Deep Dual Learning for Semantic Image Segmentation

Ping Luo\textsuperscript{2*}, Guangrun Wang\textsuperscript{1,2*}, Liang Lin\textsuperscript{1,3}, Xiaogang Wang\textsuperscript{2}

\textsuperscript{1}Sun Yat-Sen University \quad \textsuperscript{2}The Chinese University of Hong Kong \quad \textsuperscript{3}SenseTime Group (Limited)

pluo@ie.cuhk.edu.hk wanggrun@mail2.sysu.edu.cn linliang@ieee.org xgwang@ee.cuhk.edu.hk

Abstract

Deep neural networks have advanced many computer vision tasks, because of their compelling capacities to learn from large amount of labeled data. However, their performances are not fully exploited in semantic image segmentation as the scale of training set is limited, where per-pixel labelmaps are expensive to obtain. To reduce labeling efforts, a natural solution is to collect additional images from Internet that are associated with image-level tags. Unlike existing works that treated labelmaps and tags as independent supervisions, we present a novel learning setting, namely dual image segmentation (DIS), which consists of two complementary learning problems that are jointly solved. One predicts labelmaps and tags from images, and the other reconstructs the images using the predicted labelmaps. DIS has three appealing properties. 1) Given an image with tags only, its labelmap can be inferred by leveraging the images and tags as constraints. The estimated labelmaps that capture accurate object classes and boundaries are used as ground truths in training to boost performance. 2) DIS is able to clean tags that have noises. 3) DIS significantly reduces the number of per-pixel annotations in training, while still achieves state-of-the-art performance. Extensive experiments demonstrate the effectiveness of DIS, which outperforms an existing best-performing baseline by 12.6\% on Pascal VOC 2012 test set, without any post-processing such as CRF/MRF smoothing.

1. Introduction

Deep convolutional networks (CNNs) have improved performances of many computer vision tasks, because they have compelling modeling complexity to learn from large number of supervised data. However, their capabilities are not fully explored in the task of semantic image segmentation, which is to assign a semantic label such as ‘person’, ‘table’, and ‘cat’ to each pixel in an image. This is due to the training set of semantic image segmentation is limited, where the per-pixel labelmaps are difficult and expensive to obtain.

To reduce efforts of data annotations, a usual way is to automatically collect images from Internet, which are associated with image-level tags. As a result, the entire dataset contains two parts, including a small number of fully annotated images with per-pixel labelmaps and a large number of weakly annotated images with image-level tags. Learning from this dataset follows a semi-supervised scenario. To disclose the differences among existing works, we introduce necessary notations that will be used throughout this paper. Let $I$ be an image and $L, T$ denote an image, a per-pixel labelmap, and a vector of image-level tags respectively, where the labelmap $L$ can be missing in training. (a) treats $L$ as missing label in multitask learning, where its gradient is not computed in back-propagation (BP), (b) regards $L$ as latent variable that can be inferred by tags and used as ground truth in BP. We propose (c), which infers the missing label $L$ not only by recovering clean tags $T$, but also by reconstructing the image to capture accurate object shape and boundary.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure1.png}
\caption{Comparisons of recent semi-supervised learning settings. $I$, $L$, and $T$ denote an image, a per-pixel labelmap, and a vector of image-level tags respectively, where the labelmap $L$ can be missing in training. (a) treats $L$ as missing label in multitask learning, where its gradient is not computed in back-propagation (BP), (b) regards $L$ as latent variable that can be inferred by tags and used as ground truth in BP. We propose (c), which infers the missing label $L$ not only by recovering clean tags $T$, but also by reconstructing the image to capture accurate object shape and boundary.}
\end{figure}

Let $\theta$ be a set of parameters of a CNN. The objective function of semi-supervised segmentation is often formulated as maximizing a log-likelihood with respect to $\theta$. We have $\theta^* = \arg\max_{\theta} \log p(L^I, L^w, T^I, T^w | I^I, I^w; \theta)$, where the probability measures the similarity between the ground truths and the predicted labelmaps and tags pro-
duced by the CNN.

Previous works can be divided into two categories according to different factorizations of the above probability. In the first category, methods [13, 23, 28] employed multitask learning with missing labels, where the labelmap $L^w$ is missing and treated as an unobserved variable. In this case, the likelihood is factorized into two terms, $p(L^f, T^f, T^w|I^f, I^w) \propto p(L^f|I^f) \cdot p(T^f, T^w|I^f, I^w)$, which correspond to two tasks as shown in Fig.1 (a). One learns to predict labelmaps $L^f$ using the fully annotated images $I^f$, and the other is trained to classify tags, $T^f$ and $T^w$, using both fully and weakly annotated images $I^f$ and $I^w$. These two tasks are jointly optimized in a single CNN. More specific, back-propagation of the first task is not performed when a weakly labeled image is presented. In general, these approaches learn shared representation to improve segmentation accuracy by directly combining data with strong and weak supervisions.

Unlike the above approaches, $L^w$ is treated as a latent variable in the second category [21, 4, 15]. In this case, the likelihood function can be represented by $p(L^f, L^w, T^f, T^w|I^f, I^w) \propto p(L^f|I^f) \cdot p(L^w|I^w, T^w) \cdot p(T^f, T^w|I^f, I^w)$, where the first and the third terms are identical as above. The second term estimates the missing labelmap given an image and its tags. In other words, when a weakly labeled image $I^w$ is presented, the CNN produces a labelmap $L^w$, as its parameters are learned to do so from the fully annotated images. The predicted labelmap is then refined with respect to the tags $T^w$, as shown in Fig.1 (b). We take Pascal VOC12 [6] as an example, which has 20 object classes and 1 class of background. A labelmap $L^w$ is of $n \times m \times 21$, where $n, m$ denote width and height of the response map, and each entry in $L^w$ indicates the possibility of the presence of a class on a pixel. For instance, when $L^w$ is confused with ‘cat’ and ‘dog’ by assigning them similar probabilities, and $T^w$ tells that only ‘dog’ is appeared in the image but not ‘cat’, we can refine $L^w$ by decreasing the probability of ‘cat’. An implementation to achieve this is by convolving the $n \times m \times 21$ labelmap with a $1 \times 1 \times 21$ kernel (a vector of tags). After refinement, $L^w$ is used as ground truth in back-propagation, which significantly boosts performance as demonstrated in [21].

However, the recent approaches still have two weaknesses. (i) The tags $T^w$ helps refine the misclassified pixels of $L^w$, but not the object boundary and shape, because these information are not captured in image-level tags. (ii) $T^w$ may have noises when the images are automatically downloaded from the Internet. These noisy tags will hamper the learning procedure.

To resolve the above issues, this work presents Dual Image Segmentation (DIS), which is inspired by the dual learning in machine translation [9], where exists a small number of bilingual sentence pairs, such as a pair of ‘have a good day’ in English (En) and ‘bonne journée’ in French (Fr), but there exists unlimited monolingual sentences in En and Fr that are not labeled as pairs on the Internet. [9] leveraged the unlabeled data to improve performance of machine translation from En to Fr, denoted as En→Fr. This is achieved by designing two translation models, including a model of En→Fr and a reverse model of Fr→En.

In particular, when a pair of bilingual sentences is presented, both models can be updated in a fully-supervised scenario. However, when a monolingual En sentence is presented, a Fr sentence is estimated by first applying En→Fr and then Fr→En’. If the input sentence En and the reproduced sentence En’ are close, the estimated Fr sentence is likely to be a good translation of En. Thus, they can be used as a bilingual pair in training. Training models on these large number of synthesized pairs, performance can be greatly improved as shown in [9]. In general, the above two models behave as a loop to estimate the missing Fr sentences.

In this work, DIS extends [9] to four tuples, $I, L, T$, and $\hat{T}$, where $\hat{T}$ denotes the clean tags that are recovered from the noisy tags $T$. As shown in Fig.1 (c), DIS has three “translation” models as a loop, including 1) $I \rightarrow (L, T)$, 2) $\hat{T} \rightarrow (L, T)$, and 3) $L \rightarrow \hat{T}$, where the first one predicts the labelmap $L$ and clean tags $\hat{T}$ given an image $I$, the second one refines $L$ according to $\hat{T}$, and the last one reconstructs $I$ using the refined labelmap $L$.

Intuitively, DIS treats both $L^w$ and $\hat{T}^w$ as latent variables, other than only one latent variable $L^w$ as in Fig.1 (b). For example, when a weakly labeled image $I^w$ is presented, we can estimate $L^w$ and $\hat{T}^w$ by performing $I^w \rightarrow (L^w, \hat{T}^w), T^w \rightarrow (L^w, T^w)$, and $L^w \rightarrow \hat{T}^w$. As a result, when $I^w$ and $\hat{T}^w$ are close, $L^w$ and $T^w$ can be used as a strong and a weak supervisions for training, because they not only capture precise object classes that present in image, but also capture accurate object boundary and shape in order to reconstruct the image. Leveraging these synthesized ground truths in training, DIS is able to substantially reduce the number of fully annotated images, while still attaining state-of-the-art performance.

Different from Fig.1 (a,b), DIS iteratively optimizes two learning problems as visualized in (c). Its objective function contains two parts, $\log p(L^f, L^w, T^f, T^w, \hat{T}^w|I^f, I^w) + \log p(I^f, I^w|L^f, L^w)$, where the first part is for segmentation and the second part is for image reconstruction. Compared to (b), the first part has one additional variable $\hat{T}^w$ and is factorized as $p(L^f, L^w, T^f, T^w, \hat{T}^w|I^f, I^w) \propto p(L^f|I^f) \cdot p(L^w|I^w, T^w) \cdot p(T^f, T^w|I^f, I^w) \cdot p(\hat{T}^w|I^w)$. The first three terms are similar to those in (b), whilst the fourth term estimates the clean tags. The second part of the objective function reconstructs images using the predicted labelmaps.

In general, we propose a novel framework for semantic image segmentation, namely dual image segmentation.
Table 1: The numbers of training samples of recent weakly-, semi-, and fully-supervised segmentation methods are compared from top to bottom. We take VOC12 as an example to compare their experimental settings. ‘Pixel’, ‘Tag’, ‘Bbox’, and ‘Scribble’ indicate different types of supervisions, including per-pixel labelmaps, image-level tags, bounding boxes, and scribbles. ‘V’, ‘VA’, ‘I’, ‘V7’, and ‘C’ represent images are come from different data sources, where V=VOC12 [6], VA=VOC12+VOC Augment [8], I=ImageNet [25], V7=VOC07 [5], and C=COCO [16]. All these data are manually labeled, but not the data from ‘W’, which are collected from the Internet.

<table>
<thead>
<tr>
<th>Method</th>
<th>Pixel</th>
<th>Tag</th>
<th>Bbox</th>
<th>Scribble</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIL-FCN [23]</td>
<td>12k, VA</td>
<td>76k, I</td>
<td>10k, VA</td>
<td></td>
</tr>
<tr>
<td>MIL-ppx1 [20]</td>
<td>10k, VA</td>
<td>76k, I</td>
<td>10k, VA</td>
<td></td>
</tr>
<tr>
<td>CFCN [22]</td>
<td>10k, VA</td>
<td>10k, VA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WSSL (semi) [21]</td>
<td>15k, VA+I</td>
<td>118k, C</td>
<td>133k, C</td>
<td>10k, VA</td>
</tr>
<tr>
<td>BoxSup [4]</td>
<td>10k, VA</td>
<td>10k, VA</td>
<td>10k, VA</td>
<td></td>
</tr>
<tr>
<td>ScribbleSup [13]</td>
<td>12k, VA</td>
<td>12k, VA</td>
<td>12k, VA</td>
<td></td>
</tr>
<tr>
<td>DIS (ours)</td>
<td>2.9k, V</td>
<td>40k, W+10k, VA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WSSL (full) [21]</td>
<td>133k, C</td>
<td>118k, C</td>
<td>10k, VA</td>
<td></td>
</tr>
<tr>
<td>DeepLabv2-CRF [3]</td>
<td>12k, VA</td>
<td>12k, VA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CentraleSupelec [2]</td>
<td>12k, VA</td>
<td>12k, VA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DPN [17]</td>
<td>12k, VA</td>
<td>12k, VA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RNN [16]</td>
<td>12k, VA</td>
<td>12k, VA</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(DIS), which significantly reduces number of fully annotated images in training, while still achieves state-of-the-art performance. For example, we demonstrate the effectiveness of DIS on Pascal VOC 2012 (VOC12) test set, outperforming a best-performing baseline by more than 12%. When adapting to VOC12, DIS reduces the number of fully annotated images by more than 75%.

1.1. Related Works

Semantic Image Segmentation. We review recent advanced segmentation methods by dividing them into three groups, including weakly-, semi-, and fully-supervised methods, which are listed in Table 1 from top to bottom respectively. We take VOC12 as an example to compare their experimental settings. As an approach may have multiple settings, we choose the best-performing one for each approach. We can see that DIS reduces number of labelmaps by 76% and 97% compared to the fully-supervised methods [3, 2, 17, 30, 18] and [21] respectively. In comparison with the weakly- and semi-supervised methods, the number of images of DIS is also smaller than those of many previous works [24, 21, 4]. Unlike existing works that employed manually labeled data, the image-level tags in DIS are mainly collected from the Internet without manual cleaning and refinement.

Image/Labelmap Generation. Transforming between a labelmap and an image has been explored in the literature [11, 1, 29]. Table 2 compares these models with DIS, where [11] aimed at generating realistic images given labelmaps, whilst [1, 29] produced labelmaps using encoder-decoder (AE) and variational autoencoder (VAE). All these approaches treated the feature coding as latent representation to improve the quality of image or labelmap generation in a fully-supervised scenario, where large number of fully annotated images is required. In contrast, DIS infers the latent labelmaps and clean tags to improve segmentation accuracy in a semi-supervised scenario.

2. Our Approach

Weak Supervisions. To improve segmentation accuracy using weak supervisions, we employ the Image Description in the Wild (IDW) dataset [26], which has 40 thousand images selected from the Internet. These images are associated with image-level tags, object interactions, and image descriptions. In this work, we only leverage the image-level tags, but not object interactions. To improve image segmentation using object interactions and descriptions, we refer the readers to [26]. Some images and descriptions of IDW are given in Fig.2.
An appealing property of IDW is that each image can be associated with multiple tags by parsing the sentences. A deficiency is that these sentences are not sufficiently accurate, where important tags are missing or not presented in images as highlighted in orange and green in Fig.2. For instance, as shown in the top-right image, ‘birds’ are too small to be observed in image and a ‘bicycle’ is missing in the sentence. Leveraging them as supervisions may hinder the training procedure. Nevertheless, DIS is able to recover the missing labelmaps and clean tags, to boost the segmentation performance.

2.1. Network Overview

Fig.3 illustrates the diagram of DIS, which has four important components, including a ResNet101 for feature extraction, and three subnets denoted as ‘subnet-1,-2,-3’ for labelmap prediction, image reconstruction, and tag classification respectively. Best viewed in color.

Figure 3: Pipeline of DIS, which contains four key components, including a ResNet101 for feature extraction, and three subnets denoted as ‘subnet-1,-2,-3’ for labelmap prediction, image reconstruction, and tag classification respectively. Best viewed in color.

More specific, given an image $I$ in test, the objective function of inference can be written as $u_1^t, v_1^t = \arg\min_{u_1, v_1} \| z_3 - I \|_2^2$. Let $t$ be the iteration of optimization. For example, $u_1^t$ denotes a variable at the $t$-th iteration. As shown in Fig.3, when $t = 0$, $u_1^0$ and $v_1^0$ are the features extracted by ResNet101 from the input image $I$ at the beginning of inference. When $t > 0$, these features are forwarded to $z_3^t$, which can be considered as a function taken $u_1^t, v_1^t$ as inputs. Therefore, by freezing all the network parameters, we can refine these features by treating $u_1^t, v_1^t$ as variables and back-propagating gradients of the above objective function to them. After $t$ iterations,
**3. Training Algorithms**

DIS is trained with two stages. The first stage pretrains the network using the fully annotated images only. In the second stage, the network is finetuned using both the fully and weakly annotated images.

**Fully-supervised Stage.** In the first stage, given the strongly labeled data \{I^f, L^f, T^f\}, DIS is pretrained in a fully-supervised manner with three loss functions, which are indicated by the solid red arrows at the end of the three subnets as shown in Fig.3, including a softmax loss for labelmap prediction, \(L_{\text{map}}(u_3, L^f)\), an euclidian loss for image reconstruction, \(L_{\text{img}}(z_3, I^f)\), and a softmax loss for tag classification, \(L_{\text{tag}}(v_3, T^f)\). For instance, \(L_{\text{map}}(u_3, L^f)\) denotes the loss function of labelmap prediction, where \(u_3\) and \(L^f\) are the predicted and ground-truth labelmaps respectively. More specific, the four components of DIS is progressively trained in three steps. First, we train three components, including ResNet101, subnet-1, and -3 to predict labelmaps and tags. Second, subnet-2 is learned to reconstruct images by freezing the parameters of the above components. Finally, all four components are jointly updated. The rationale behind the above multi-step training scheme is that it encourages a smooth course of multitask minimization. Similar idea has been demonstrated in many previous works [12, 10].

**3.1. Semi-supervised Stage**

In the second stage, DIS is finetuned by two sources of data, the weakly and fully annotated data, to improve segmentation performance. They are represented by \(\{I, L, T\}\), where \(I = \{I^f, I^w\}\), \(L = \{L^f, L^w\}\), and \(T = \{T^f, T^w, T^w\}\). The missing labelmap \(L^w\) and clean tags \(T^w\) are inferred as ground truths for training. This is another main contribution of DIS. We first explain the objective function and then discuss the training steps.

Let \(\theta\) be a set of parameters of the entire network. Then the learning problem can be formulated as

\[
\arg \min_{\theta} \left\{ L_{\text{map}}(u_3, L) + L_{\text{img}}(z_3, I) + L_{\text{tag}}(v_3, T) \right\} + \arg \min_{u_1, v_1} \left\{ L_{\text{img}}(z_3, I^w) + L_{\text{tag}}(v_3, T^w) \right\}, \tag{1}
\]

which contains two parts that are optimized iteratively. The first part is identical to the fully-supervised stage above, where each input image \(I\), regardless of which data source it comes from, is trained to update \(\theta\) by minimizing the losses between its predictions \(\{u_3, z_3, v_3\}\) and ground truths \(\{I, L, T\}\). However, for an weakly annotated image \(I^w\), its labelmap \(L^w\) and clean tags \(T^w\) are unobserved. Thus, they are estimated by minimizing the second part of Eqn.(1).

In general, given an image \(I^f\) with strong supervision, DIS can be simply finetuned by optimizing the first part of Eqn.(1). Otherwise, when \(I^w\) is presented, DIS is finetuned by iteratively minimizing both parts of Eqn.(1) following two steps, (i) inferring \(L^w\) and \(T^w\) by freezing \(\theta\) and updating \(u_1\) and \(v_1\), and (ii) updating the network parameters \(\theta\) by using the inferred \(L^w\) and \(T^w\) as ground truths.

**Inference in Training.** We introduce the first step in detail. The inference in training is similar to that in test. The main difference is that \(u_3\) and \(v_1\) are updated by minimizing the losses of both image reconstruction and tag classification, other than only the image reconstruction as we did in test. In this case, \(u_3\) and \(v_1\) receive gradients from two paths as shown in Fig.3. One is from \(v_3\) and the other is from \(u_3\) and \(z_3\). As the subnets have been pretrained on strongly annotated data, iteratively updating \(u_3\) and \(v_1\) makes them captured the accurate spatial representation and clean tags, because noises in them are removed in order to produce the labelmaps and reconstruct the images.

Let \(t\) be the iteration of optimization. Then \(u_3^0\) and \(v_1^0\) are the outputs of ResNet101 when \(t = 0\) at the beginning of inference. For \(t > 0\), we update \(u_3^t\) and \(v_1^t\) by propagating the gradients of the two loss functions back to them, while keeping the network parameters fixed. After \(t\) iterations, we obtain \(u_3^*\) and \(v_1^*\), which are forwarded to \(u_3^*\) and \(v_3^*\). The final prediction of \(u_3^*\) is refined by \(v_3^*\) using convolution following the inference in test, as shown in blue dashed arrows of Fig.3. Then the ground truth labelmap and clean tags can be achieved by applying the softmax function on...
Implementation Details. For a fair comparison, the parameters of ResNet101 are initialized the same as [3] by training on ImageNet and COCO. When adapting to VOC12, our baseline model is trained on 2.9k fully supervised data, including images and their labelmaps. The weakly-supervised data with image-level tags are chosen from IDW and VOC augment datasets. The parameters of three subnets are initialized by sampling from a normal distribution. The entire network is trained by using stochastic gradient descent with momentum.

An Interesting Finding. The number of iteration $t$ in inference of training and test can be different. In general, more iterations are performed in training, less iterations are required in test, and vice versa. In other words, computation time in test can be simply reduced by increasing inference iterations in training. On the contrary, training time can be reduced by growing number of iterations in test. We find that both strategies provide remarkable segmentation accuracies. More details are presented in experiments.

4. Experiments

We evaluate the effectiveness of DIS in three aspects. In section 4.1, DIS is compared with existing segmentation methods on VOC12 test set [6]. When adapting to VOC12, DIS is trained on 2.9k pixel-level annotations and 50k image-level tags. In particular, 40k tags are from IDW and the other 10k tags are from VOC augment dataset. In contrast, previous works typically combined VOC12 and VOC augment dataset [8], resulting in 12k pixel-level annotations. For initialization, existing methods initialized their networks by pretraining on ImageNet [25] and COCO [16], which typically brings 2~3% performance gain. For a fair comparison, we mainly report and compare to results that were pretrained on the above two datasets.

In section 4.2, we study the impact of the number of iterations in inference, for both the training and testing stages. In section 4.3, we examine the generalization of DIS on a test set of IDW.

4.1. Comparisons between Previous Works

The segmentation accuracy of DIS is compared to those of the state-of-the-art methods on VOC12 test set. We adopt 11 representative fully-supervised methods, including SegNet [1], FCN [19], Zoom-out [20], WSSL (full) [21], RNN [30], Piecewise [14], DPN [17], DeepLabv2 [3], LRR-4x-Res [7], HP [27], and CentraleSupelec [2]. We also employ two best-performing semi-supervised approaches, WSSL (semi) [21] and BoxSup [4].

Results are reported in the upper three blocks of Table 3, where the superscript $^t$ indicates methods whose baseline models are pretrained on both ImageNet and COCO. We can see that the recent fully-supervised methods generally achieved better results than those of the semi-supervised methods, such as WSSL (semi) and BoxSup. However, when training WSSL on additional images from IDW, it obtains performance of 81.9% that outperforms all previous works. A new record of 86.8% is achieved by DIS, which significantly outperforms the baseline and WSSL+IDW by 12.6% and 4.9% respectively, and reduces the number of fully annotated images by 75% and 97% compared to them.

Properties of several representative works are compared in Table 4, in terms of number of network parameters, training on manually labeled data, CRF post-processing, multiscale testing, and runtime per image on a Titan-X GPU. We have four main observations. First, fusing multiscale features to improve performance is a common practice, which typically leads to 1~2% gain [17, 4]. Second, all approaches except DIS are trained on manually labeled data, whilst weakly-supervised data for DIS are automatically collected from Web. Third, methods that have smaller numbers of parameters execute faster, but sometimes sacrifice performance such as [1]. Fourth, most of computation time in existing methods such as [3, 17] is occupied by CRF post-processing. Note that the runtime of CRF is not counted in Table 4.

4.2. Ablation Study

We study the effect of using different iterations in inference, as shown at the bottom of Table 3, where $t_{tr}$ and $t_{ts}$ represent numbers of inference in training and test respectively. We evaluate 10 different setups. In the first setup when $t_{tr} = 0$, DIS degenerates to the recent semi-supervised learning scheme without refining labelmaps and tags. We can see that it outperforms the baseline by 5.8%. For the remaining setups when $t_{tr} = 5, 10, $ and 30, DIS attains more than 5% gain compared to the first one.

We have several important observations. First, when $t_{ts} = 0$, setup id ‘2’, ‘5’, and ‘8’ outperform ‘1’ and the baseline by more than 5% and 10% respectively, demonstrating the importance of inference in training. Second, when the numbers of $t_{tr}$ are the same, increasing $t_{ts}$ improves performances. For example, the seventh setup outperforms the fifth one by 0.9%, showing the usefulness of inference in test. Third, when the numbers of $t_{ts}$ are the same, larger $t_{tr}$ obtains better performances. For instance, setup id ‘9’ has 1.1% improvement over ‘3’. Finally, the best performance of 86.8% is achieved when $t_{tr} = 30$ and $t_{ts} = 30$.

Fig. 5 visualizes some segmentation results. As our purpose is not to generate high quality images, DIS is supervised by the downsampled images of $45 \times 45$ in order to save computations. These images are able to catch sufficient details of object shapes and boundaries as shown in the second column. In general, the predicted labelmaps become more accurate when more inferences are performed. Note
that the original images are put in the first column for better visualization.

In Table 3, we also see that the results of ‘tstr = 30, ttr = 0’, ‘tstr = 5, ttr = 30’, and ‘tstr = 10, ttr = 30’ are comparable, which are 86.3, 86.2, and 86.4 respectively. We find that this is an useful feature of DIS. In particular, the first one indicates we can reduce runtime in test by increasing iterations of inference in training, when computational cost is a priority. To the extreme, inference is not performed in test when ttr = 0, while still maintaining high performance. The last two tell us when model deployment is urgent, we can reduce iterations in training and increase those in test, such as tstr = 30.

Another interesting finding from Table 3 is that different iterations in inference induces diversity among these models, as disclosed by the top performance of each class in bold. In other words, accuracy can be further boosted by ensembling models with different iterations in inference.

### 4.3. Weakly-supervised Segmentation

We examine the generalization of DIS on IDW test set, which consists of one thousand manually labeled images. This test set is challenging, because the number of object categories per image is more than those in existing datasets, i.e., 2.23 compared to 1.48 in VOC12 and 1.83 in COCO. We evaluate 7 setups of DIS similar as above, and compare them to the baseline ResNet101 and WSSL+IDW. The results are reported in Table 5. In general, since only image-level tags are available in training, the performances in IDW test set are much lower than those in VOC12 test set. However, similar trends can be observed from these two datasets, representing the effectiveness of DIS. It outperforms the baseline and WSSL+IDW by 9.2% and 7.8% respectively, when tstr = 30 and ttr = 30.

### 5. Conclusion

This work presented a novel learning setting for semi-supervised semantic image segmentation, namely dual image segmentation (DIS). DIS is inspired by the dual learning in machine translation [9], but has two uniqueness. First, DIS extends two tuples of En and Fr translation in [9] to multiple tuples in the task of semantic image segmentation, by modeling a close loop of generation among images, per-pixel labelmaps, and image-level tags. Second, different ‘translation’ models of DIS can be plugged into a single
Different from existing semi-supervised segmentation methods, where a missing labelmap is treated as an unobserved variable, or as a latent variable but inferred only based on the tags, DIS estimates the missing labelmaps by not only satisfying the tags, but also reconstructing the images. These two problems are iteratively solved, making the inferred labelmaps captured the accurate object classes that present in the images, as well as the accurate object shapes and boundaries so as to generate the images. Extensive experiments demonstrated the effectiveness of DIS on VOC12 and IDW test sets, where it establishes new records of 86.8% and 59.8% respectively, outperforming the baseline by 12.6% and 9.2% and reducing number of fully annotated images by more than 75%.

**Acknowledgement.** This work is supported in part by SenseTime Group Limited, the Hong Kong Innovation and Technology Support Programme, and the National Natural Science Foundation of China (61503366, 91320101, 61472410, 61622214). Correspondence to: Ping Luo and Liang Lin.

---

**References**


