Transferable Semi-supervised Semantic Segmentation

Huaxin Xiao[1,2], Yunchao Wei[3], Yu Liu[1], Maojun Zhang[1], Jiashi Feng[2]
[1]Department of System Engineering, National University of Defense Technology
[2]Department of ECE, National University of Singapore
[3]Beckman Institute, University of Illinois at Urbana-Champaign
{xiaohuaxin, jasonyuliu, mjzhang}@nudt.edu.cn, wychao1987@gmail.com, elefjia@nus.edu.sg

Abstract

The performance of deep learning based semantic segmentation models heavily depends on sufficient data with careful annotations. However, even the largest public datasets only provide samples with pixel-level annotations for rather limited semantic categories. Such data scarcity critically limits scalability and applicability of semantic segmentation models in real applications. In this paper, we propose a novel transferable semi-supervised semantic segmentation model that can transfer the learned segmentation knowledge from a few strong categories with pixel-level annotations to unseen weak categories with only image-level annotations, significantly broadening the applicable territory of deep segmentation models. In particular, the proposed model consists of two complementary and learnable components: a Label transfer Network (L-Net) and a Prediction transfer Network (P-Net). The L-Net learns to transfer the segmentation knowledge from strong categories to the images in the weak categories and produces coarse pixel-level semantic maps, by effectively exploiting the similar appearance shared across categories. Meanwhile, the P-Net tailors the transferred knowledge through a carefully designed adversarial learning strategy and produces refined segmentation results with better details. Integrating the L-Net and P-Net achieves 96.5% and 89.4% performance of the fully-supervised baseline using 50% and 0% categories with pixel-level annotations respectively on PASCAL VOC 2012. With such a novel transfer mechanism, our proposed model is easily generalizable to a variety of new categories, only requiring image-level annotations, and offers appealing scalability in real applications.

Introduction

Fully-supervised deep learning algorithms for semantic segmentation (Long, Shelhamer, and Darrell 2015; Chen et al. 2015; Pan et al. 2017) generally demand a large amount of high-quality pixel-level annotation. However, such annotation is only available for a small number of categories to date, e.g., 20 categories in PASCAL VOC 2012 (Everingham et al. 2014) and 80 categories in MS-COCO (Lin et al. 2014). Scarcity of annotated data severely limits the deployment of advanced segmentation models in real applications. The semi-supervised learning based semantic segmentation models are developed to provide an alternative for comparable segmentation quality with less annotation cost.

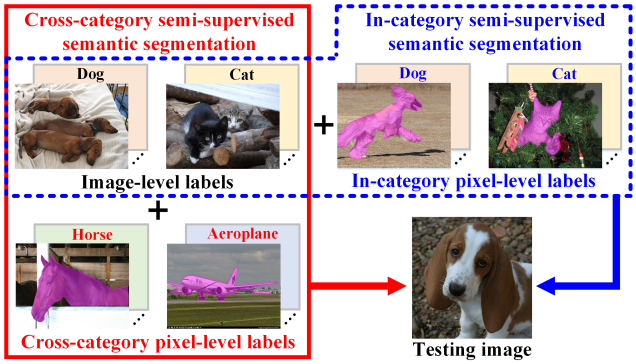

Figure 1: Illustration on two different settings for semi-supervised image semantic segmentation: conventional In-category Semi-supervised Semantic Segmentation (I3S) and the novel Cross-category Semi-supervised Semantic Segmentation (C3S) (which we consider in this work). Different from the I3S problem where each category (e.g., `dog`) has a few in-category pixel-level annotations as well as considerable image-level labels, we introduce a more general and realistic C3S problem where some categories (e.g., `horse` and `aeroplane`) have pixel-level annotations and some other categories to segment (e.g., `dog` and `cat`) only have image-level labels. The C3S problem is more challenging and requires the segmentation model to have a strong transferable learning ability. Best viewed in color.

In the setting of conventional semi-supervised semantic segmentation (Papandreou et al. 2015; Hong, Noh, and Han 2015), namely the In-category Semi-supervised Semantic Segmentation (I3S) as illustrated in the top panel of Figure 1, each category in the training set must be provided with a few pixel-level annotations as well as considerable image-level annotations. This setting however deviates from real applications because a new introduced category still requires extra efforts on re-labeling the category samples. This scheme thus becomes impractical for dealing with hundreds of thousands of categories. For instance, over 20,000 categories are included in ImageNet (Russakovsky et al. 2015) and people can recognize much more categories.

To mitigate such a gap and essentially enhance scalability and applicability of segmentation models, in this work we introduce a more general learning scheme of semi-supervised semantic segmentation, i.e., the Cross-category Semi-supervised Semantic Segmentation (C3S), as illustrated in the left panel of Figure 1. Within C3S scheme, different categories have supervision at different levels, or more concretely some categories have pixel-level annotations (called "strong" categories) and some only have class labels (called "weak" categories). More importantly, there is no overlap between the strong and weak categories.

To solve C3S induced problems, the key point lies in how to effectively learn and transfer re-usable knowledge from strong categories to the segmentation of weak categories. To this end, we develop a novel transferable semi-supervised semantic segmentation model. It contains two complementary components, i.e., a Label transfer Network (L-Net) and a Prediction transfer Network (P-Net), to transfer and adapt the learned segmentation knowledge from strong categories to the weak ones. More concretely, the L-Net learns the segmentation knowledge from strong categories explicitly at first and then transfers the knowledge to produce pixel-level but coarse annotations for the images from weak categories. Upon the coarse annotations, the P-Net conducts another knowledge transfer by learning implicit structural fitting patterns between the predicted and manually annotated segmentations in the strong categories to refine the prediction of the weak categories.

In practice, we notice that segmentation knowledge can be transferred more easily among the categories sharing similar appearances, e.g., from `bicycle` to `motorcycle`. Based on this intuitive yet important observation, we devise following learning scheme for the L-Net: we first familiarize the L-Net with the segmentation knowledge learned from strong categories and utilize the knowledge to predict class-agnostic segmentation maps of weak categories with similar appearance but only image-level labels. Conditioned on the segmentation maps by L-Net, a self-diffusion algorithm on the localized semantic seeds is employed to produce the pixel-level annotations of images from weak categories.

The P-Net learns to transfer the verifiable segmentation structural patterns from strong categories and refine segmentations via adversarial training (Goodfellow et al. 2014). Concretely, the P-Net is trained on the strong categories to implicitly learn the fitting patterns between the predicted segmentation map and the raw images, taking ground truth as the adversarial reference. Such knowledge is class-agnostic and well transferable from strong to weak categories. P-Net cannot only tune the prediction to approach the ground truth but also refine details to reduce discrepancies introduced by the inaccurate annotations from L-Net.

We conduct experiments on the PASCAL VOC 2012 dataset, and in case of only 50% (30%) categories with pixel-level annotations, our proposed model achieves 96.5% (91.4%) performance of the fully-supervised baseline. Moreover, we conduct a cross-dataset C3S experiment on transferring the knowledge from completely new categories in MS-COCO to PASCAL VOC 2012 where only image-level labels are available. The proposed model can still retain 89.4% performance of the fully-supervised baseline. Benefiting from the transferable segmentation knowledge from L-Net and the tailored prediction by P-Net, the proposed model can easily produce high-quality pixel-wise masks for a large number of categories, which undoubtedly broadens image semantic segmentation applications in practice.

Related Work

To relieve the high demand of pixel-level annotation in semantic segmentation, weakly- and semi-supervised learning approaches have attracted much attention. For weakly-supervised approaches, the image-level label is the simplest way to collect and label. To learn a promising model only with image-level annotations, Kolesnikov and Lampert (2016) defined three loss functions to constrain the model from coarse seeds to fine boundary. Saleh et al. (2016) extracted the activations from higher-level layers as initial segmentation masks. Kwak, Hong, and Han (2017) utilized superpixels of the input image as a pooling layout to learn and infer semantic segmentation. Wei et al. (2017) progressively mined semantic regions from classification activations to prevent the network from focusing on a small part of an object. Due to the limited information provided by image-level labels, Wei et al. (2016) employed saliency maps from extra simple images to provide annotations for learning the semantic segmentation model. Similar to the setting of C3S problem, Hong et al. (2016) pre-trained an attention model with irrelevant pixel-level annotations for transferring the segmentation knowledge to the weakly labeled targets. Recently, Hong et al. (2017) generated segmentation labels automatically from the web-crawled videos as strong supervision for weakly-supervised semantic segmentation.

Semi-supervised semantic segmentation gives a trade-off between decent performance and labeling efficiency. Papandreou et al. (2015) inferred the segmentation model by bundling a fixed proportion of strongly/weakly annotated images in one mini-batch with the expectation maximization methods. Hong, Noh, and Han (2015) separately learned classification and segmentation networks which correspond to different annotations, and transferred the class-specific activations from classification network to segmentation network. Souly, Spampinato, and Shah (2017) adopted Generative Adversarial Networks (GANs) to provide extra training samples as a fake class, and the segmentation model acted as a discriminator to classify each pixel to a semantic label or fake label. The above-mentioned semi-supervised approaches are I3S-centric models, meaning they focuses on learning category-specific segmentation knowledge and would fail in case of a new added category. In this work, we attempt to solve the more general and practical C3S problem where annotations at different supervision levels are available across different categories.

Proposed Model

The proposed model includes two novel components, i.e., the L-Net for learning to produce label maps for weak categories from strong categories and the P-Net for predicting sharp and detailed semantic segmentation. Suppose the

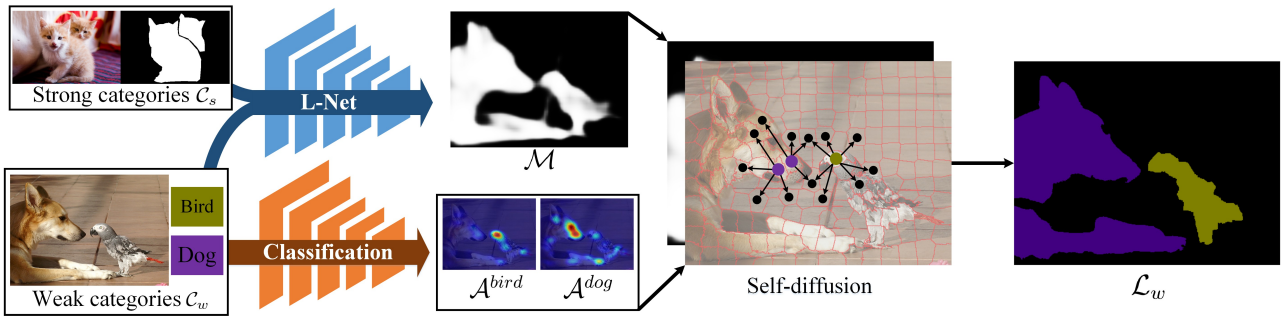

Figure 2: The flowchart of L-Net to produce the pixel-level annotations $\mathcal{L}_w$ for an image from weak categories $\mathcal{C}_w$. L-Net is trained on the images from strong categories with pixel-level annotations (with semantic information removed). Such more transferable knowledge enables L-Net to produce a class-agnostic segmentation map $\mathcal{M}$ for the image from $\mathcal{C}_w$. Based on the coarse segmentation $\mathcal{M}$, we propagate the class-wise activation maps $\mathcal{A}^{bird}$ and $\mathcal{A}^{dog}$ to generate the final annotations $\mathcal{L}_w$ by a self-diffusion algorithm. Best viewed in color.

weak categories and strong categories are denoted as $\mathcal{C}_w$ and $\mathcal{C}_s$ respectively. The pixel-level annotations for $\mathcal{C}_s$ are denoted as $\mathcal{L}_s$. For the weak categories $\mathcal{C}_w$ provided with only the image-level annotations, the pixel-level annotations $\mathcal{L}_w$ are generated by the L-Net.

L-Net: Generating Label Maps for Weak Categories

To learn the semantic segmentation model, the first step is to produce pixel-level annotations $\mathcal{L}_w$ for the images from weak categories $\mathcal{C}_w$. In order to provide relatively complete $\mathcal{L}_w$, we introduce the L-Net to learn to perform class-agnostic segmentation as category-agnostic knowledge is easier to learn and transfer among different categories. The learning process of L-Net is illustrated in Figure 2. Formally, given training images from the categories $\mathcal{C}_s$ that have pixel-level annotations, the objective for training L-Net (parameterized by θ_L) is defined as follows:

$$\min_{\theta_L} \sum_{\mathcal{C}_s} \mathcal{J}_b\big(\mathcal{L}'_s, \mathcal{O}_L(\mathcal{C}_s;\theta_L)\big), \tag{1}$$

where $\mathcal{O}_L(\mathcal{C}_s;\theta_L)$ denotes the output of L-Net, $\mathcal{L}'_s$ is the non-semantic ground truth derived by binarizing $\mathcal{L}_s$ and $\mathcal{J}_b$ denotes the standard element-wise binary cross-entropy loss.

The semantic information of $\mathcal{L}_s$ is removed in obtaining $\mathcal{L}'_s$ in order to learn more transferable knowledge across categories. Such a strategy can fully exploit the object-level information shared among strong categories and benefit segmentation over objects from the weak categories. After training, L-Net is applied to the images of $\mathcal{C}_w$ to produce the class-agnostic segmentation map $\mathcal{M} = \mathcal{O}_L(\mathcal{C}_w;\theta_L)$.

To recover the class-agnostic segmentation map $\mathcal{M}$ to $\mathcal{L}_w$ with rich semantic information, we employ an approach to predict class-discriminative activation by utilizing the image-level annotations available for weak categories. In particular, we employ a pre-trained image classification network to localize class-specific activations over the image plane. The bottom panel of Figure 2 visualizes the activation maps $\mathcal{A}^{bird}$ and $\mathcal{A}^{dog}$ produced by a classification network (Zhou et al. 2016) for two weak categories, `bird` and `dog` respectively. We take such localization results as reliable seeds for semantic segmentation and diffuse the semantic information originating from these seeds by a Random Walk (RW) based self-diffusion algorithm (Kong et al. 2016). Given an image from $\mathcal{C}_w$, we oversegment it into superpixels $\mathbf{p} = \{p_1, p_2, \cdots, p_N\}$ which are collectively described by a graph model G where each node corresponds to a particular superpixel. Then, the self-diffusion algorithm is performed on this undirected graph model G. Conditioned on $\mathcal{M}$, the objective function of the self-diffusion process for a specific category $\mathcal{A}^c$ is defined as

$$\min_{\mathbf{q}} \frac{1}{2} \sum_{i,j} z_{ij}(q_i - q_j)^2, \tag{2}$$

where $\mathbf{q} = [q_1, q_2, \cdots, q_N]$ denotes the label vector of all superpixels $\mathbf{p}$. If $p_i \in \mathcal{A}^c$, q_i is fixed to 1, and otherwise it takes an initial value of 0. $z_{ij} = \exp(-\|\mathcal{F}(p_i) - \mathcal{F}(p_j)\|/2\sigma^2)$ denotes the Gaussian distance between two adjacent superpixels. $\mathcal{F}(p_i) \in \mathbb{R}^4$ denotes the mean feature of superpixel p_i in the normalized CIELAB color space and the segmentation map $\mathcal{M}$.

Eqn. (2) formulates the conventional RW algorithm that strengthens label consistency of nodes with large affinity.

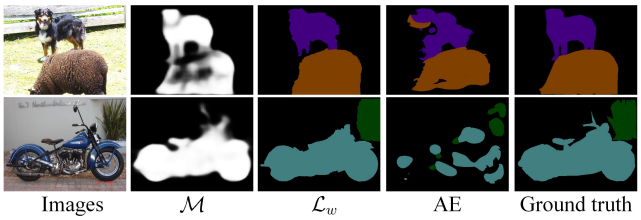

Figure 3: Comparison of generated label maps. $\mathcal{M}$ denotes the class-agnostic segmentation maps while $\mathcal{L}_w$ means the pixel-level annotations generated by the proposed L-Net. AE denotes the Adversarial Erasing approach (Wei et al. 2017) to generate the pixel-level annotations for weak categories. One can find that $\mathcal{L}_w$ provides sharp and complete semantic context, even if the $\mathcal{M}$ is noisy. Best viewed in color.

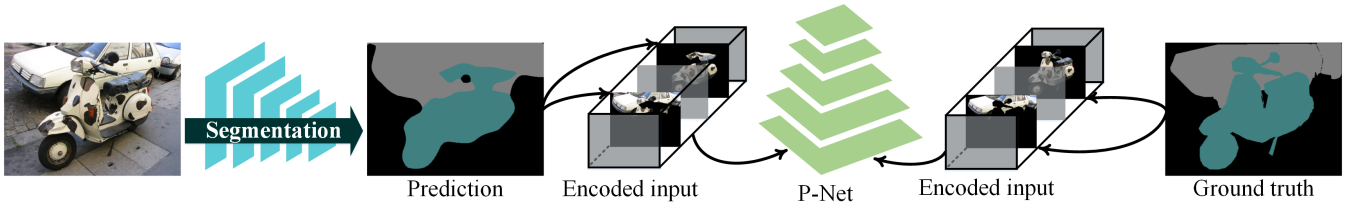

Figure 4: The framework of semantic segmentation with the P-Net. We propose to learn the semantic segmentation model through adversarial training. The input of the P-Net comes from the prediction of the semantic segmentation model and the ground truth, which is encoded by multiplying the training image with the mask of each category.

Considering there are some examples hard to be transferred within the L-Net, e.g., the `plant` in the second example of Figure 3, we impose no extra constraint on the segmentation map $\mathcal{M}$ in Eqn. (2). When the L-Net cannot segment all objects out, the high confidence class activation maps can still reveal and propagate segmentation information well. If there are more than two weak categories in the image to segment, we assign the category label c to the superpixel p_i with a larger q_i. As shown in Figure 3, we compare the generated label map by self-diffusion with the State-Of-The-Art (SOTA) Adversarial Erasing (AE) approach (Wei et al. 2017). Observing the generated pixel-level annotations, one can find that $\mathcal{L}_w$ provides semantic context at a satisfactory level, even if the segmentation map $\mathcal{M}$ is noisy.

P-Net: Semantic Segmentation with Adversarial Learning

Once L-Net generates the coarse pixel-annotations of weak categories, the semantic segmentation model can be trained upon such annotations. However, to get sharper and more accurate segmentation results, we introduce the P-Net component that learns to refine the semantic segmentation with adversarial training (Goodfellow et al. 2014), as shown in Figure 4. The generator within the adversarial learning framework is the semantic segmentation model in the left of Figure 4 which tries to predict label maps to match the joint data distribution of the ground truth and input images. The discriminator called P-Net acts to distinguish the input drawn from the generator or from the ground truth. On the one hand, the adversarial training forces the prediction of the semantic segmentation model to be as close as possible to the ground truth. On the other hand, the adversarial training learns to capture and utilize the implicit fitting patterns between the prediction and the ground truth which can be transferred to the weak categories.

Formally, for a given training sample I and its corresponding label map $\mathcal{L}_I$, we define the objective of adversarial training as follows:

$$\begin{aligned}&\min_{\theta_S}\max_{\theta_P}\sum_I \mathcal{J}_m\big(\mathcal{L}_I,\mathcal{O}_S(I;\theta_S)\big)-\\&\lambda\Big[\mathcal{J}_b\big(1,\mathcal{O}_P(\mathcal{L}_I;\theta_P)\big)+\mathcal{J}_b\big(0,\mathcal{O}_P(\mathcal{O}_S(I;\theta_S);\theta_P)\big)\Big],\end{aligned} \tag{3}$$

where θ_S and θ_P denote the parameters of the semantic segmentation model and the P-Net respectively. $\mathcal{J}_m$ and $\mathcal{J}_b$ denote the multi-class and binary cross-entropy loss respectively. $\mathcal{O}_S$ and $\mathcal{O}_P$ denote the output of the semantic segmentation model and the P-Net respectively. We use 1 and 0 to denote the label of P-Net when its input comes from the ground truth $\mathcal{L}_I$ and the prediction $\mathcal{O}_S(I;\theta_S)$ respectively.

For training the semantic segmentation model, we minimize the loss in Eqn. (3) w.r.t. θ_S:

$$\min_{\theta_S}\sum_I \mathcal{J}_m\big(\mathcal{L}_I,\mathcal{O}_S(I;\theta_S)\big)+\lambda\mathcal{J}_b\big(1,\mathcal{O}_P(\mathcal{O}_S(I;\theta_S);\theta_P)\big), \tag{4}$$

where the term $\lambda\mathcal{J}_b\big(1,\mathcal{O}_P(\mathcal{O}_S(I;\theta_S);\theta_P)\big)$ replaces the term $-\lambda\mathcal{J}_b\big(0,\mathcal{O}_P(\mathcal{O}_S(I;\theta_S);\theta_P)\big)$ in Eqn. (3). The first term in Eqn. (4) encourages the prediction of semantic segmentation to be consistent with the ground truth at each position while the second term penalizes the unfitting structure between the prediction and the ground truth.

For training the P-Net, we minimize the loss in Eqn. (3) w.r.t. θ_P:

$$\min_{\theta_P}\sum_I \Big[\mathcal{J}_b\big(1,\mathcal{O}_P(\mathcal{L}_I;\theta_P)\big)+\mathcal{J}_b\big(0,\mathcal{O}_P(\mathcal{O}_S(I;\theta_S);\theta_P)\big)\Big]. \tag{5}$$

Inspired by Luc et al. (2016), we do not directly input the probability maps predicted by the semantic segmentation network to P-Net. Instead, as shown in Figure 4, we encode the input of P-Net by multiplying the training image I with the predicted segmentation mask $\mathcal{O}_S(I;\theta_S)$ or the ground truth mask $\mathcal{L}_I$. This encoding makes the P-Net observe different objects and does not emphasize too much on the semantic label, which facilitates knowledge transfer across categories. Considering the unreliable label maps generated by the L-Net, directly training the whole network in Figure 4 could lead to poor performance of the P-Net, because the generated label maps may fall in conflict with the ground truth from strong categories. Therefore, we first pretrain the P-Net with the strong categories to encourage the P-Net to learn the real high-order fitting patterns and then fine-tune the whole training set. Experiments in the following section prove that it is indeed helpful to improve performance on weak categories.

Experiments

Implementation Details

Datasets We evaluate the performance of the proposed model on the PASCAL VOC 2012 benchmark (Everingham et al. 2014) which contains one background category and 20 object categories. The training set contains 10,582 images

Table 1: Layer configuration of P-Net

Layer	Channels	Kernel	Activation
conv1	16 32	3×3 3×3	ReLU
pool1	-	2× 2, stride 2	-
conv2	64 64	3×3 3×3	ReLU
pool2	-	2× 2, stride 2	-
conv3	128 128	3×3 3×3	ReLU
pool3		2× 2, stride 2	-
fc4	256-d	-	tanh
fc5	512-d	-	tanh
fc6	1-d	-	sigmoid

with pixel-level annotations, which is extended by Hariharan et al. (2011). We evaluate the performance in terms of mean Intersection over Union (mIoU) on other two subsets, i.e., validation and test, including 1,449 and 1,456 images respectively. According to the appearance similarity, we divide the 20 object categories into two super-categories, i.e., strong categories and weak categories, to guarantee each super-category contains similar categories. We provide four split-sets of the training images. Split-set 1 consists of 10 strong categories and 10 weak categories while split-set 2 is reversed based on split-set 1. Split-set 3 is a harder case which contains 6 strong categories and 14 weak categories. Similar to the setting in Hong et al. (2016), split-set 4 only provides image-level annotations for all 20 categories in PASCAL VOC 2012 while the strong categories are derived from MS-COCO (Lin et al. 2014). The training images containing PASCAL VOC 2012 categories are removed from MS-COCO and the remaining 16,241 images from 60 exclusive categories are employed as strong categories.

Network Architecture In this paper, we focus on transfer learning across various categories with different types of annotations. Therefore, extensive engineering on the segmentation network architecture is out of the scope of this work. We adopt the popular architecture of DeepLab-LargeFOV (Chen et al. 2015) as the backbone network for the L-Net in Figure 2 and the semantic segmentation network in Figure 4. DeepLab-LargeFOV is initialized by the weights of VGG-16 model (Simonyan and Zisserman 2014) which is pre-trained on the ImageNet. The L-Net differs from the semantic segmentation network in the loss function as shown in Eqn. (1) and Eqn. (5). The classification model in Figure 2 that provides category-specific activation maps is identical with the VGG-16 based CAM model (Zhou et al. 2016) and is fine-tuned on PASCAL VOC 2012 dataset with image-level labels. The P-Net in Figure 4 consists of six 3×3 convolutional layers and three fully connected layers. Details on layer configuration are provided in Tabel 1.

Training For the training of L-Net, we convert the semantic label maps from strong categories to a binary mask. We take a mini-batch size of 30, in which patches of 321×321 pixels are randomly cropped from images. We totally perform 30 epochs for training the L-Net with an initial learning rate of 5e-8. Momentum and weight decay are set to 0.9 and 0.0005 respectively. We train the semantic segmentation model in the same setting as DeepLab-LargeFOV. When the semantic segmentation model is trained, we fine-tune it with scratched P-Net and set the learning rate of the semantic segmentation model and the P-Net to 1e-5 and 1e-3 respectively. All the experiments are performed on NVIDIA TITAN X PASCAL GPU with 12G memory.

Comparison with Baselines

We evaluate various models on the PASCAL VOC 2012 validation set with four different strong/weak category splits. The results are summarized in Table 2. In particular, we compare the proposed model with following four baselines.

1. We use the fully-supervised DeepLab-LargeFOV (Chen et al. 2015) to gain performance upper bound for the compared weakly-/semi-supervised segmentation methods.
2. We also compare with an I3S-centric model, WSSL (Papandreou et al. 2015). It has the same segmentation network, i.e., DeepLab-LargeFOV, as our proposed model and is directly applied for the C3S problem introduced in this work. Following the practice in Papandreou et al. (2015), the semantic segmentation for weak categories is inferred via the adaptive EM algorithm.
3. We adopt a SOTA weakly-supervised approach, i.e., AE (Wei et al. 2017), as the third baseline, aiming to thoroughly compare model's capability of predicting pixel-level annotations for weak categories. During evaluation of AE, we apply AE to generate semantic label maps of the weak categories at first. Then we train and evaluate the DeepLab-LargeFOV model using the AE generated label maps (for weak categories) and the provided ground truths (for strong categories).
4. For the fourth split-set, we also compare with the TransferNet (Hong et al. 2016). We evaluate its performance using a stronger segmentation model DeconvNet (Noh, Hong, and Han 2015), under the same setting as our proposed model.

In Table 2, the numbers in gray block represent the segmentation performance of strong categories. The "L-Net" denotes the results obtained by the semantic segmentation model trained on both the strong categories and the label maps generated by L-Net. The "P-Net" denotes the final results by applying the P-Net to refine the semantic segmentation results. From the results, one can make following observations. The results of WSSL† in Table 2 demonstrate that the I3S-centric WSSL (Papandreou et al. 2015) performs poorly for the weak categories because it is incapable of transferring knowledge across categories. For the newly introduced category, WSSL performs not so well without additional pixel-level annotations. The proposed L-Net outperforms SOTA AE (Wei et al. 2017) by 13.1%, 10.6% and 16.3% on the first three split-sets respectively, confirming effectiveness of the L-Net on predicting high-quality

Table 2: Performance on PASCAL VOC 2012 validation set. The number in gray block represents the performance of categories with *pixel-level* annotations.

	bkg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbk	prsn	plnt	sheep	sofa	train	tv	mIoU
(a) Fully supervised baseline																						
DeepLab	90.2	73.2	31.0	73.7	58.6	63.5	81.4	74.5	76.9	28.4	62.7	49.9	69.4	61.5	67.1	76.1	46.2	67.8	42.0	74.1	52.6	62.9
(b) Split-set 1																						
WSSL†	83.3	31.1	31.2	29.1	59.9	31.9	80.6	74.3	78.2	29.7	61.6	43.7	39.6	38.6	29.5	74.8	21.5	33.4	19.1	45.5	48.4	46.9
AE	85.8	50.5	31.0	44.6	56.8	40.7	78.7	74.1	75.1	27.6	61.2	50.0	54.1	42.8	51.1	75.8	28.2	53.3	26.9	46.2	53.1	52.7
L-Net	89.6	74.0	30.8	66.8	58.7	43.1	80.8	75.6	76.3	28.3	61.7	48.8	66.5	60.7	68.2	76.0	30.4	68.8	27.4	65.8	53.1	59.6
P-Net	90.0	74.0	30.9	64.3	59.6	43.4	82.8	76.7	77.7	27.8	66.0	51.2	68.4	63.7	68.5	76.7	33.4	71.6	28.4	64.3	55.0	60.7
(c) Split-set 2																						
WSSL†	82.1	77.3	17.4	73.5	29.1	63.1	45.3	40.2	43.2	16.5	35.4	27.3	69.0	56.3	61.2	27.4	45.1	69.4	28.9	73.9	33.1	48.3
AE	84.2	72.1	22.9	71.8	32.6	61.0	63.6	30.0	59.5	16.0	43.2	22.7	67.5	58.7	65.4	53.0	45.4	66.9	37.7	70.9	39.9	51.7
L-Net	87.1	72.8	30.7	71.7	50.6	62.4	76.3	71.3	73.2	17.5	59.1	15.4	68.3	60.9	65.5	50.6	43.5	67.9	39.5	71.4	45.7	57.2
P-Net	87.7	74.4	31.1	72.6	53.9	62.7	77.1	73.0	73.9	17.5	61.8	16.6	70.3	62.2	66.8	51.5	45.0	69.3	40.0	72.8	49.0	58.5
(d) Split-set 3																						
WSSL†	82.4	31.0	15.2	30.1	26.2	33.2	44.7	42.2	45.1	23.6	64.4	25.6	70.3	38.4	54.4	76.5	25.1	33.6	20.6	75.9	33.8	42.5
AE	85.1	50.0	23.4	46.6	31.8	39.7	62.8	30.7	59.4	28.8	61.7	25.7	67.6	42.6	65.0	76.6	28.8	52.6	28.2	72.3	39.0	48.5
L-Net	89.1	58.3	33.5	71.1	34.8	42.0	75.4	67.7	74.9	27.1	62.2	24.5	70.5	61.1	67.5	76.4	31.2	68.4	25.5	74.2	47.5	56.4
P-Net	89.3	57.2	34.5	71.2	38.6	44.5	77.7	67.5	76.2	24.3	64.5	25.8	72.9	63.8	69.5	77.3	32.1	71.8	27.4	77.9	44.1	57.5
(e) Split-set 4																						
TransferNet	85.3	68.5	26.4	69.8	36.7	49.1	68.4	55.8	77.3	6.2	75.2	14.3	69.8	71.5	61.1	31.9	25.5	74.6	33.8	49.6	43.7	52.1
L-Net	86.5	70.9	26.3	70.1	46.3	55.3	73.9	67.6	72.9	20.6	61.4	21.2	66.1	60.5	63.2	55.2	32.7	66.3	34.2	64.4	44.9	55.3
P-Net	87.1	71.2	25.1	69.9	48.5	56.2	75.7	67.1	75.0	19.2	63.8	22.3	67.0	64.2	62.9	55.1	35.2	69.6	34.4	67.2	43.1	56.2

label maps. For some weak categories (e.g., the category `motorbike` in split-set 1 and the category `horse` in split-set 3), L-Net even performs slightly better than the fully-supervised model. We attribute this surprising superiority to the useful knowledge transferred across categories with similar appearance. For split-set 4, L-Net improves over TransferNet (Hong et al. 2016) by 6.1% under the same setting, proving the transferable segmentation knowledge in L-Net is more appropriate than the attention-based mechanism in TransferNet (Hong et al. 2016).

As shown in Table 2, employing adversarial training over a semantic segmentation model further improves the results by 1.9%. The P-Net pre-trained by the strong categories can learn the implicit fitting patterns between the prediction and the "real" pixel-level annotations. The learned suitable knowledge can be transferred to the weak categories and alleviate high-level disparities in the prediction of images from weak categories. We observe that pre-training on the strong categories is useful for stabilizing training process of P-Net. This is because some pixel-level annotations in weak categories are not reliable and may contaminate P-Net. If we directly train P-Net with the whole training set (consisting of provided pixel-level annotations and predicted ones from L-Net), we find the improvement brought by P-Net on split-set 1 is only 0.5%—on the other three split-sets the performance may even drop. Overall, the proposed model provides a very promising solution for segmenting the categories without pixel-level annotations and approaches the performance of the fully-supervised baseline.

Table 3: Comparison with weakly- and semi-supervised semantic segmentation models on PASCAL VOC 2012 test set.

Methods	#Training Set	mIoU
(a) Weakly-supervised methods		
DCSM (2016)	10k	45.1
BFBP (2016)	10k	48.0
STC (2016)	50k	51.2
SEC (2016)	10k	51.7
FCL (2017)	10k	53.7
AE (2017)	10k	55.7
Hong et al. (2017)	970k	58.7
(b) In-category semi-supervised methods		
WSSL (2015)	10k	66.2
DecoupledNet (2015)	10k	62.5
(c) Cross-category semi-supervised methods		
Ours (Split-set 1)	10k	64.6
Ours (Split-set 2)	10k	61.9
Ours (Split-set 3)	10k	59.5
Ours (Split-set 4)	27k	58.0
TransferNet (Split-set 4)	27k	51.2

Comparison with State-of-the-arts

We further compare our proposed model with several SOTA weakly- and semi-supervised semantic segmentation models, provided with different levels of annotations. Table 3 presents relevant results on PASCAL VOC 2012 test set. Among the compared models, Ours (Split-set 4), TransferNet (Hong et al. 2016), STC (Wei et al. 2016) and Hong et

Figure 5: Segmentation results on unseen categories from ImageNet (Russakovsky et al. 2015). All the results are produced by the L-Net which is trained on the split-set 1. Best viewed in color.

al. (2017) employ extra data (16k, 16k, 40k and 960k) for segmentation while the other methods are based on the 10k training samples of PASCAL VOC 2012. The pixel-level annotations in DecoupledNet (Hong, Noh, and Han 2015) are provided for 500 images while the number in WSSL is 1,464. For fair comparison, we apply post-processing over the results from P-Net with CRF (Krähenbühl and Koltun 2011). Compared with the latest weakly-supervised method (Hong et al. 2017), Ours (Split-set 4) performs competitively well as the model trained using 960k extra images in Hong et al. (2017). However, the proposed model only uses the image-level annotations of PASCAL VOC 2012 and 16k irrelevant pixel-level annotations.

For the semi-supervised semantic segmentation, even though as few as $1/2$ categories have pixel-level annotations in Ours (Split-set 1), the performance of proposed model only degrades by 2.5% compared with I3S-centric WSSL. Actually, based on the results in Table 2, the I3S-centric approaches (Papandreou et al. 2015; Hong, Noh, and Han 2015) cannot handle the C3S problem well and fail to generalize the weak categories. Such deficiency will restrain their application to the newly introduced categories. Compared with the attention-based TransferNet (Hong et al. 2016), the proposed model (Split-set 4) is advantageous. It introduces two complementary transferable components on segmentation knowledge and can provide superior semantic segmentation results as shown in Table 3.

Running Time

In this paper, training the L-Net with 3,000 images for 30 epochs takes about 3 hours while the inference of self-diffusion algorithm takes only 1 second for an input image. Training the P-Net with 10,000 images for 30 epochs takes about 12 hours. The total training time of the proposed method is about 17 hours (the training time of L-Net and P-Net plus self-diffusion inference on 7,000 weakly labeled images). The time cost is comparable with WSSL (Papandreou et al. 2015) which takes about 10 hours with the same setting. For testing, the proposed model has the same computational complexity as WSSL and it takes about 0.2 second to process a 300×400 image.

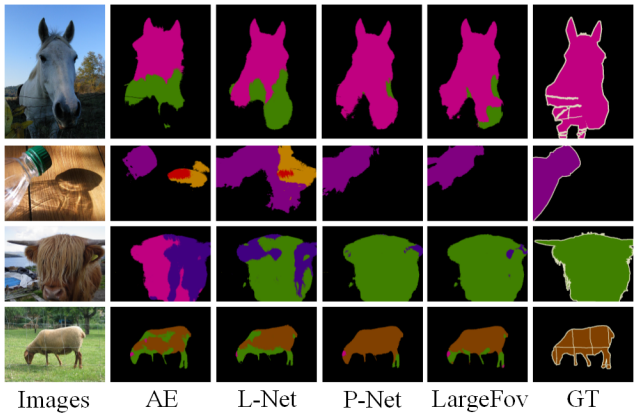

Figure 6: Visual comparison of semantic segmentation results. AE denotes the segmentation results by the weakly-supervised baseline (Wei et al. 2017) while LargeFov denotes the results by the DeepLab-LargeFov. P-Net denotes the refined results of L-Net by adversarial training. The first two examples come from the weak categories of split-set 1 while the last two examples come from the weak categories of split-set 2 and split-set 3 respectively. Best viewed in color.

Qualitative Results

To verify the effectiveness of the learned L-Net, we apply L-Net on the unseen categories from ImageNet as shown in Figure 5. All the results in Figure 5 are produced by the L-Net trained on the split-set 1 of PASCAL VOC 2012. One can find that the L-Net generalizes well on those unseen categories and provides sharp and complete segmentation masks. The L-Net generalizes well and provides practical solution for transferring a segmentation model from familiar objects to unseen ones. In Figure 6, we provide visual comparisons of the semantic segmentation results by AE (Wei et al. 2017), L-Net, P-Net and DeepLab-LargeFov (Chen et al. 2015). The first two examples come from the weak categories of split-set 1 and the last two examples come from the weak categories of split-set 2 and split-set 3 respectively. From the results of P-Net, one can observe that the adversar-

ial training can clean the noisy regions of L-Net and maintain consistency with the ground truth.

Conclusion

In this paper we tackle a more general problem in semi-supervised semantic segmentation where the strong categories and weak categories do not have overlap. We propose a novel transferable semi-supervised semantic segmentation model which contains two networks capable of learning and transferring segmentation knowledge, i.e., L-Net and P-Net. The L-Net generates the label maps of weak categories while the P-Net further refines the transferred knowledge by correcting high-level discrepancies between the prediction and ground truth. Benefited from the cross-category transferring, the proposed model provides superior performance over SOTA weakly-supervised approaches on the newly added category. Though only a small fraction of categories are with pixel-level annotations, the proposed model can still achieve 90% performance of the fully-supervised baseline. It enhances the applicability and scalability of semantic segmentation models in real applications.

Acknowledgments This work was supported in part by the National Natural Science Foundation of China under Grant 61403403, China Postdoctoral Science Foundation under Grant 2015M52707, and the China Scholarship Council under Grant 201603170287. The work of Jiashi Feng was partially supported by NUS startup R-263-000-C08-133, MOE Tier-I R-263-000-C21-112 and IDS R-263-000-C67-646.

References

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2015. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *ICLR*.

Everingham, M.; Eslami, S. A.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2014. The pascal visual object classes challenge: A retrospective. *IJCV* 111(1):98–136.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *NIPS*.

Hariharan, B.; Arbeläez, P.; Bourdev, L.; Maji, S.; and Malik, J. 2011. Semantic contours from inverse detectors. In *ICCV*.

Hong, S.; Oh, J.; Lee, H.; and Han, B. 2016. Learning transferrable knowledge for semantic segmentation with deep convolutional neural network. In *CVPR*.

Hong, S.; Yeo, D.; Kwak, S.; Lee, H.; and Han, B. 2017. Weakly supervised semantic segmentation using web-crawled videos. In *CVPR*.

Hong, S.; Noh, H.; and Han, B. 2015. Decoupled deep neural network for semi-supervised semantic segmentation. In *NIPS*.

Kolesnikov, A., and Lampert, C. H. 2016. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *ECCV*.

Kong, Y.; Wang, L.; Liu, X.; Lu, H.; and Ruan, X. 2016. Pattern mining saliency. In *ECCV*.

Krähenbühl, P., and Koltun, V. 2011. Efficient inference in fully connected crfs with gaussian edge potentials. In *NIPS*.

Kwak, S.; Hong, S.; and Han, B. 2017. Weakly supervised semantic segmentation using superpixel pooling network. In *AAAI*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollär, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *CVPR*.

Luc, P.; Couprie, C.; Chintala, S.; and Verbeek, J. 2016. Semantic segmentation using adversarial networks. *arXiv preprint arXiv:1611.08408*.

Noh, H.; Hong, S.; and Han, B. 2015. Learning deconvolution network for semantic segmentation. In *ICCV*.

Pan, T.; Wang, B.; Ding, G.; and Yong, J.-H. 2017. Fully convolutional neural networks with full-scale-features for semantic segmentation. In *AAAI*.

Papandreou, G.; Chen, L.-C.; Murphy, K. P.; and Yuille, A. L. 2015. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *ICCV*.

Roy, A., and Todorovic, S. 2017. Combining bottom-up, top-down, and smoothness cues for weakly supervised image segmentation. CVPR.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *IJCV* 115(3):211–252.

Saleh, F.; Akbarian, M. S. A.; Salzmann, M.; Petersson, L.; Gould, S.; and Alvarez, J. M. 2016. Built-in foreground/background prior for weakly-supervised semantic segmentation. In *ECCV*.

Shimoda, W., and Yanai, K. 2016. Distinct class-specific saliency maps for weakly supervised semantic segmentation. In *ECCV*.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. In *ICLR*.

Souly, N.; Spampinato, C.; and Shah, M. 2017. Semi and weakly supervised semantic segmentation using generative adversarial network. *arXiv preprint arXiv:1703.09695*.

Wei, Y.; Liang, X.; Chen, Y.; Shen, X.; Cheng, M.-M.; Feng, J.; Zhao, Y.; and Yan, S. 2016. Stc: A simple to complex framework for weakly-supervised semantic segmentation. *IEEE TPAMI*.

Wei, Y.; Feng, J.; Liang, X.; Cheng, M.-M.; Zhao, Y.; and Yan, S. 2017. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *CVPR*.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *CVPR*.