heat maps as two-stream inputs to train the proposed $DCTN_1$, achieving 72.7% mIoU, which is 18.9%, 15.2% and 3.3% higher than those under the single-stream settings respectively.

When gradually appending CLF_1, $DCTN_2$ and CLF_2, the quality of generated pseudo labels is improved correspondingly. At last, our method reaches 75.7% mIoU, which outperforms CAM by a large margin (28.4%). Notice that the pseudo labels in our baseline method [16] is produced by DSRG [17]. DSRG obtains an mIoU of 60.1%, which is also far behind our method. This experiment demonstrates the effectiveness of each proposed learning stage. Some visual examples can be found in Fig. 1.

Analysis of the DCTN model. Since the difference between the $DCTN_1$ and $DCTN_2$ is only the prediction head, we will emphasize the analysis on $DCTN_1$ and the prediction head. Table 3 shows that employing a dual Transformer network without cross attention worsens the mIoU of the pseudo labels at the first stage from 72.7% to 62.2%. This validates the importance of cross attention inside the proposed DCTN, since the information interaction between the two streams is essential for feature fusion and complementation. When we discard the warm-up module, the performance decreases by 4.2%. When adding FPN in the prediction head in $DCTN_1$, the performance drops by 4.4%. This indicates that proper prediction heads suit the different learning stages.

Table 4
Segmentation results on the VOC val and test sets. FullSup is trained with full 10.5k pixel-level labels, where its result is produced with the code from Luo and Yang [16]. Other results are copied from the original papers. † denotes our baseline.

Method	Model	Backbone	Val-	Test
FullSup [39]	DeepLabV1	ResNet101	77.7	-
WSSL [42]	DeepLabV1-CRF	VGG16	64.6	66.2
DSRG [17]	DeepLabV2	VGG16	64.3	-
MDC [43]	DeepLabV1-CRF	VGG16	65.7	67.6
FickleNet [27]	DeepLabV2	ResNet101	65.8	-
CCT [21]	PSP-Net	ResNet50	73.2	-
PseudoSeg [22]	DeepLabV3+	ResNet50	73.8	-
DCC [23]	DeepLabV3+	ResNet50	76.1	-
AdvCAM [29]	DeepLabV2	ResNet101	77.8	76.0
SWDN† [16]	DeepLabV1	ResNet101	76.6	77.1
TRL (Ours)	DeepLabV1	ResNet101	**78.5**	**78.6**

4.3. Comparison with state-of-the-art methods

Results on VOC. As shown in Table 4, we compare the proposed TRL with the state-of-the-art methods. For a fair comparison, the employed models and backbones of all methods are listed in the table.

Our TRL achieves 78.5% and 78.6% mIoU on the VOC val set and test set respectively, which are 1.9% and 1.5% higher than the baseline SWDN. The improvement of our method over the baseline is mainly attributed to the high quality of the pseudo labels generated by our method. Surprisingly, we observe that our method even surpasses the fully-supervised method (i.e., FullSup [39]). This is because under the dual-branch supervisions [16], our method is possible to surpass FullSup when the quality of the pseudo labels generated by our method is good enough.

Moreover, our TRL outperforms several consistency learning based methods, e.g., CCT [21] and DCC [23], as they do not consider promoting the quality of pseudo labels. TRL also surpasses the weakly-supervised learning based methods (i.e., DSRG [17] and FickleNet [27]) by a large margin (i.e., 14.2% and 12.7% gains) on the val set, because their pseudo labels are unsatisfactory. AdvCAM [29] produces relatively high-quality pseudo labels, which are still inferior to ours, and it obtains the best result among the other competitors. Unlike AdvCAM applying consistency training, our method does not use the consistency training, but it outperforms AdvCAM by 2.6% gains on the test set.

The superior performance of our TRL method can be attributed to the fact that it explores the heat map refinement learning to improve the quality of pseudo labels, which is neglected by most of the state-of-the-art methods.

Results on COCO. We further validate the proposed TRL on the COCO dataset. COCO includes a large amount of data and it is much more challenging than VOC. We apply TRL to refine the heat maps yielded by CAM and generate the pseudo labels, which are used to train the segmentation model. Table 5 shows that our method outperforms the baseline by 2.1% when using VGG16 as backbone, and 5.5% when using ResNet101 as backbone. Our method also outperforms LPLN [44] by a large margin (i.e., 18.1%). Notice that our method again surpasses the fully-supervised learning method, i.e., FullSup [39], which is with the same backbone as ours. These results on COCO further verify the generalization of TRL.

Different data proportions. To validate the robustness of the proposed method, we evaluate our method on different propor-

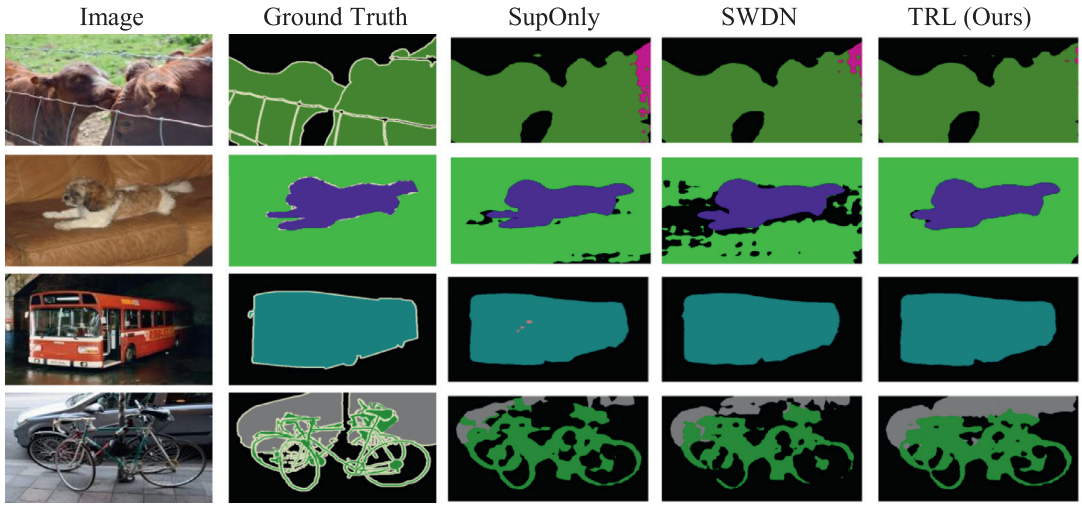

Fig. 5. Visual examples of segmentation results obtained by the competing methods. The baseline supervised only by 1.5k strong annotations (i.e., SupOnly) and the baseline SWDN are compared with our TRL.

Table 5
Segmentation results on the COCO val set. † denotes our baseline. FullSup means the fully-supervised method for the DeepLab model.

Method	Backbone	Data	mIoU
FullSup [39]	VGG16	s20k	46.1
FullSup [39]	VGG16	s118k	48.9
LPLN [44]	VGG16	s20k+w60k	31.6
SWDN† [16]	VGG16	s20k+w98k	47.6
TRL (Ours)	VGG16	s20k+w98k	**49.7**
TRL (Ours)	ResNet101	s20k+w98k	**53.1**

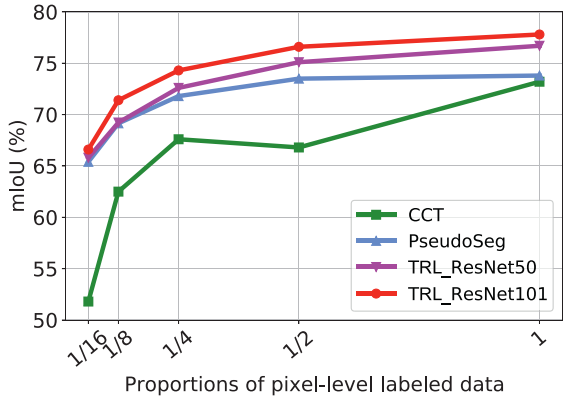

Fig. 6. Segmentation results on the VOC val set by different methods trained on different splits. The values of each point for CCT and PseudoSeg are taken from Zou et al. [22].

tions of strong annotations. Following the practice in [22], we randomly subsample 1/2, 1/4, 1/8, and 1/16 of images from the standard VOC training set as the strong annotations. As shown in Fig. 6, our method consistently outperforms the competitors on all data splits, especially when the scale of strong annotations becomes larger. When the backbone ResNet50 is replaced with ResNet101, our method obtains further performance improvement.

4.4. Visual results

Fig. 5 presents some visual examples of the segmentation results on VOC 2012. From the first row of Fig. 5, we can observe that the baseline supervised only by 1.5k strong annotations (i.e., SupOnly) and the baseline SWDN predict some false regions inside the object "cow", while the segmentation map obtained by our TRL contains less regions of the wrong category. Similarly, in the third row, SupOnly outputs several false points in the object "bus", while the output of our TRL is correct for the object "bus". Combining the observations in the first and third rows, our TRL can predict fewer wrong foreground classes compared with SupOnly and SWDN. Moreover, from the second row, we can see that SupOnly and SWDN mistakenly recognize some regions in the object "sofa" as background, while our TRL can segment the "sofa" regions more completely. Similarly in the fourth row, TRL outputs larger true regions of "car" and "bicycle" compared with SupOnly and SWDN. According to the second and fourth rows, our TRL can predict more true foreground regions, even under a very challenging circumstance with occlusions (e.g., "car" and"bicycle" in the fourth row).

5. Conclusion

This paper formally defines the learning paradigm of hybrid-supervised semantic segmentation. We introduce a novel and simple method, i.e., Transformer based refinement learning (TRL), to advance the quality of pseudo labels for hybrid-supervised semantic segmentation. We design a dual-cross Transformer network (DCTN) to perform the heat map refinement learning. By using TRL, we obtain high-quality pseudo labels, which are used to more effectively train the segmentation model. Extensive experiments on the PASCAL VOC and COCO datasets demonstrate that our method is superior to several state-of-the-art competitors for hybrid-supervised semantic segmentation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the National Natural Science Foundation of China [Grant numbers U21A20514, 61872307, 62002302, 62071404]; the China Fundamental Research Funds for the Central Universities [Grant number 20720210099].

References

[1] S. Cha, Y. Wang, Zero-shot semantic segmentation via spatial and multi-scale aware visual class embedding, Pattern Recognit. Lett. 158 (2022) 87–93.
[2] W. Huang, Z. Shao, M. Luo, P. Zhang, Y. Zha, A novel multi-loss-based deep adversarial network for handling challenging cases in semi-supervised image semantic segmentation, Pattern Recognit. Lett. 146 (2021) 208–214.
[3] N. Luo, Y. Wang, Y. Gao, Y. Tian, Q. Wang, C. Jing, kNN-based feature learning network for semantic segmentation of point cloud data, Pattern Recognit. Lett. 152 (2021) 365–371.
[4] J. Hong, W. Li, J. Han, J. Zheng, P. Fang, M. Harandi, L. Petersson, GOSS: towards generalized open-set semantic segmentation, 2022. arXiv preprint arXiv:2203.12116v1.
[5] G. Dong, Y. Yan, C. Shen, H. Wang, Real-time high-performance semantic image segmentation of urban street scenes, IEEE Trans. Intell. Transp. Syst. 22 (6) (2021) 3258–3274.
[6] O. Ronneberger, P. Fischer, T. Brox, U-Net: convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention, vol. 9351, 2015, pp. 234–241.
[7] Z. Huang, Z. Zheng, C. Yan, H. Xie, Y. Sun, J. Wang, J. Zhang, Real-world automatic makeup via identity preservation makeup net, in: IJCAI, 2021, pp. 652–658.
[8] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The cityscapes dataset for semantic urban scene understanding, in: IEEE CVPR, 2016, pp. 3213–3223.
[9] A.L. Bearman, O. Russakovsky, V. Ferrari, L. Fei-Fei, What's the point: semantic segmentation with point supervision, in: ECCV, vol. 9911, 2016, pp. 549–565.
[10] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: IEEE CVPR, 2016, pp. 2921–2929.
[11] A. Kolesnikov, C.H. Lampert, Seed, expand and constrain: three principles for weakly-supervised image segmentation, in: ECCV, vol. 9908, 2016, pp. 695–711.
[12] J. Ahn, S. Kwak, Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation, in: IEEE CVPR, 2018, pp. 4981–4990.
[13] Z. Ke, D. Qiu, K. Li, Q. Yan, R.W.H. Lau, Guided collaborative training for pixel–wise semi-supervised learning, in: ECCV, vol. 12358, 2020, pp. 429–445.
[14] G. French, S. Laine, T. Aila, M. Mackiewicz, G.D. Finlayson, Semi-supervised semantic segmentation needs strong, varied perturbations, BMVC, 2020.
[15] X. Chen, Y. Yuan, G. Zeng, J. Wang, Semi-supervised semantic segmentation with cross pseudo supervision, in: IEEE CVPR, 2021, pp. 2613–2622.
[16] W. Luo, M. Yang, Semi-supervised semantic segmentation via strong-weak dual-branch network, in: ECCV, vol. 12350, 2020, pp. 784–800.
[17] Z. Huang, X. Wang, J. Wang, W. Liu, J. Wang, Weakly-supervised semantic segmentation network with deep seeded region growing, in: IEEE CVPR, 2018, pp. 7014–7023.
[18] P. Krähenbühl, V. Koltun, Efficient inference in fully connected CRFs with gaussian edge potentials, in: NeurIPS, 2011, pp. 109–117.
[19] M. Everingham, L.V. Gool, C.K.I. Williams, J.M. Winn, A. Zisserman, The pascal visual object classes (VOC) challenge, Int. J. Comput. Vis. 88 (2) (2010) 303–338.
[20] T. Lin, M. Maire, S.J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft COCO: common objects in context, in: ECCV, vol. 8693, 2014, pp. 740–755.
[21] Y. Ouali, C. Hudelot, M. Tami, Semi-supervised semantic segmentation with cross-consistency training, in: IEEE CVPR, 2020, pp. 12671–12681.
[22] Y. Zou, Z. Zhang, H. Zhang, C. Li, X. Bian, J. Huang, T. Pfister, PseudoSeg: designing pseudo labels for semantic segmentation, ICLR, 2021.
[23] X. Lai, Z. Tian, L. Jiang, S. Liu, H. Zhao, L. Wang, J. Jia, Semi-supervised semantic segmentation with directional context-aware consistency, in: IEEE CVPR, 2021, pp. 1205–1214.
[24] W. Hung, Y. Tsai, Y. Liou, Y. Lin, M. Yang, Adversarial learning for semi-supervised semantic segmentation, in: BMVC, 2018, p. 65.
[25] N. Souly, C. Spampinato, M. Shah, Semi supervised semantic segmentation using generative adversarial network, in: IEEE ICCV, 2017, pp. 5689–5697.
[26] B. Zoph, G. Ghiasi, T. Lin, Y. Cui, H. Liu, E.D. Cubuk, Q. Le, Rethinking pre-training and self-training, NeurIPS, 2020.
[27] J. Lee, E. Kim, S. Lee, J. Lee, S. Yoon, FickleNet: weakly and semi-supervised semantic image segmentation using stochastic inference, IEEE CVPR, 2019.
[28] J. Ahn, S. Cho, S. Kwak, Weakly supervised learning of instance segmentation with inter-pixel relations, in: IEEE CVPR, 2019, pp. 2209–2218.
[29] J. Lee, E. Kim, S. Yoon, Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation, in: IEEE CVPR, 2021, pp. 4071–4080.
[30] L. Cheng, P. Fang, Y. Liang, L. Zhang, C. Shen, H. Wang, TSGB: target-selective gradient backprop for probing CNN visual saliency, IEEE Trans. Image Process. 31 (2022) 2529–2540.
[31] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: visualising image classification models and saliency maps, ICLR, 2014.
[32] T. Lin, P. Dollár, R.B. Girshick, K. He, B. Hariharan, S.J. Belongie, Feature pyramid networks for object detection, in: IEEE CVPR, 2017, pp. 936–944.
[33] A. Kirillov, R.B. Girshick, K. He, P. Dollár, Panoptic feature pyramid networks, in: IEEE CVPR, 2019, pp. 6399–6408.
[34] L.J. Ba, J.R. Kiros, G.E. Hinton, Layer normalization, CoRR (2016) abs/1607.06450.
[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: NeurIPS, 2017, pp. 5998–6008.
[36] Y. Wu, K. He, Group normalization, Int. J. Comput. Vis. 128 (3) (2020) 742–755.
[37] X. Chu, Z. Tian, Y. Wang, B. Zhang, H. Ren, X. Wei, H. Xia, C. Shen, Twins: revisiting spatial attention design in vision transformers, CoRR (2021) abs/2104.13840.
[38] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: hierarchical vision transformer using shifted windows, in: IEEE ICCV, 2021, pp. 10012–10022.
[39] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Semantic image segmentation with deep convolutional nets and fully connected CRFs, ICLR, 2015.
[40] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, ICLR, 2015.
[41] B. Hariharan, P. Arbelaez, L.D. Bourdev, S. Maji, J. Malik, Semantic contours from inverse detectors, in: IEEE ICCV, 2011, pp. 991–998.
[42] G. Papandreou, L. Chen, K.P. Murphy, A.L. Yuille, Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation, in: IEEE ICCV, 2015, pp. 1742–1750.
[43] Y. Wei, H. Xiao, H. Shi, Z. Jie, J. Feng, T.S. Huang, Revisiting dilated convolution: a simple approach for weakly- and semi-supervised semantic segmentation, in: IEEE CVPR, 2018, pp. 7268–7277.
[44] R. Yi, Y. Huang, Q. Guan, M. Pu, R. Zhang, Learning from pixel-level label noise: a new perspective for semi-supervised semantic segmentation, IEEE Trans. Image Process. 31 (2022) 623–635.