Boundary Knowledge Translation based Reference Semantic Segmentation

Lechao Cheng[1] [†], **Zunlei Feng**[2] [†], **Xinchao Wang**[3], **Ya Jie Liu** [1], **Jie Lei** [4], **Mingli Song**[2] [*]
[1]Zhejiang Lab
[2]Zhejiang University
[3]National University of Singapore
[4]Zhejiang University Of Technology
{chenglc,liuyj}@zhejianglab.com, {zunleifeng,brooksong}@zju.edu.cn, xinchao@nus.edu.sg,
jasonlei@zjut.edu.cn

Abstract

Given a reference object of an unknown type in an image, human observers can effortlessly find the objects of the same category in another image and precisely tell their visual boundaries. Such visual cognition capability of humans seems absent from the current research spectrum of computer vision. Existing segmentation networks, for example, rely on a humongous amount of labeled data, which is laborious and costly to collect and annotate; besides, the performance of segmentation networks tend to downgrade as the number of the category increases. In this paper, we introduce a novel Reference semantic segmentation Network (Ref-Net) to conduct visual boundary knowledge translation. Ref-Net contains a Reference Segmentation Module (RSM) and a Boundary Knowledge Translation Module (BKTM). Inspired by the human recognition mechanism, RSM is devised only to segment the same category objects based on the features of the reference objects. BKTM, on the other hand, introduces two boundary discriminator branches to conduct inner and outer boundary segmentation of the target object in an adversarial manner, and translate the annotated boundary knowledge of open-source datasets into the segmentation network. Exhaustive experiments demonstrate that, with tens of finely-grained annotated samples as guidance, Ref-Net achieves results on par with fully supervised methods on six datasets.

1 Introduction

In recent years, deep neural networks have triumphed over many computer vision problems, including semantic segmentation, which is critical in emerging autonomous driving and medical image diagnostics applications. In general, training deep neural networks requires a humongous amount of labeled data, which is laborious and costly to collect and annotate. To alleviate the annotation burden, some learning techniques, such as few-shot learning and transfer learning, have been proposed. The former aims to train models using only a few annotated samples, while the latter focuses on transferring the models learned on one domain to another novel one. Despite the recent progress in few-shot and transfer learning, existing methods are still prone to either inferior results, or the rigorous requirement that the two tasks are strongly related and a large number of annotated samples are required.

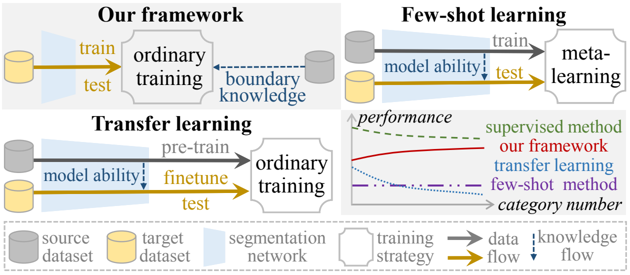

Figure 1: The differences between different frameworks. The main differences include three aspects. First, the source sample does not pass through the segmentation network in our framework, while others' input contains both source and target samples. Second, few-shot learning adopts meta-learning as the training strategy, while others not. Third, our framework translates dataset-level knowledge into the segmentation network, while others transfer model-level knowledge for target dataset; this leads to the fact that the performance of our framework enhances as the number of the category increases.

For many tasks, including few-shot and transfer learning, the performances of existing approaches, such as segmentation networks, deteriorate as the number of object categories increase, as demonstrated by prior works [Vinyals *et al.*, 2016; Chen *et al.*, 2019], and also by our experiments. The root cause lies in that, existing approaches are devised to recognize the category-wise features and segment the corresponding objects. Recently, boundary-aware features have been introduced to enhance the segmentation results, yet their core frameworks still focus on classifying category-wise features and segmenting the corresponding objects.

In the paper, we propose a novel Reference semantic segmentation Network (Ref-Net) based on visual boundary knowledge translation. In Ref-Net, only the target samples pass through the segmentation network, while the boundary knowledge of open-source datasets is translated into the segmentation network in an adversarial manner. That means only

[*]Corresponding authors.

[†]These authors contributed equally to this work.

the data flow of the target dataset will pass through the segmentation network. Fig. 1 shows the differences between the proposed framework and related frameworks. More importantly, Ref-Net is the first proposed dataset-level knowledge translation framework, which is different from the model-level knowledge transfer of related frameworks. The fatal limitation of model-level knowledge transfer is that the specific segmentation ability of the model for the source categories is beneficial for the segmentation of target categories, but also will limit the bound of the performance on the target categories. The most direct evidence is that the segmentation performance will drop dramatically when the number of the category increases. By contrast, it is noteworthy that the performance will increase when the number of the category increase for Ref-Net.

With a reference object, the human vision system can effortlessly find the same category objects in another image and precisely tell their visual boundary, even if they belong to an unknown category. Inspired by the above fact, the Ref-Net is devised with a Reference Segmentation Module (RSM) and a Boundary Knowledge Translation Module (BKTM), as shown in Fig. 2. In RSM, the reference object guides the segmentation network to find objects of the same category. Meanwhile, BKTM is proposed to assist the segmentation network in handling the accurate boundary segmentation, through translating the annotated boundary knowledge of open-source datasets. The object category of open-source datasets can be totally different from the target dataset.

RSM is devised for segmenting the same category objects in the target samples with several annotated reference objects. For the annotated reference image, the extracted features by the first branch are set as a condition, which will be concatenated with the target image features. Based on the condition, the semantic segmentation network learns to find and segment the same category objects in the target images.

To alleviate the burden of laborious and costly annotations, we adopt BKTM to translate the general boundary knowledge of abundant open-source segmentation datasets into the boundary segmentation ability of segmentation network. For accurate segmentation, the segmented object should not contain any background feature; meanwhile, the segmented background should not have residual object features. Inspired by this fact, BKTM is designed to comprise an inner boundary discriminator and an outer discriminator, as shown in Fig. 2. The outer boundary discriminator distinguishes whether the segmented objects contain the features of the outer background. Meanwhile, the inner boundary discriminator distinguishes whether the segmented background contains features of the inner objects.

Our contribution is therefore the first dataset-level knowledge translation based Ref-Net for semantic segmentation, which brings increased performance as the number of target category increases. Also, a boundary-aware self-supervision and a category-wise constraint are proposed to enhance the segmentation consistency on both the image-level and representation level, respectively. We evaluate the proposed Ref-Net on a wide domain of image datasets, and show that, with only ten annotated samples, our method achieves close results on par with fully supervised ones.

2 Related Works

For **boundary-aware semantic segmentation**, the commonly adopted framework is a two-branch network that simultaneously predicts segmentation maps and boundaries [Takikawa *et al.*, 2019]. Unlike predicting the boundary directly, some strategies, such as pixel's distance to boundary [Hayder *et al.*, 2017], boundary-aware filtering [Khoreva *et al.*, 2017], boundary refinement [Zhang *et al.*, 2017], boundary weights [Qin *et al.*, 2019], are proposed for improving the segmentation performance on boundary. Unlike the above methods, the proposed BKTM focuses on the outer boundary of the object and inner boundary of the background. Meanwhile, two boundary discriminators are devised for discriminating whether the outer boundary of the object and inner boundary of the background contain residual features.

GAN based segmentation contains two categories: mask distribution-based methods [Arbelle and Raviv, 2018] and composition fidelity based methods [Chen *et al.*, 2019]. The former discriminates mask distribution between the predicted mask and GT mask, while the latter adopts discriminator to discriminate fidelity of the composite images. Unlike the above GAN-based methods, an adversarial strategy is adopted to translate the annotated boundary knowledge of the open-source dataset into the segmentation network by two boundary discriminators.

Few-shot Segmentation (FSS) aims at training a segmentation network that can segment new category well with only a few labeled samples of those new category. It contains parametric matching-based methods [Zhang *et al.*, 2018; Xian *et al.*, 2019], prototype-based methods [Wang *et al.*, 2019a], GCN-based methods [Liu *et al.*, 2021], R-CNN based methods [Yan *et al.*, 2019] and proposal-free based methods [Gao *et al.*, 2019]. Ref-Net is expected to gain a general segmentation ability and focus on the segmentation ability of the target categories. In addition, the category number of support category in FSS is usually larger than two, while Ref-Net can handle a single category source dataset.

Transfer learning based segmentation contains pseudo-sample generation methods [Han and Yin, 2017], iterative optimization methods [Zou *et al.*, 2018], graph-based methods [Yang *et al.*, 2020a], and distillation methods [Yang *et al.*, 2020b]. Those methods aim at transferring the models' ability on source dataset into target datasets, where the capacity for source datasets still occupy the part of the segmentation ability of the model. However, the Ref-Net aims at learning a general segmentation ability with a reference image as a condition. Training on target datasets will bring more focused segmentation ability on the target categories. In addition, transfer learning based methods require the two domains as similar as possible, but Ref-Net has no such requirement.

3 Proposed Method

The proposed Ref-Net is composed of a Reference Segmentation Module (RSM) and a Boundary Knowledge Translation Module (BKTM), as shown in Fig. 2. RSM is designed for segmenting the same category objects in the target samples with the reference image as condition, which is inspired by the human recognition mechanism with a reference object.

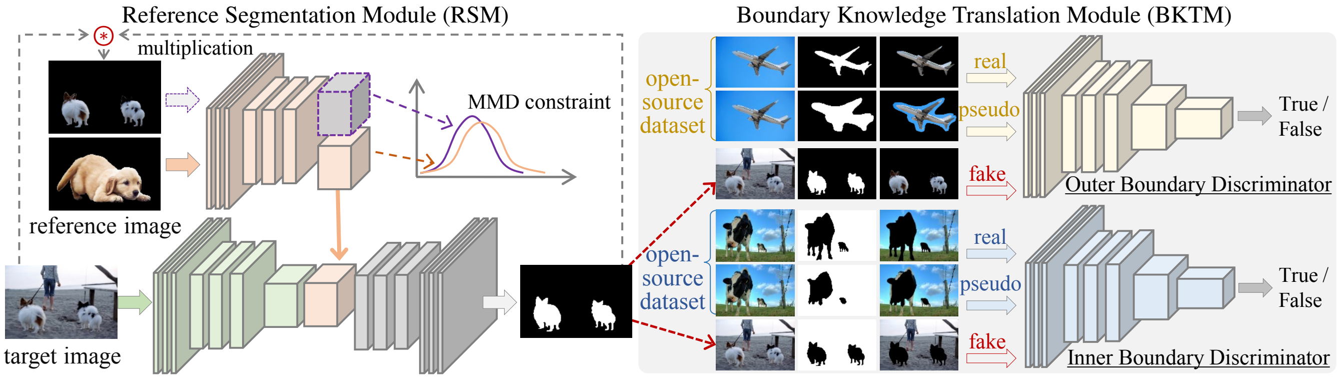

Figure 2: The framework of Ref-Net, which is composed of RSM and BKTM. In RSM, with the extracted features of reference image as condition, the segmentation network branch will find and segment the same category objects in the target image. Meanwhile, the MMD constraint is adopted to constrain the distribution consistency between the representation of the reference object and representation of segmented objects. The feature extraction branch and the encoder of the segmentation network branch share the same structure and parameters. The BKT contains the outer boundary discriminator and the inner boundary discriminator, which are devised for distinguishing whether the segmented objects and background contain residual features.

BKTM contains two boundary discriminators, which are devised for distinguishing whether segmented objects contain outer features of the background and segmented background contains inner features objects, respectively. BKTM aims to translate the annotated boundary knowledge of open-source datasets into the ability of the segmentation network, which will dramatically reduce the requirement of labeled samples for the target categories.

Formally, given a target dataset $\mathbb{S}$ with object category labels, a reference image dataset $\mathbf{S}$ with fine-grained object mask annotation and an open-source dataset $\mathbf{O}$ with fine-grained object mask annotation, the goal of Ref-Net is to learn a segmentation network for $\mathbb{S}$ with $\mathbf{S}$ as condition and guidance, while BKTM translates the annotated boundary knowledge of $\mathbf{O}$ into the segmentation network. It is noticed that the reference image dataset $\mathbf{S}$ are chosen from the target dataset $\mathbb{S}$, which has the same object category. The object category of open-source dataset $\mathbf{O}$ can be totally different from the category of the target dataset $\mathbb{S}$.

3.1 Reference Segmentation Module

Inspired by the human recognition mechanism with a reference object, RSM is devised to be composed of two network branches: the reference feature extraction branch and the target segmentation branch. The target segmentation branch is designed to be an encoder-decoder architecture. To maintain the consistency of the feature space, the reference feature extraction branch $\mathcal{F}_\theta^e$ has the same network architecture and parameter as the encoder of the target segmentation branch.

Give a target image $\bar{\boldsymbol{x}} \in \mathbb{S}$ and a reference image $(\boldsymbol{x}_k^c, \boldsymbol{m}_k^c) \in \mathbf{S}$, the extracted representations can be denoted as $\bar{\boldsymbol{r}} = \mathcal{F}_\theta^e(\bar{\boldsymbol{x}})$ and $\boldsymbol{r}_k^c = \mathcal{F}_\theta^e(\boldsymbol{x}_k^c * \boldsymbol{m}_k^c)$, where the $*$ denotes pixel-wise multiplication. With the concatenated representation $[\bar{\boldsymbol{r}}, \boldsymbol{r}_k^c]$, the decoder $\mathcal{F}_\theta^d$ of the target segmentation branch predicts the mask $\tilde{\boldsymbol{m}} = \mathcal{F}_\theta^d([\bar{\boldsymbol{r}}, \boldsymbol{r}_k^c])$. For simplicity, the reference segmentation is formulated as $\tilde{\boldsymbol{m}} = \mathcal{F}_\theta(\bar{\boldsymbol{x}}, \boldsymbol{x}_k^c * \boldsymbol{m}_k^c)$, where the $\mathcal{F}_\theta$ denotes the two-branch segmentation network.

Limited Sample Supervision. In the training stage, the limited annotated samples are also fed to the target segmentation branch, which will generate direct supervision information. So, given a random image $(\boldsymbol{x}, \boldsymbol{m}) \in \mathbf{S}$ and a reference image $(\boldsymbol{x}_k^c, \boldsymbol{m}_k^c) \in \mathbf{S}$, the segmented result $\tilde{\boldsymbol{m}} = \mathcal{F}_\theta(\boldsymbol{x}, \boldsymbol{m}_k^c * \boldsymbol{x}_k^c)$ is expected to approximate the GT mask $\boldsymbol{m}$, which can be achieved by minimizing the pixel-wise two-class Dice loss $\mathcal{L}_{\mathbf{dic}}$:

$$\mathcal{L}_{\mathbf{dic}} = 1 - \frac{2|\tilde{\boldsymbol{m}} \cap \boldsymbol{m}| + \tau}{|\tilde{\boldsymbol{m}}| + |\boldsymbol{m}| + \tau}, \tag{1}$$

where, τ is the Laplace smoothing parameter for preventing zero error and reducing overfitting, the GT mask $\boldsymbol{m}$ will be an all-zero mask when the random image $\boldsymbol{x}$ and reference image $\boldsymbol{x}_k^c$ has no same category object. For the limited annotated samples, the data augmentation strategy is adopted to increase the samples' diversity.

Representation Consistency Constraint. The same category objects in the reference image and target image should have a similar representation distribution. So, we adopt the Maximum Mean Discrepancy ($\mathcal{MMD}$) to constrain the representation distribution consistency. With the reference image $(\boldsymbol{x}_k^c, \boldsymbol{m}_k^c) \in \mathbf{S}$ and the same category target image $\bar{\boldsymbol{x}}$, the predicted mask is $\tilde{\boldsymbol{m}} = \mathcal{F}_\theta(\bar{\boldsymbol{x}}, \boldsymbol{m}_k^c * \boldsymbol{x}_k^c)$. The representation consistency loss $\mathcal{L}_{\mathbf{rep}}$ is defined as follows:

$$\mathcal{L}_{\mathbf{rep}} = \mathcal{MMD}\{f(\tilde{\boldsymbol{m}} * \bar{\boldsymbol{x}}), f(\boldsymbol{m}_k^c * \boldsymbol{x}_k^c)\}, \tag{2}$$

where f is the feature encoder for reference image.

Boundary-aware Self-supervision. To reduce the number of labeled samples, inspired by [Wang *et al.*, 2019b], we propose a boundary-aware self-supervision strategy, which can strengthen the boundary consistency of target objects. The core idea is that for the same network, the predicted mask of the *transformed input image* should be equal to the *transformed mask* predicted by the network with the original image as input. Formally, for the robust segmentation network, given an affine transformation matrix $\boldsymbol{A}$, the segmented result $\tilde{\boldsymbol{m}}' = \mathcal{F}_\theta(\boldsymbol{A}\bar{\boldsymbol{x}}, \boldsymbol{x}_k^c * \boldsymbol{m}_k^c)$ of the transformed image $\boldsymbol{A}\bar{\boldsymbol{x}}$ and the transformed result $\boldsymbol{A}\mathcal{F}_\theta(\bar{\boldsymbol{x}}, \boldsymbol{x}_k^c * \boldsymbol{m}_k^c)$ should be consistent in the following way: $\mathcal{F}_\theta(\boldsymbol{A}\bar{\boldsymbol{x}}, \boldsymbol{x}_k^c * \boldsymbol{m}_k^c) = \boldsymbol{A}\mathcal{F}_\theta(\bar{\boldsymbol{x}}, \boldsymbol{x}_k^c * \boldsymbol{m}_k^c)$. Furthermore, we obtain the boundary neighborhood weight map $\boldsymbol{w}'$ and $\boldsymbol{w}$ as follows:

$$\begin{aligned} \boldsymbol{w}' =& \mathfrak{D}_r(\tilde{\boldsymbol{m}}') - \mathfrak{E}_r(\tilde{\boldsymbol{m}}'), \\ \boldsymbol{w} =& \mathfrak{D}_r(\tilde{\boldsymbol{m}}) - \mathfrak{E}_r(\tilde{\boldsymbol{m}}), \end{aligned} \tag{3}$$

where, $\mathfrak{D}_r$ and $\mathfrak{E}_r$ denote the dilation and erosion operation with a disk strel of radius r, respectively. The weight map $\boldsymbol{w}'$ and $\boldsymbol{w}$ can strengthen the boundary consistency. The boundary-aware self-supervision loss $\mathcal{L}_{\mathbf{sel}}$ is defined as follows:

$$\mathcal{L}_{\mathbf{sel}} = ||\boldsymbol{w}' * \mathcal{F}_\theta(\boldsymbol{A}\bar{\boldsymbol{x}}, \boldsymbol{x}_k^c * \boldsymbol{m}_k^c) - \boldsymbol{A}\{\boldsymbol{w} * \mathcal{F}_\theta(\bar{\boldsymbol{x}}, \boldsymbol{x}_k^c * \boldsymbol{m}_k^c)\}||_2^2, \tag{4}$$

where, $\boldsymbol{w}'$ and $\boldsymbol{w}$ are the weight maps of the predict masks $\tilde{\boldsymbol{m}}'$ and $\tilde{\boldsymbol{m}}$, respectively. The boundary-aware self-supervision mechanism not only strengthens the boundary consistency but also eliminates the unreasonable holes.

3.2 Boundary Knowledge Translation Module

Inspired by the fact that humans can segment the boundary of the object through distinguishing whether the inner and outer of boundary contain redundant features, we devise two boundary discriminators, which can translate the boundary knowledge of the open-source dataset $\mathbf{O}$ into the reference segmentation network.

Outer Boundary Discriminator. Randomly sampling a pair of samples $\bar{\boldsymbol{x}}^c \in \mathbb{S}$ and $(\boldsymbol{x}_k^c, \boldsymbol{m}_k^c) \in \mathbf{S}$ from the target image set $\mathbb{S}$ and the reference image set $\mathbf{S}$. Next, the segmentation network predicts the mask $\tilde{\boldsymbol{m}} = \mathcal{F}_\theta(\bar{\boldsymbol{x}}^c, \boldsymbol{x}_k^c * \boldsymbol{m}_k^c)$. Then, the segmented objects $\bar{\boldsymbol{x}}_o^c$ are computed by: $\bar{\boldsymbol{x}}_o^c = \tilde{\boldsymbol{m}} * \bar{\boldsymbol{x}}^c$. The concatenated triplet $\boldsymbol{I}_a^o = [\bar{\boldsymbol{x}}^c, \tilde{\boldsymbol{m}}, \bar{\boldsymbol{x}}_o^c]$ is fed to the outer boundary discriminator $\mathcal{D}_\phi^o$, which discriminates whether the segmented objects $\bar{\boldsymbol{x}}_o^c$ contain the outer features of the background. In the paper, $\boldsymbol{I}_a^o$ is regarded as a fake triplet. Meanwhile, choosing an annotated sample $(\underline{\boldsymbol{x}}, \underline{\boldsymbol{m}}) \in \mathbf{O}$ from the open-source dataset $\mathbf{O}$, the corresponding $\boldsymbol{I}_e^o = [\underline{\boldsymbol{x}}, \underline{\boldsymbol{m}}, \underline{\boldsymbol{x}}_o]$ is labeled as a real triplet. Furthermore, we reprocess the GT mask $\underline{\boldsymbol{m}}$ of samples $\underline{\boldsymbol{x}}$ by dilation operation $\mathfrak{D}_r$ and get the pseudo triplet $\boldsymbol{I}_s^o = [\underline{\boldsymbol{x}}, \mathfrak{D}_r(\underline{\boldsymbol{m}}), \underline{\boldsymbol{x}}_o^{\mathfrak{D}}]$, where $\underline{\boldsymbol{x}} \in \mathbf{O}$ and $\underline{\boldsymbol{x}}_o^{\mathfrak{D}} = \mathfrak{D}_r(\underline{\boldsymbol{m}}) * \underline{\boldsymbol{x}}$. The generated pseudo triplet $\boldsymbol{I}_s^o$ will assist the outer boundary discriminator in distinguishing the outer features of the background. The adversarial optimization between the segmentation network and outer boundary discriminator will translate the outer boundary knowledge of the source dataset $\mathbf{O}$ into the segmentation network with the following outer boundary adversarial loss $\mathcal{L}_{\mathbf{adv}}^{\mathbf{out}}$:

$$\begin{aligned} \mathcal{L}_{\mathbf{adv}}^{\mathbf{out}} =& \frac{1}{2} \mathop{\mathbb{E}}_{\boldsymbol{I}_a^o \sim \mathbb{P}_a^o} [\mathcal{D}_\phi^o(\boldsymbol{I}_a^o)] + \frac{1}{2} \mathop{\mathbb{E}}_{\boldsymbol{I}_s^o \sim \mathbb{P}_s^o} [\mathcal{D}_\phi^o(\boldsymbol{I}_s^o)] - \mathop{\mathbb{E}}_{\boldsymbol{I}_e^o \sim \mathbb{P}_e^o} [\mathcal{D}_\phi^o(\boldsymbol{I}_e^o)] \\ & + \lambda \mathop{\mathbb{E}}_{\boldsymbol{I}^o \sim \mathbb{P}_{\boldsymbol{I}^o}} [(\|\nabla_{\boldsymbol{I}^o} \mathcal{D}_\phi^o(\boldsymbol{I}^o)\|_2 - 1)^2], \end{aligned} \tag{5}$$

where, the $\mathbb{P}_a^o$, $\mathbb{P}_s^o$, $\mathbb{P}_e^o$ are the segmented outer boundary distribution, pseudo outer boundary distribution, and real outer boundary distribution, respectively. The $\mathbb{P}_{\boldsymbol{I}^o}$ is sampled uniformly along straight lines between pairs of points sampled from the distribution $\mathbb{P}_e^o$ and the segmentation network distribution $\mathbb{P}_a^o$. The $\boldsymbol{I}^o = \varepsilon \boldsymbol{I}_e^o + (1 - \varepsilon)\boldsymbol{I}_a^o$, where the ε is a random number between 0 and 1. The gradient penalty term is firstly proposed in WGAN-GP [Gulrajani *et al.*, 2017]. The λ is the gradient penalty coefficient.

Inner Boundary Discriminator. The inner boundary discriminator $\mathcal{D}_\varphi^i$ is devised for discriminating whether the segmented background contains the inner features of the object. To obtain the segmented background, the predict background mask $\tilde{\boldsymbol{m}}'$ and GT mask $\boldsymbol{m}'$ are reprocessed with the Not-operation as follows: $\tilde{\boldsymbol{m}}' = [\mathbf{1}] - \tilde{\boldsymbol{m}}$, $\boldsymbol{m}' = [\mathbf{1}] - \boldsymbol{m}$, where the $[\mathbf{1}]$ denotes the unit matrix of $\boldsymbol{m}$'s size. Then, the corresponding fake triplet $\boldsymbol{I}_a^i = [\bar{\boldsymbol{x}}, \tilde{\boldsymbol{m}}', \bar{\boldsymbol{x}}_b]$, real triplet $\boldsymbol{I}_e^i = [\underline{\boldsymbol{x}}, \underline{\boldsymbol{m}}', \underline{\boldsymbol{x}}_b]$ and pseudo triplet $\boldsymbol{I}_s^i = [\underline{\boldsymbol{x}}, \mathfrak{D}_r(\underline{\boldsymbol{m}}'), \underline{\boldsymbol{x}}_b^{\mathfrak{D}}]$ are computed in the same manner as done in the outer boundary discriminator. The generated pseudo triplet $\boldsymbol{I}_s^i$ will also assist the inner boundary discriminator in distinguishing the inner features of objects. Similarly, the inner boundary adversarial loss $\mathcal{L}_{\mathbf{adv}}^{\mathbf{in}}$ is defined as follows:

$$\begin{aligned} \mathcal{L}_{\mathbf{adv}}^{\mathbf{in}} =& \frac{1}{2} \mathop{\mathbb{E}}_{\boldsymbol{I}_a^i \sim \mathbb{P}_a^i} [\mathcal{D}_\varphi^i(\boldsymbol{I}_a^i)] + \frac{1}{2} \mathop{\mathbb{E}}_{\boldsymbol{I}_s^i \sim \mathbb{P}_s^i} [\mathcal{D}_\varphi^i(\boldsymbol{I}_s^i)] - \mathop{\mathbb{E}}_{\boldsymbol{I}_e^i \sim \mathbb{P}_e^i} [\mathcal{D}_\varphi^i(\boldsymbol{I}_e^i)] \\ & + \lambda \mathop{\mathbb{E}}_{\boldsymbol{I}^i \sim \mathbb{P}_{\boldsymbol{I}^i}} [(\|\nabla_{\boldsymbol{I}^i} \mathcal{D}_\varphi^i(\boldsymbol{I}^i)\|_2 - 1)^2], \end{aligned} \tag{6}$$

where, the $\mathbb{P}_a^i$, $\mathbb{P}_s^i$, $\mathbb{P}_e^i$ are the segmented inner boundary distribution, pseudo inner boundary distribution, and real inner boundary distribution. $\boldsymbol{I}^i = \varepsilon \boldsymbol{I}_e^i + (1 - \varepsilon)\boldsymbol{I}_a^i$. The optimization on $\mathcal{L}_{\mathbf{adv}}^{\mathbf{in}}$ will translate the outer boundary knowledge of the open-source dataset into the segmentation network.

3.3 Complete Algorithm

To sum up, two boundary adversarial losses $\mathcal{L}_{\mathbf{adv}}^{\mathbf{out}}$ and $\mathcal{L}_{\mathbf{adv}}^{\mathbf{in}}$ are used to translate the visual boundary knowledge of the source dataset $\mathbf{O}$ into the segmentation network. The basic reconstruction loss $\mathcal{L}_{\mathbf{rec}}$ is adopted to supervise the segmentation on tens of finely-grained annotated samples of reference image dataset $\mathbb{S}$. The representation consistency loss $\mathcal{L}_{\mathbf{rep}}$ and the self-supervision loss $\mathcal{L}_{\mathbf{sel}}$ are devised for strengthening the category-wise representation consistency and the boundary-aware segmentation consistency on target datasets $\mathbb{S}$. During training, we alternatively optimize the segmentation network $\mathcal{F}_\theta$ and two boundary discriminators $\mathcal{D}_\phi^o, \mathcal{D}_\varphi^i$ using the randomly sampled samples from the reference image dataset $\boldsymbol{S}$, target dataset $\mathbb{S}$ and the open-source dataset $\mathbf{O}$, respectively. For training the segmentation network $\mathcal{F}_\theta$, the total loss $\mathcal{L}_{\mathbf{seg}}$ is calculated as follows:

$$\mathcal{L}_{\mathbf{seg}} = \xi \mathcal{L}_{\mathbf{dic}} + \zeta \mathcal{L}_{\mathbf{rep}} + \eta \mathcal{L}_{\mathbf{sel}} - \mathcal{D}_\phi^o(\boldsymbol{I}_a^o) - \mathcal{D}_\varphi^i(\boldsymbol{I}_a^i). \tag{7}$$

For training the inner and outer discriminators, the Eq.(5) and Eq.(6) are adopted.

4 Experiments

Dataset. The target datasets we adopted contain Cityscapes, SBD, THUR, Bird, Flower, Human. What's more, the open-source datasets (MSRA10K, MSRA-B, CSSD, ECSSD, DUT-OMRON, PASCAL-Context, HKU-IS, SOD, SIP1K) are merged into *MixAll*, which contains multiple categories. We do not use coarse data during training, due to our boundary loss which requires fine boundary annotation.

Type ($\mathbb{S} \leftarrow$ O)	$\mathcal{P}$.(*Cityscapes* $\leftarrow$ *SYNTHIA*)			*Multiple Category* (*Half-category* $\leftarrow$ *Half-category*)						*Single Category* (*One* $\leftarrow$ $MixAll^{-}$)								
Dataset	**Cityscapes**			**SBD**			**THUR**			**Bird**			**Human**			**Flower**		
Method\Index	**MPA**	**MIoU**	**FWIoU**	**MPA**	**MIoU**	**FWIoU**	**MPA**	**MIoU**	**FWIoU**	**MPA**	**MIoU**	**FWIoU**	**MPA**	**MIoU**	**FWIoU**	**MPA**	**MIoU**	**FWIoU**
CAC	–	–	–	–	–	–	-	-	-	48.28	26.42	27.06	47.14	24.28	23.72	36.96	24.24	35.34
ReDO	–	–	–	–	–	–	-	-	-	50.00	38.53	33.01	50.00	35.72	36.49	70.00	58.58	43.82
SG-One	–	–	–	32.61	24.91	47.30	84.11	71.51	69.27	78.66	61.43	74.68	72.46	56.72	56.93	87.05	74.73	76.77
PANet	–	–	–	31.03	20.09	49.78	66.90	55.75	84.39	66.94	57.83	69.42	76.49	60.54	63.29	68.43	69.25	69.94
SPNet	–	–	–	30.10	21.92	39.09	84.11	71.51	69.27	79.65	76.92	78.01	75.90	60.42	62.90	88.43	79.21	80.43
CANet	–	–	–	40.17	32.02	40.82	59.51	50.49	79.02	85.36	76.02	85.01	95.29	90.98	90.99	81.68	70.83	73.96
ALSSS	50.62	41.19	77.53	33.45	23.86	54.95	80.04	60.28	77.63	51.54	39.48	66.09	76.26	60.42	63.90	79.32	85.21	87.32
USSS	51.36	40.68	71.99	55.26	41.07	68.53	84.12	71.51	84.39	49.64	41.27	67.95	77.75	62.40	62.28	95.15	90.81	91.36
Trans.(10)	37.99	30.92	70.22	20.75	14.00	70.22	74.09	62.25	87.73	79.83	65.07	74.71	85.44	75.71	75.72	82.69	78.66	76.39
Trans.(100)	46.65	37.57	71.75	31.67	23.71	71.25	84.37	75.28	91.92	91.56	83.23	84.06	95.00	90.46	90.48	89.29	81.33	89.05
Gated-SCNN	52.89	39.37	71.45	49.44	38.44	84.74	91.11	78.77	90.41	94.90	90.71	96.40	98.82	97.61	97.60	95.61	92.83	93.92
BFP	51.03	36.43	70.20	50.24	42.31	45.20	75.02	77.32	84.25	92.83	87.48	91.45	97.81	96.30	96.41	95.44	92.30	93.81
Unet	52.80	42.98	74.66	73.48	61.45	89.17	91.04	84.53	94.85	91.88	86.41	92.06	97.88	95.86	95.87	96.39	93.44	93.89
FPN	55.10	45.12	74.38	72.78	62.24	89.16	88.78	81.80	93.91	92.86	86.53	92.06	98.19	96.45	96.46	97.16	94.22	94.59
LinkNet	43.99	35.02	73.96	74.33	62.37	89.31	90.74	83.66	94.52	93.04	86.03	91.77	97.41	94.97	94.98	96.82	94.26	94.65
PSPNet	39.76	32.70	69.01	49.23	39.65	82.43	82.22	73.13	90.95	87.01	79.47	87.97	97.02	94.22	94.23	95.77	91.46	91.99
PAN	55.60	45.13	74.17	72.30	60.43	88.74	90.75	81.61	93.76	93.86	87.07	92.38	98.15	96.37	96.38	96.71	93.64	94.06
DeeplabV3+	57.49	46.45	75.19	74.32	63.06	89.38	92.48	84.92	94.94	94.88	89.62	93.95	98.28	96.62	96.63	96.65	93.80	94.23
***R*(0)**	–	–	–	–	–	–	–	–	–	86.02	70.69	85.24	69.95	53.03	53.00	82.88	71.89	75.13
***R*(10)**	53.44	45.01	81.61	75.64	62.10	71.31	88.42	74.84	88.78	90.85	77.14	89.34	95.86	92.06	92.07	95.86	91.02	92.09
***R*(fully)**	58.21	46.91	86.34	79.14	63.07	71.68	92.19	76.84	90.01	94.76	87.27	94.15	97.14	94.44	94.44	97.03	92.05	92.94

Table 1: The performance comparison of different methods. '$\mathbb{S} \leftarrow \mathbf{O}$' denotes knowledge of open-source dataset **O** are translated into the network for target dataset $\mathbb{S}$. '–' denotes that the dateset is not suitable for the method. '$\mathcal{P}$.' denotes the panoramic. '*Half-category*' denotes half category of each dataset. '$MixAll^{-}$' denotes mix category open-source dataset without the target category. 'Trans.(K)' denotes transfer learning method with K labeled samples. Green and Blue indicate the best and second-best performance among all methods. Red indicates the best performance among all non-fully supervised methods. $\boldsymbol{R}(K)$ denotes the Ref-Net with K labeled samples (All scores in %).

Network architecture. In the paper, the segmentation network we adopted is the DeeplabV3+ (backbone: resnet50) [Chen *et al.*, 2017]. Some popular network architectures (Unet [Ronneberger *et al.*, 2015], FPN [Lin *et al.*, 2017], Linknet [Chaurasia and Culurciello, 2017], PSPNet [Zhao *et al.*, 2017], PAN [Li *et al.*, 2018]) are also tested.

Parameter setting. The parameters are set as follows: $\tau = 1, \lambda = 10, \xi = 1, \zeta = 1, \eta = 1$. In the generation of pseudo samples, the disk strel of radius r for the dilation and erosion operation is randomly sampled integer between 11 and 55. The interval iteration number between the segmentation network and discriminators is 5, the batch size is 64, Adam hyperparameters for two discriminators $\alpha = 0.0001, \beta_1 = 0, \beta_2 = 0.9$. The learning rate for the segmentation network and two discriminators are set as $1e^{-4}$.

Metric. The metrics we adopted include Pixel Accuracy (PA), Mean Pixel Accuracy (MPA), Mean Intersection over Union (MIoU), and Frequency Weighted Intersection over Union (FWIoU). Since the Dice index and IoU are positively correlated, the Dice index is omitted.

4.1 Comparing with SOTA Methods

In this section, the Ref-Net is compared with the SOTA methods, including *unsupervised methods* (CAC [Hsu *et al.*, 2018], ReDO [Chen *et al.*, 2019]), *few-shot methods* (SG-One [Zhang *et al.*, 2018], PANet [Wang *et al.*, 2019a], SPNet [Xian *et al.*, 2019], CANet [Zhang *et al.*, 2019]), *weakly-/semi-supervised methods* (USSS [Kalluri *et al.*, 2019], ALSSS [Hung *et al.*, 2018]) and *fully supervised methods* (boundary-aware methods:{ Gated-SCNN [Takikawa *et al.*, 2019], BFP [Ding *et al.*, 2019]}, Unet [Ronneberger *et al.*, 2015], FPN [Lin *et al.*, 2017], LinkNet [Chaurasia and Culurciello, 2017], PSPNet [Zhao *et al.*, 2017], PAN [Li *et al.*, 2018] and DeeplabV3+ [Chen *et al.*, 2017]) on six datasets. For the semi-supervised methods, ten labeled samples are provided. Except for the panoramic, the target dataset and open-source dataset have no overlapped object category. For a fair comparison, the categories of each multiple category dataset (SBD and THUHR) are split into two non-overlapping parts. The fully supervised methods are trained with both two parts. The transfer learning based methods are initially trained on the half-category samples and then trained with specified labeled samples of the rest half-category. Table 1 shows the quantitative results, where we can see that most scores of $\boldsymbol{R}(10)$ achieve the state-of-the-art results on par with existing non-fully supervised methods. Even with more labeled samples, **Trans.**(100) only achieves higher scores than $\boldsymbol{R}(10)$ on THUR and Bird dataset. Moreover, with only 10 labeled samples, the Ref-Net can achieve better results than some fully supervised methods and close results on par with the best fully supervised method. Meanwhile, with fully supervised samples, the Ref-Net achieves higher scores on the complex dataset (Cityscapes and SBD), which demonstrates the advantage of Ref-Net for handling datasets with more categories. Note that the resolution of the Cityscapes dataset we adopted is $128 * 128$. The above two causes lead to the relatively low-scores of all the methods. However, it still validates the superior performance and wide application of the Ref-Net.

4.2 Results on Incremental Category Number

Table 2 gives the IoU scores of Ref-Net and DeepLabV3+ with the incremental category number. For the Ref-Net, the eight categories are set as target dataset, and ten labeled samples are provided for those categories. The rest of the categories are set as the open-source dataset. From Table 2, we

Category	**C-Num.**$=1$	**C-Num.**$=2$	**C-Num.**$=3$	**C-Num.**$=4$	**C-Num.**$=5$	**C-Num.**$=6$	**C-Num.**$=7$	**C-Num.**$=8$	**Average**
bicycle	44.83 / 62.52	42.93 / 56.75	30.85 / 61.01	33.11 / 61.84	35.78 / 59.21	36.52 / 57.25	36.66 / 55.72	37.12 / 51.91	37.23 / 58.28
train	×	45.67 / 68.03	48.07 / 67.93	50.68 / 68.47	49.86 / 69.69	51.35 / 69.52	47.20 / 68.37	53.95 / 66.94	49.54 / 68.42
airplane	×	×	48.67 / 62.21	50.04 / 63.83	50.29 / 65.20	53.13 / 64.79	51.80 / 65.35	49.15 / 63.49	50.51 / 64.15
bird	×	×	×	41.08 / 63.15	39.40 / 63.20	41.74 / 59.58	42.50 / 60.53	41.68 / 61.00	41.28 / 61.49
person	×	×	×	×	32.39 / 72.29	35.89 / 72.62	34.54 / 71.79	38.08 / 70.12	35.22 / 71.71
cat	×	×	×	×	×	56.74 / 79.52	61.58 / 80.58	61.75 / 74.86	60.03 / 78.32
car	×	×	×	×	×	×	39.53 / 71.38	42.06 / 71.27	40.80 / 71.33
dog	×	×	×	×	×	×	×	59.04 / 66.89	59.04 / 66.89

Table 2: The IoU scores of (Ref-Net / DeepLabV3+) with incremental category number. 'C-Num.' denotes the category number of training dataset. '×' denotes the training dataset without the corresponding category (All scores in %).

Index\Ablation	$\boldsymbol{R}^-_{self}$	$\boldsymbol{R}^-_{cond}$	$\boldsymbol{R}^-_{pseu}$	$\boldsymbol{R}^-_{inner}$	$\boldsymbol{R}^-_{outer}$	$\boldsymbol{R}^-_{disc}$	$\boldsymbol{R}^-_{dice}$	$\boldsymbol{R}(1)$	$\boldsymbol{R}(5)$	$\boldsymbol{R}(10)$	$\boldsymbol{R}(20)$	$\boldsymbol{R}(50)$	$\boldsymbol{R}(100)$	$\boldsymbol{R}(fully)$
MPA	88.31	88.01	87.34	87.78	86.96	83.23	65.40	74.39	83.38	88.42	91.82	89.92	92.06	92.19
MIoU	66.12	63.75	70.61	61.20	72.32	66.51	45.08	55.86	65.11	74.84	65.36	70.43	67.90	76.84
FWIoU	88.90	90.35	88.15	87.69	86.31	89.09	82.88	84.49	88.38	88.78	91.58	90.78	91.70	90.01

Table 3: The ablation study of Ref-Net on THUR. $\boldsymbol{R}^-_{self}$, $\boldsymbol{R}^-_{cond}$, $\boldsymbol{R}^-_{pseu}$, $\boldsymbol{R}^-_{inner}$, $\boldsymbol{R}^-_{outer}$, $\boldsymbol{R}^-_{disc}$ and $\boldsymbol{R}^-_{dice}$ denote the Ref-Net without boundary-aware self-supervision, condition images, pseudo triplet, inner boundary discriminator, outer boundary discriminator, two discriminators, and supervised loss (dice loss). $\boldsymbol{R}(K)$ denotes the Ref-Net with K labeled samples.

can see that the IoU scores of *bicycle* increase with the incremental category number when the 'C-Num.' larger than 3. The first two scores (44.83 and 56.75) of the *bicycle* are larger than the score 30.85 ('C-Num.=3'). The reason is that the network trained with a single category will learn more category-aware features and focus on the single category. When the category number increases, the IoU scores first decrease then increase, which indicates that the Ref-Net can learn more general segmentation ability with the incremental category number. For the rest seven categories, most of the IoU scores increase with the incremental category number. In contrast, most IoU scores of DeeplabV3+ decrease with the incremental category number, which verifies the drawback of model-level knowledge translation.

4.3 Ablation Study

To verify each component's effectiveness, we conduct an ablation study on the boundary-aware self-supervision, the pseudo triplet, two boundary discriminators, supervised loss, and the different numbers of labeled samples. In the experiment, the knowledge translation is set as $\{\text{THUR} \leftarrow \text{MixAll}^-\}$. For all $\boldsymbol{R}^-$, ten labeled samples of each category are provided. From Table 3, we can see that $\boldsymbol{R}(10)$ achieves higher scores than others, which verifies the effectiveness of each component. In addition, $\boldsymbol{R}(10)$ achieves about 2% increase on the scores of $\boldsymbol{R}^-_{inner}$ and $\boldsymbol{R}^-_{outer}$ and about 5% increase on the scores of $\boldsymbol{R}^-_{disc}$, which demonstrates that the two-discriminator framework is useful for improving the segmentation performance. For the different numbers of labeled samples, we find that 10-labeled-samples is a critical cut-off point, which can supply relatively sufficient guidance. With more labeled samples, Ref-Net achieves better performance.

5 Conclusion

In this paper, we propose the Ref-Net for segmenting target objects with a reference image as condition, which is inspired by the human recognition mechanism with a reference object. The Ref-Net contains two modules: a Reference Segmentation Module (RSM) and a Boundary Knowledge Translation Module (BKTM). Given a reference object, RSM is trained for finding and segmenting the same category object with tens of finely-grained annotations. Meanwhile, MMD and boundary-aware self-supervision are introduced to constrain the representation consistency of the same category and boundary consistency of the segmentation mask, respectively. Furthermore, in BKTM, two boundary discriminators are devised for distinguishing whether the segmented objects and background contain residual features. BKTM is able to translate the boundary annotation of open-source samples into the segmentation network. The open-source dataset and target dataset can have totally different object categories. Exhaustive experiments demonstrate that the Ref-Net achieves close results on par with fully supervised methods on six datasets. The most important of all, the segmentation performance of all categories will improve with the increasing categories in the target dataset, while existing methods are opposite. The root reason is that the proposed method is based on dataset-level knowledge translation, where the data flow of open-source samples will not pass the segmentation network. It brings a new perspective for the framework design of the image segmentation task.

Acknowledgments

This work is supported by National Natural Science Foundation of China (No.62002318), Zhejiang Provincial Natural Science Foundation of China (LQ21F020003), Key Research and Development Program of Zhejiang Province (2020C01023), Zhejiang Provincial Science and Technology Project for Public Welfare (LGF21F020020), Programs Supported by Ningbo Natural Science Foundation 202003N4318), Major Scientific Research Project of Zhejiang Lab (No.2019KD0AC01), and Zhejiang Lab (No.2020KD0AA06).

References

[Arbelle and Raviv, 2018] Assaf Arbelle and Tammy Riklin Raviv. Microscopy cell segmentation via adversarial neural networks. *ISBI*, 2018.

[Chaurasia and Culurciello, 2017] Abhishek Chaurasia and Eugenio Culurciello. Linknet: Exploiting encoder representations for efficient semantic segmentation. *VCIP*, 2017.

[Chen *et al.*, 2017] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *CoRR*, abs/1706.05587, 2017.

[Chen *et al.*, 2019] Mickael Chen, Thierry Artieres, and Ludovic Denoyer. Unsupervised object segmentation by redrawing. *NeurIPS*, 2019.

[Ding *et al.*, 2019] Henghui Ding, Xudong Jiang, Ai Qun Liu, Nadia Magnenat Thalmann, and Gang Wang. Boundary-aware feature propagation for scene segmentation. *ICCV*, 2019.

[Gao *et al.*, 2019] Naiyu Gao, Yanhu Shan, Yupei Wang, Xin Zhao, Yinan Yu, Ming Yang, and Kaiqi Huang. Ssap: Single-shot instance segmentation with affinity pyramid. *ICCV*, 2019.

[Gulrajani *et al.*, 2017] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. *NeurIPS*, 2017.

[Han and Yin, 2017] Liang Han and Zhaozheng Yin. Transferring microscopy image modalities with conditional generative adversarial networks. *CVPR*, 2017.

[Hayder *et al.*, 2017] Zeeshan Hayder, Xuming He, and Mathieu Salzmann. Boundary-aware instance segmentation. *CVPR*, 2017.

[Hsu *et al.*, 2018] Kuangjui Hsu, Yenyu Lin, and Yungyu Chuang. Co-attention cnns for unsupervised object cosegmentation. *IJCAI*, 2018.

[Hung *et al.*, 2018] Weichih Hung, Yihsuan Tsai, Yanting Liou, and Yenyu Lin. Adversarial learning for semi-supervised semantic segmentation. *BMVC*, 2018.

[Kalluri *et al.*, 2019] Tarun Kalluri, Girish Varma, Manmohan Chandraker, and C V Jawahar. Universal semi-supervised semantic segmentation. *ICCV*, 2019.

[Khoreva *et al.*, 2017] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. *CVPR*, 2017.

[Li *et al.*, 2018] Hanchao Li, Pengfei Xiong, Jie An, and Lingxue Wang. Pyramid attention network for semantic segmentation. *BMVC*, 2018.

[Lin *et al.*, 2017] Tsungyi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. *CVPR*, 2017.

[Liu *et al.*, 2021] Huihui Liu, Yiding Yang, and Xinchao Wang. Overcoming catastrophic forgetting in graph neural networks. *AAAI*, 2021.

[Qin *et al.*, 2019] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. *CVPR*, 2019.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *MICCAI*, 2015.

[Takikawa *et al.*, 2019] Towaki Takikawa, David Acuna, Varun Jampani, and Sanja Fidler. Gated-scnn: Gated shape cnns for semantic segmentation. *CoRR*, abs/1907.05740, 2019.

[Vinyals *et al.*, 2016] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. *NeurIPS*, 2016.

[Wang *et al.*, 2019a] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. *ICCV*, 2019.

[Wang *et al.*, 2019b] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised scale equivariant network for weakly supervised semantic segmentation. *CoRR*, abs/1909.03714, 2019.

[Xian *et al.*, 2019] Yongqin Xian, Subhabrata Choudhury, Yang He, Bernt Schiele, and Zeynep Akata. Semantic projection network for zero- and few-label semantic segmentation. *CVPR*, 2019.

[Yan *et al.*, 2019] Xiaopeng Yan, Ziliang Chen, Anni Xu, Xiaoxi Wang, Xiaodan Liang, and Liang Lin. Meta r-cnn: Towards general solver for instance-level low-shot learning. *ICCV*, 2019.

[Yang *et al.*, 2020a] Yiding Yang, Zunlei Feng, Mingli Song, and Xinchao Wang. Factorizable graph convolutional networks. *NeurIPS*, 2020.

[Yang *et al.*, 2020b] Yiding Yang, Jiayan Qiu, Mingli Song, Dacheng Tao, and Xinchao Wang. Distilling knowledge from graph convolutional networks. *CVPR*, 2020.

[Zhang *et al.*, 2017] Rui Zhang, Sheng Tang, Min Lin, Jintao Li, and Shuicheng Yan. Global-residual and local-boundary refinement networks for rectifying scene parsing predictions. *IJCAI*, 2017.

[Zhang *et al.*, 2018] Xiaolin Zhang, Yunchao Wei, Yi Yang, and Thomas S Huang. Sg-one: Similarity guidance network for one-shot semantic segmentation. *TCYB*, 2018.

[Zhang *et al.*, 2019] Chi Zhang, Guosheng Lin, Fayao Liu, Rui Yao, and Chunhua Shen. Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. *CVPR*, 2019.

[Zhao *et al.*, 2017] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. *CVPR*, 2017.

[Zou *et al.*, 2018] Yang Zou, Zhiding Yu, B V K Vijaya Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. *ECCV*, 2018.