Boundary Knowledge Translation based Reference Semantic Segmentation

Lechao Cheng¹†, Zunlei Feng²†, Xinchao Wang³, Ya Jie Liu¹, Jie Lei⁴, Mingli Song²∗
¹Zhejiang Lab
²Zhejiang University
³National University of Singapore
⁴Zhejiang 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.

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

### 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 \( \mathcal{F}_\theta \) has the same network architecture and parameter as the encoder of the target segmentation branch.

Give a target image \( \boldsymbol{x} \in \mathcal{S} \) and a reference image \( (\boldsymbol{x}_k^c, \boldsymbol{m}_k^c) \in \mathcal{S} \), the extracted representations can be denoted as \( \tilde{r} = \mathcal{F}_\theta^e(\boldsymbol{x}) \) and \( r_k^c = \mathcal{F}_\theta(x_k^c \ast m_k^c) \), where \( \ast \) denotes pixel-wise multiplication. 

With the concatenated representation \( \tilde{r} \) \( r_k^c \), the decoder \( \mathcal{F}_\theta^d \) of the target segmentation branch predicts the mask \( \hat{m} = \mathcal{F}_\theta^d(\tilde{r}, r_k^c) \). For simplicity, the reference segmentation is formulated as \( \hat{m} = \mathcal{F}_\theta(x_k^c \ast m_k^c) \), 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 \((\boldsymbol{x}, \boldsymbol{m}) \in \mathcal{S}\) and a reference image \((\boldsymbol{x}_k^c, \boldsymbol{m}_k^c) \in \mathcal{S}\), the segmented result \( \hat{m} = \mathcal{F}_\theta(\boldsymbol{x}, m_k^c \ast x_k^c) \) is expected to approximate the GT mask \( \boldsymbol{m} \), which can be achieved by minimizing the pixel-wise two-class Dice loss \( \mathcal{L}_{\text{dice}} \):

\[
\mathcal{L}_{\text{dice}} = 1 - \frac{2|\hat{m} \cap \boldsymbol{m}| + \tau}{|\hat{m}| + |\boldsymbol{m}| + \tau}, \tag{1}
\]

where, \( \tau \) 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 (MMD) to constrain the representation distribution consistency. With the reference image \((\boldsymbol{x}_k^c, m_k^c) \in \mathcal{S}\) and the same category target image \( \boldsymbol{x} \), the predicted mask is \( \hat{m} = \mathcal{F}_\theta(\boldsymbol{x}, m_k^c \ast x_k^c) \). The representation consistency loss \( \mathcal{L}_{\text{rep}} \) is defined as follows:

\[
\mathcal{L}_{\text{rep}} = \text{MMD}\{(\hat{f}(\hat{m} \ast \tilde{x})), (\hat{f}(m_k^c \ast x_k^c))\}, \tag{2}
\]

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 \( \hat{m} = \mathcal{F}_\theta(\boldsymbol{A} \boldsymbol{x}, m_k^c \ast x_k^c) \) of the transformed image \( \boldsymbol{A} \boldsymbol{x} \) and
the transformed result $A\mathcal{F}_\theta(x, x'_k \ast m'_k)$ should be consistent in the following way: $\mathcal{F}_\theta(Ax, x'_k \ast m'_k) = A\mathcal{F}_\theta(x, x'_k \ast m'_k)$. Furthermore, we obtain the boundary neighborhood weight map $w'$ and $w$ as follows:

$$
\begin{align*}
  w' &= \mathcal{D}_r(m') - \mathcal{E}_r(m'), \\
  w &= \mathcal{D}_r(m) - \mathcal{E}_r(m),
\end{align*}
$$

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

$$
L_{\text{sel}} = ||w' \ast \mathcal{F}_\theta(Ax, x'_k \ast m'_k) - A\{w \ast \mathcal{F}_\theta(x, x'_k \ast m'_k)\}||^2,
$$

where, $w'$ and $w$ are the weight maps of the predict masks $m'$ and $m$, 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 $O$ into the reference segmentation network.

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

$$
L_{\text{adv}}^{\text{out}} = \frac{1}{2I_{\sim p^o_a}} E_{p^o_a}[D^o_\phi(I'_o)] + \frac{1}{2I_{\sim p^o_c}} E_{p^o_c}[D^o_\phi(I'_o)]
$$

$$
\quad + \lambda E_{I'_{\sim p^c}}[(\|\nabla I' \ast D^o_\phi(I')\| ||2-1)^2],
$$

where the $p^o_a, p^o_c, p^c$ are the segmented outer boundary distribution, pseudo outer boundary distribution, and real outer boundary distribution, respectively. The $p^o_c$ is sampled uniformly along straight lines between pairs of points sampled from the distribution $p^o_a$ and the segmentation network distribution $p^c$. The $I'_o = \epsilon I'_o + (1- \epsilon)I'_o$, where the $\epsilon$ is a random number between 0 and 1. The gradient penalty term is firstly proposed in WGAN-GP [Gulrajani et al., 2017]. The $\lambda$ is the gradient penalty coefficient.

**Inner Boundary Discriminator.** The inner boundary discriminator $D^i_\phi$ is devised for discriminating whether the segmented background contains the inner features of the object. To obtain the segmented background, the predict background mask $m'$ and GT mask $m$ are reprocessed with the Not-operation as follows: $m' = [1] - m$, $m = [1] - m$, where the $[1]$ denotes the unit matrix of $m$’s size. Then, the corresponding fake triplet $I'_i = [x, m', x]$, real triplet $I'_i = [x, m, x]$, and pseudo triplet $I'_i = [x, \mathcal{D}_r(m'), x]$ are computed in the same manner as done in the outer boundary discriminator. The generated pseudo triplet $I'_i$ will also assist the inner boundary discriminator in distinguishing the inner features of objects. Similarly, the inner boundary adversarial loss $L_{\text{adv}}^{\text{in}}$ is defined as follows:

$$
L_{\text{adv}}^{\text{in}} = \frac{1}{2I_{\sim p^i_a}} E_{p^i_a}[D^i_\phi(I'_i)] + \frac{1}{2I_{\sim p^i_c}} E_{p^i_c}[D^i_\phi(I'_i)] - E_{I'_i_{\sim p^c}}[D^i_\phi(I'_i)]
$$

$$
\quad + \lambda E_{I'_{\sim p^c}}[(\|\nabla I' \ast D^i_\phi(I')\| ||2-1)^2],
$$

where, the $p^i_a, p^i_c, p^c$ are the segmented inner boundary distribution, pseudo inner boundary distribution, and real inner boundary distribution. $I'_i = \epsilon I'_i + (1- \epsilon)I'_i$. The optimization on $L_{\text{adv}}^{\text{in}}$ 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 $L_{\text{adv}}^{\text{out}}$ and $L_{\text{adv}}^{\text{in}}$ are used to translate the visual boundary knowledge of the source dataset $O$ into the segmentation network. The basic reconstruction loss $L_{\text{rec}}$ is adopted to supervise the segmentation on tens of finely-grained annotated samples of reference image dataset $S$. The representation consistency loss $L_{\text{rep}}$ and the self-supervision loss $L_{\text{sel}}$ are devised for strengthening the category-wise representation consistency and the boundary-aware segmentation consistency on target datasets $S$. During training, we alternatively optimize the segmentation network $\mathcal{F}_\theta$ and two boundary discriminators $D^o_\phi, D^i_\phi$ using the randomly sampled samples from the reference image dataset $S$, target dataset $S$ and the open-source dataset $O$, respectively. For training the segmentation network $\mathcal{F}_\theta$, the total loss $L_{\text{seg}}$ is calculated as follows:

$$
L_{\text{seg}} = \xi L_{\text{dic}} + \eta L_{\text{rep}} + \eta_{\text{sel}} - D^o_\phi(I'_a) - D^i_\phi(I'_a). \quad (7)
$$

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.
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 panoptic, 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 $R(10)$ 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 $R(10)$ on THUR and Bird dataset. Moreover, with only ten 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 $128 \times 128$. 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


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 %).

| C-Num. | Index | R_{self} | R_{cond} | R_{pseudo} | R_{inner} | R_{outer} | R_{disc} | R_{dice} | R_{(1)} | R_{(5)} | R_{(10)} | R_{(20)} | R_{(50)} | R_{(100)} | R_{(fully)} |
|--------|-------|---------|---------|-----------|-----------|----------|---------|---------|--------|--------|---------|---------|--------|--------|-----------|
| 1      | 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     |
| 2      | MIoU  | 66.12   | 66.75   | 70.61     | 61.20     | 72.32    | 66.51   | 45.08   | 55.86  | 65.11  | 74.84   | 70.43   | 67.90  | 76.84  |           |
| 3      | FWIoU | 88.90   | 90.35   | 88.15     | 87.69     | 86.31    | 89.09   | 82.88   | 84.49  | 88.38  | 91.88   | 90.78   | 91.70  | 90.01  |           |

Table 3: The ablation study of Ref-Net on THUR. R_{self}, R_{cond}, R_{pseudo}, R_{inner}, R_{outer}, R_{disc}, R_{dice}, R_{(1)}, R_{(5)}, R_{(10)}, R_{(20)}, R_{(50)}, R_{(100)}, and R_{(fully)} 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). 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 drawbacks 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 ‘THUR ← MixAll’. For all R_{-}, ten labeled samples of each category are provided. From Table 3, we can see that R_{(10)} achieves higher scores than others, which verifies the effectiveness of each component. In addition, R_{(10)} achieves about 2% increase on the scores of R_{inner} and R_{outer} and about 5% increase on the scores of R_{dice}, 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 (BKT). 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 BKT, two boundary discriminators are devised for distinguishing whether the segmented objects and background contain residual features. BKT 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 (LG21F020020), Ningbo Natural Science Foundation 202003N4318, Major Scientific Research Project of Zhejiang Lab (No.2019KD0AC01), and Zhejiang Lab (No.2020KD0AA06).
