Image-level to Pixel-wise Labeling: From Theory to Practice

Tiezhu Sun[1], **Wei Zhang**[1*], **Zhijie Wang**[1], **Lin Ma**[2*] and **Zequn Jie**[2]
[1]School of Control Science and Engineering, Shandong University
[2]Tencent AI Lab, Shenzhen, China
{suntiezhu, wangzj}@mail.sdu.edu.cn, {davidzhangsdu, forest.linma, zequn.nus}@gmail.com

Abstract

Conventional convolutional neural networks (CNNs) have achieved great success in image semantic segmentation. Existing methods mainly focus on learning pixel-wise labels from an image directly. In this paper, we advocate tackling the pixel-wise segmentation problem by considering the image-level classification labels. Theoretically, we analyze and discuss the effects of image-level labels on pixel-wise segmentation from the perspective of information theory. In practice, an end-to-end segmentation model is built by fusing the image-level and pixel-wise labeling networks. A generative network is included to reconstruct the input image and further boost the segmentation model training with an auxiliary loss. Extensive experimental results on benchmark dataset demonstrate the effectiveness of the proposed method, where good image-level labels can significantly improve the pixel-wise segmentation accuracy.

1 Introduction

Semantic segmentation aims to assign a categorical label to every pixel in an image, which plays an important role in autonomous navigation, human-machine interaction, and virtual reality. Earlier attempts on semantic segmentation focused on designing hand-crafted features together with flat classifiers, such as Boosting [Shotton *et al.*, 2009] and Support Vector Machines [Fulkerson *et al.*, 2009]. The recent success of deep convolutional neural networks (CNNs) [Krizhevsky *et al.*, 2012; Szegedy *et al.*, 2015] on object recognition [He *et al.*, 2016], localization [Jie *et al.*, 2016], re-identification [Zhang *et al.*, 2017], multimodal learning [Ma *et al.*, 2016], and segmentation [Li *et al.*, 2017; Jin *et al.*, 2017], has empowered the development of pixel-wise semantic segmentation due to the rich hierarchical features and end-to-end training architectures [Long *et al.*, 2015; Lin *et al.*, 2016; Wang *et al.*, 2017; Chen *et al.*, 2016; Papandreou *et al.*, 2015; Zhang *et al.*, 2018]. Among these deep models, fully convolutional network (FCN) [Long *et al.*, 2015] is prevalent, as it can allow arbitrary input size and produce results with efficient inference. Besides, PSPNet [Zhao *et al.*, 2016] addressed the capability of global context aggregation through a pyramid pooling module together with a scene parsing network. Two similar works [Liu *et al.*, 2015; Pan *et al.*, 2017] seek a better representation by fusing the global features obtained from the whole image with additional feature learning layers. To conclude, most existing methods share a similar view of learning pixel-wise categorical labels from the image directly.

However, when looking at an image, the first reaction of human is to find out the objects involved roughly ($e.g.$, a car or an animal) and then figure out what they look like ($e.g.$, shape and boundary). Inspired by such cognitive psychology clue, we advocate regarding semantic segmentation as an image-level to pixel-wise labeling task. In practice, we decompose semantic segmentation into two sub-tasks: estimating object categories existing in input image and learning the shapes and boundaries of these specific objects, which could be referred as a process of image-level labeling to pixel-wise labeling. The key insight is to address the pixel-wise segmentation problem progressively rather than accomplish it directly. Based on the information theory, we first give the theoretical proof that accurate image labels can eliminate the invalid information in pixel labeling and make the task easier. Besides, analysis is made to discuss what effects may be caused on pixel-wise labeling when image labels are partially accurate.

Specifically, the proposed segmentation model is composed of two streams: an image-level labeling network to find existing object categories and a pixel-wise labeling network ($i.e.$, the segmentation network) to label a pixel with a specific object category, as illustrated in Figure 1. The category information extracted by the image-labeling network will be fused with the output of pixel-labeling network by dot product. It is found that the fusion can boost the segmentation performance by reducing the invalid information that the pixel-labeling network needs to learn, $i.e.$, the uncertainty about pixel labels.

The loss function is defined as follows. First, multi-class cross-entropy loss is used to measure the pixel-wise difference between the network prediction and the groundtruth. Second, as a generative network is introduced to reproduce the input image from its predicted segmentation result, a reconstruction loss between the input image and the recon-

*Corresponding authors.

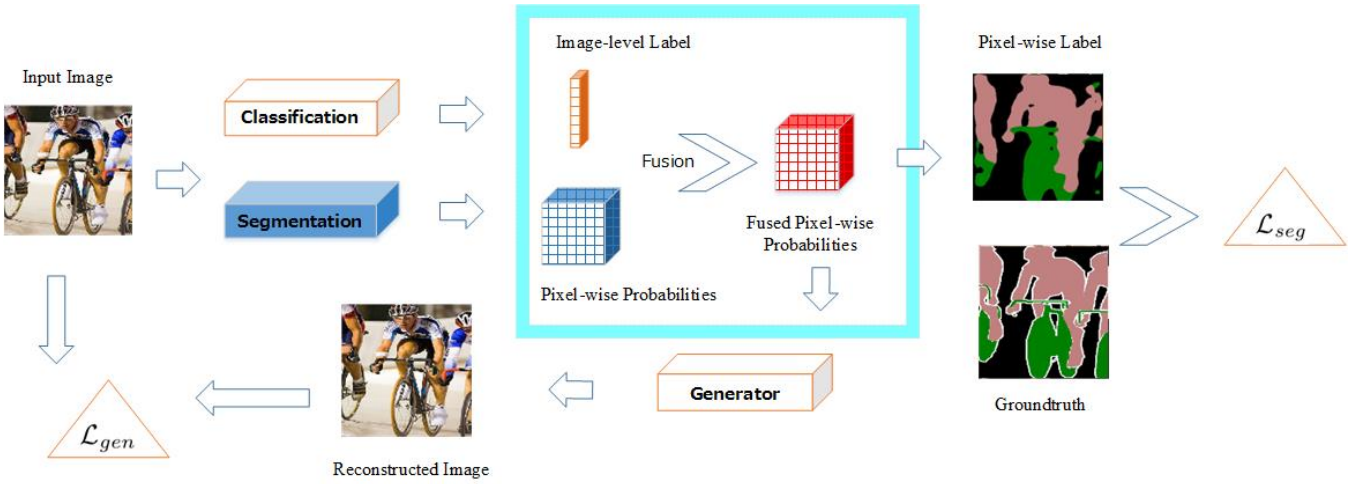

Figure 1: Illustration of the proposed segmentation architecture. The input image simultaneously undergoes two networks to yield the image-level label and the pixel-wise probabilities specified for each category. The two intermediate results are fused to yield the corrected probabilities, based on which the final pixel-wise segmentation results are generated. A generative network is included to reconstruct the input image and provides an auxiliary loss to further boost the model.

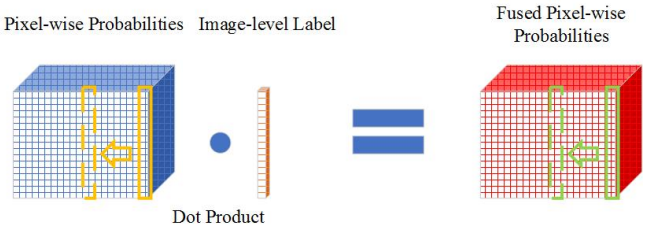

Figure 2: Illustration of image-level label and pixel-wise probability fusion with dot product operation.

structed one can be yielded. With the proposed two loss functions, the whole network is trained in an end-to-end manner.

As a summary, the main contributions of this work are threefold. First, we present a progressive image-to-pixel solution to address the semantic segmentation problem. Second, theoretical analysis is given to prove that good image labels are beneficial to reduce the uncertainty in pixel-wise labeling problem. An additional generative network is included to provide auxiliary loss to further constrain the model. Last, the proposed architecture is compatible to current CNN-based models, and significant segmentation gains can be obtained on benchmark dataset.

2 Methodology

The whole structure of our approach is illustrated in Figure 1. The input image simultaneously undergoes two networks to yield the image-level label and the pixel-wise probabilities specified for each category. The two intermediate results are fused to yield the corrected probabilities, based on which the final pixel-wise segmentation results are generated. Moreover, we propose a generative network to reconstruct the image based on the fused probability feature maps. Therefore, our training objective is defined as $\mathcal{L}_{overall}(\theta_s, \theta_g) = \mathcal{L}_{seg} + \mathcal{L}_{gen}$, where $\mathcal{L}_{seg}$ and $\mathcal{L}_{gen}$ denote the pixel-wise segmentation loss and reconstruction loss, respectively. Different segmentation networks can be employed for pixel-wise labeling, with the parameters tuned together with the generator. Due to the limited number of training samples, we employ the classification network trained in [Wei *et al.*, 2014] to generate the image-level label, with the parameters of classification network fixed during our training procedure. It is worth noting that the classification network can also be tuned jointly as long as enough training samples are available.

2.1 Fusion of Image-level and Pixel-wise Labels

The idea of the image-level and pixel-wise labels fusion is to decrease the influence of a category as much as possible, if it does not exist in the given image. Thus, the segmentation network does not need to consider such category and the uncertainty of the task could be significantly reduced.

Concretely, dot product is employed to perform the fusion as shown in Figure 2. The pixel-wise probability maps predicted from the segmentation model are represented by a three dimensional matrix, while the classification network generates the image-level label in vector form. Suppose the segmentation model yields a semantic prediction $P \in [0,1]^{w \times h \times N}$, each pixel in P is represented by a vector which has N elements indicating the probabilities that it belongs to a specific category: $\sum_p P(i,j,p) = 1$. The image-level label is denoted as a vector $L \in \{0,1\}^N$. Each element in L is either 1 or 0, which means that the image is either associated with the corresponding category or not. The dot product operation is performed as follows.

$$P'(i,j,\cdots) = P(i,j,\cdots) \bullet L, \tag{1}$$

where $i \in \{1,...,w\}$ and $j \in \{1,...,h\}$ denote the horizontal and vertical pixel positions, respectively. The result of the product of P and L is assigned to P' as the corrected prediction. After the dot product, the P' is re-normalized to make

sure that $\sum_p P'(i,j,p) = 1$. With such fusion, the pixel-wise prediction could be improved, based on which we can generate the segmentation results.

As aforementioned, different networks can be employed to realize the segmentation. In this paper, FCNs are used as the baseline segmentation networks. Take FCN-32s for example, for an input image u of size $w \times h \times 3$, FCN-32s together with the classification network generate the fused probability maps $s(u)$, where s denotes the mapping function learned by the segmentation architecture. Given an image u_q and corresponding label mapping v_q in a dataset of Q training samples, the segmentation loss $\mathcal{L}_{seg}$ is defined as

$$\mathcal{L}_{seg}(\theta_s) = \sum_{q=1}^{Q} l_{mce}(s(u_q), v_q), \tag{2}$$

where θ_s denotes the parameters of FCN-32s, and l_{mce} denotes the multi-class cross-entropy loss for the predicted $s(u)$.

2.2 Auxiliary Loss

With the pixel-wise segmentation results, we can introduce a generative network to reconstruct the input image. Thus, a reconstruction loss between the input image and the reconstructed one can be yielded to further constrain the model. The idea is similar to those in [Zhu *et al.*, 2017; Yi *et al.*, 2017; Luo *et al.*, 2017]. Let g denote the mapping function learned by the generative network, $g(s(u))$ is the reconstructed image with the size of $w \times h \times 3$. The auxiliary reconstruction loss $\mathcal{L}_{gen}$ can be defined as

$$\mathcal{L}_{gen}(\theta_g) = \sum_{q=1}^{Q} l_{euc}(g(s(u_q)), u_q), \tag{3}$$

where θ_g denotes the parameters of the proposed generative network, and l_{euc} denotes the Euclidean distance which is better at preventing over-fitting. The auxiliary reconstruction loss obtained by the generative network and the calculated segmentation loss are jointly considered to train the segmentation and generative networks. The generative network consists of three convolutional layers, where the sizes of the kernels are $21 \times 3 \times 3 \times 18$, $18 \times 3 \times 3 \times 9$, $9 \times 3 \times 3 \times 3$, respectively.

2.3 Training Procedure

For each epoch, we train both segmentation and generative networks with Q iterations, where Q denotes the size of dataset. In each iteration, the training proceeds in three steps. First, the segmentation network is trained individually based on $\mathcal{L}_{seg}$. Second, the generative network is trained to reconstruct the input image referring to $\mathcal{L}_{gen}$ with the parameters of the segmentation network fixed. Third, the segmentation and generative networks are jointly trained by considering both $\mathcal{L}_{seg}$ and $\mathcal{L}_{gen}$. Such training strategy is employed for the sake of smoothness in multi-network optimization as proved in [Lee *et al.*, 2015].

3 Theoretical Analysis

To investigate the effects of image labeling on pixel-wise segmentation, we make a theoretical analysis from the information theory perspective in this section. Suppose the whole image dataset is associated with N categories, the category set is $X = \{x_1, x_2, ..., x_N\}$. Given a specific input image associating with n categories ($1 \leq n \leq N$) with a resolution of $w \times h$, we analyze the amount of invalid information that the image-level labels can decrease for segmentation model to generate pixel-wise segmentation under two assumptions.

3.1 Assumption I: Accurate Image-level Label

When the image label is absolutely accurate, it can help to eliminate all invalid information which may be learned by the pixel-wise segmentation network. Thus, such network could learn the categories for each pixel only based on the valid information, which is related to existing categories.

Let x be the label of one pixel in a given image, the initial self-information of the event $x = x_i$ ($1 \leq i \leq N$) is defined as

$$\begin{aligned} I(x) &= -\mathbf{log}P(x) = -\mathbf{log}\frac{1}{N} \\ &= \mathbf{log}N, \end{aligned} \tag{4}$$

where $\mathbf{log}$ denotes the natural logarithm with base e, and $I(x)$ denotes the amount of information that the segmentation model needs to learn from the input image. The definition of $I(x)$ is therefore written in unit as $nats$. One nat is the amount of information gained by observing an event of probability $1/e$.

The image-level label (represented in vector form) is fused with the pixel-wise label by dot product, the idea is to set the pixel-wise probabilities of the categories that do not exist in the image to zeros, as shown in Figure 2. After dot product, the category set turns into $X' \sqsubseteq X$, and X' is associated with n categories which are all valid. The self-information of the event $x = x_j$ ($1 \leq j \leq n$) becomes

$$\begin{aligned} I'(x) &= -\mathbf{log}P'(x) = -\mathbf{log}\frac{1}{n} \\ &= \mathbf{log}n. \end{aligned} \tag{5}$$

Now the amount of information that segmentation model needs to learn are decreased to $I'(x) = \mathbf{log}n$ nats, which denote all valid information. The difference between $I(x)$ and $I'(x)$ is calculated as follows.

$$\begin{aligned} \hat{d} &= I(x) - I'(x) \\ &= \mathbf{log}N - \mathbf{log}n \\ &= \mathbf{log}\frac{N}{n}. \end{aligned} \tag{6}$$

In fact, $\hat{d}$ denotes the amount of invalid information coming from those $N - n$ outlier categories, which do not exist in the given image.

In one extreme case, when there are only one existing category in the given image ($i.e.$, $n = 1$),

$$\hat{d} = \mathbf{log}\frac{N}{1} = \mathbf{log}N = I(x). \tag{7}$$

This means that all $\mathbf{log}N$ nats of information leaned by the segmentation model are invalid. In other words, the segmentation model can predict the label x of each pixel correctly relying only on the image-level label.

In the other case, when all the N categories exist in the given image ($i.e.$, $n = N$),

$$\hat{d} = \mathbf{log}\frac{N}{N} = 0. \tag{8}$$

This means that the information learned by the segmentation model is all valid, so no information needs to be eliminated. Hence, the dot product does not exert any influence on the segmentation model.

Finally, as for the whole image, the amount of invalid information that the image label helps to decrease are $w \times h \times \mathbf{log}\frac{N}{n}$ nats.

3.2 Assumption II: Noising Image-level Labels

In general, the image-level labels contain three types of noises.

Case 1. Besides the n correct categories, the image label also contains some outlier categories that do not exist in the given image. Let m ($1 \leq m \leq N - n$) denote the number of these outlier categories, the amount of information that segmentation model needs to learn after dot product are decreased from $I(x)$ to

$$\begin{aligned} I_m(x) &= -\mathbf{log}P_m(x) \\ &= -\mathbf{log}\frac{1}{n+m} \\ &= \mathbf{log}(n+m). \end{aligned} \tag{9}$$

The difference between $I(x)$ and $I_m(x)$ is calculated as

$$\begin{aligned} d_m &= I(x) - I_m(x) \\ &= \mathbf{log}N - \mathbf{log}(n+m) \\ &= \mathbf{log}\frac{N}{n+m}. \end{aligned} \tag{10}$$

Since $n < n + m \leq N$, we get $0 \leq \mathbf{log}\frac{N}{n+m} = d_m < \hat{d} = \mathbf{log}\frac{N}{n}$, which means that the image-level label can still decrease $\mathbf{log}\frac{N}{n+m}$ nats of invalid information. Nevertheless, these decreased $d_m = \mathbf{log}\frac{N}{n+m}$ nats of invalid information are less than $\hat{d}$ defined in Equation (6), and $\hat{d} - d_m = \mathbf{log}\frac{n+m}{n}$. Specially, when m reaches its maximum value $N - n$, d_m reaches its minimum value $\mathbf{log}\frac{N}{n+N-n} = 0$, implying that image label cannot decrease any invalid information.

All the situations we discussed before share a common point that image labels cover all existing categories, which do not have any loss of valid information.

Case 2. When the image label is partially accurate, which including only a part of n existing categories but no outlier categories, some negative effects may be caused. Let k ($1 \leq k < n$) be the number of the part existing categories that one image label contains, the amount of information that segmentation model can learn after dot product operation is

$$\begin{aligned} I_k(x) &= -\mathbf{log}P_k(x) = -\mathbf{log}\frac{1}{k} \\ &= \mathbf{log}k. \end{aligned} \tag{11}$$

Since $I'(x) = \mathbf{log}n$ denotes all the valid information and $k < n$, we should compute the difference between $I'(x)$ and $I_k(x)$ as

$$\begin{aligned} d_k &= I'(x) - I_k(x) \\ &= \mathbf{log}n - \mathbf{log}k \\ &= \mathbf{log}\frac{n}{k}. \end{aligned} \tag{12}$$

It is found that $d_k = \mathbf{log}\frac{n}{k} > 0$, as $1 \leq k < n$. So d_k denotes the loss of valid information provided by those $n - k$ categories, which should be maintained yet are eliminated by dot product. Therefore, the segmentation model cannot learn these $\mathbf{log}\frac{n}{k}$ nats of valid information. It means that the pixel-wise labeling on these eliminated categories will be wrong.

Case 3. The last case is a combination of the two aforementioned cases. Hence, the amount of information that the segmentation model can learn is

$$\begin{aligned} I_{m,k}(x) &= -\mathbf{log}P_{m,k}(x) \\ &= -\mathbf{log}\frac{1}{m+k} \\ &= \mathbf{log}(m+k). \end{aligned} \tag{13}$$

As we discussed in Case 2, since k correct categories exist in the image label, the amount of valid information is $I_k(x) = \mathbf{log}k$. So, the amount of lost valid information is $d_k = \mathbf{log}\frac{n}{k}$, which is the same as that in Case 2. The difference between $I_{m,k}(x)$ and $I_k(x)$ is calculated as

$$\begin{aligned} d_{m,k} &= I_{m,k}(x) - I_k(x) \\ &= -\mathbf{log}\frac{1}{m+k} - (-\mathbf{log}\frac{1}{k}) \\ &= \mathbf{log}\frac{m+k}{k}. \end{aligned} \tag{14}$$

The $d_{m,k}$ represents the amount of invalid information resulted from those m outlier categories. However, the whole segmentation performance in this case and Case 2 is not completely bad. To suppress the amount of invalid information $d_{m,k}$, we can decrease the number of outlier labels ($i.e.$, m) by increasing the classification threshold τ as shown in Section 4.1. But the side effect is that the number of valid labels k will be reduced as well. Thus, it is unknown $d_{m,k}$ will increase or decrease, similar when decreasing τ. So, the suggestion obtained in these two cases is that a too large or too small τ are both undesired. There should be a tradeoff for τ as demonstrated in Figure 4.

On the other hand, increasing the threshold may have some positive effects, though more valid labels may be removed (k is smaller). First, the number of invalid labels ($i.e.$, m) will be decreased as well. Second, the labels to be inferred turn to be less. So, the segmentation task will be easier and the labeling accuracy for the rest k categories could be promoted.

The following sections are presented to validate the above theoretical analysis based on experimental results.

4 Experimental Results

In this section, the evaluation is conducted on the benchmark segmentation dataset PASCAL VOC 2012 [Everingham *et*

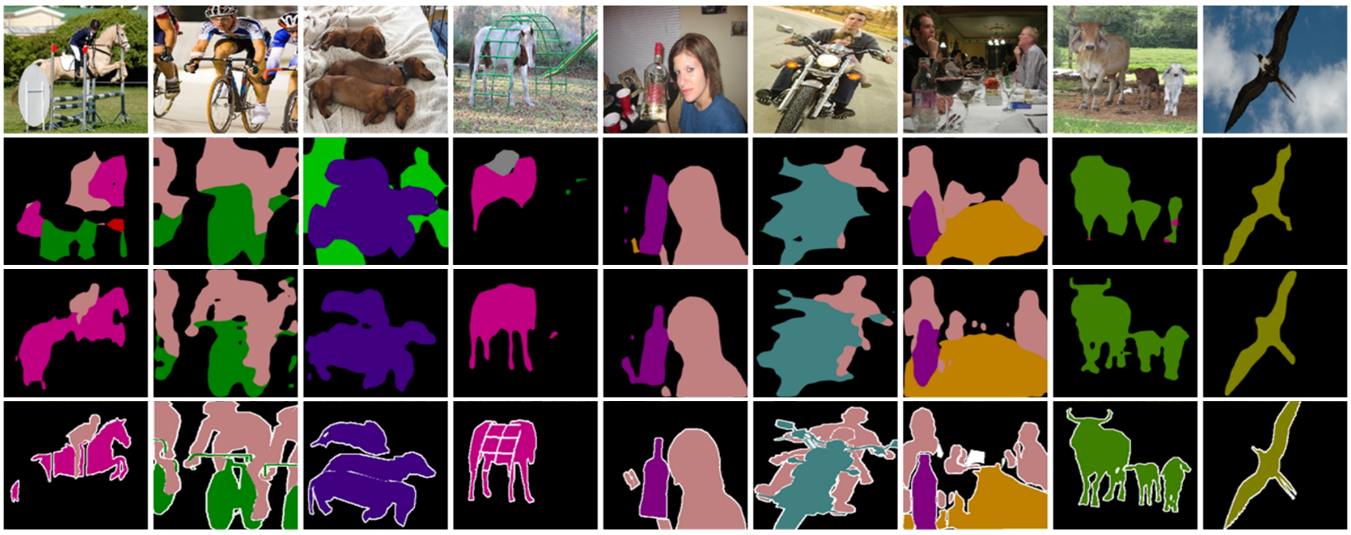

Figure 3: Visualization of segmented examples in PASCAL VOC 2011 validation set. From top to bottom, each row represents the input image, segmentation results of FCN-8s, segmentation results of the proposed method and the groundtruth, respectively.

	PA	MPA	FWAA	MIOU
FCN-32s	89.1	73.3	81.4	59.4
FCN-32s+image label	93.6	87.4	88.5	74.9
FCN-32s+generator	90.7	75.5	83.8	63.8
FCN-32s+label+gen	**93.9**	**87.5**	**89.1**	**75.7**

Table 1: Ablation study on PASCAL VOC 2011 validation set.

	PA	MPA	FWAA	MIOU
FCN-32s	89.1	73.3	81.4	59.4
FCN-32s+dis	88.7	68.9	80.4	60.6
FCN-32s+gen	**90.7**	**75.5**	**83.8**	**63.8**

Table 2: Analysis of auxiliary loss on PASCAL VOC 2011 validation set.

al., 2010], which consists of 21 classes of objects (including background). Similar to [Zhao *et al.*, 2016], we use the augmented data of PASCAL VOC 2012 with annotation of [Hariharan *et al.*, 2011] resulting 11,295, 736, 1456 samples for training, validation and testing, respectively. To test the benefits of the image-to-pixel labeling strategy, typical model FCNs were employed as the baseline segmentation models. In the training stage, SGD and Adam were employed as optimizers to train the segmentation and generative networks with the same learning rate of 10^{-10}, respectively. The iteration number 100,000 is set for all experiments.

The evaluation metrics used in this work include: Pixel Accuracy (PA), Mean Pixel Accuracy (MPA), Frequency Weighted Average Accuracy (FWAA) and Mean Intersection over Union (MIOU). PA represents the proportion of the correctly labeled pixels to the total pixels. MPA denotes the mean of pixel accuracy of all categories. FWAA is the PA weighted by the pixel ratio of each category in one image. MIOU is a standard measure commonly used in segmentation problem [Rahman and Wang, 2016], which gives the similarity between the predicted region and the groundtruth region for an object.

4.1 Effect of Image Label Accuracy

To verify the amount of information that image labels decrease for segmentation as studied in Section 3, experimental analysis is performed as follows.

Image Label is Absolutely Accurate

Since the groundtruth image labels are available in the validation set of PASCAL VOC 2011, experiments were conducted on it to test the proposed framework when the image labels are completely accurate (using the groundtruth image labels). As listed in Table 1, the results are consistent with the conclusion drawn from Assumption I in Section 3.1. That is, accurate image labels can eliminate all amount of invalid information for pixel labeling and the segmentation performance is improved significantly in all metrics, $e.g.$, the gains of MIOU reach 15.5%. Besides, the generator is also proved to be beneficial, providing the gains of 4.4% in MIOU. Certainly, the best segmentation performance is yielded by incorporating both image labels and generator.

Image Label is Partially Accurate

The experimental validation about Assumption II was conducted on the testing set of PASCAL VOC 2012 whose groundtruth labels are unavailable. Even though with state-of-the-art image-level labeling network, the image labels may have errors inevitably. In the experiments, the trained image-labeling network [Wei *et al.*, 2014] was employed to generate the image labels, which achieved about 90% in mean average precision (MAP) when classifying the testing set of PASCAL VOC 2012. Since the image-labeling network only outputs the probabilities of categories that may exist in the given image, the threshold value used to decide the image labels is vital in discussing the influence of errors in image labels to

	PA	MPA	FWAA	MIOU
FCN-32s	89.1	73.3	81.4	59.4
FCN-32s+image label (addition)	91.2	77.7	84.5	66.3
FCN-32s+image label (dot product)	**93.6**	**87.4**	**88.5**	**74.9**

Table 3: Analysis of image-level and pixel-wise feature fusion on PASCAL VOC 2011 validation set.

Method	MIOU
FCN-32s	59.3
FCN-32s+image label	65.1
FCN-32s+generator	63.6
FCN-32s+image label+generator	**66.4**
FCN-8s	62.2
FCN-8s+image label	66.7
FCN-8s+generator	65.4
FCN-8s+image label+generator	**68.2**

Table 4: Comparison on PASCAL VOC 2012 testing set.

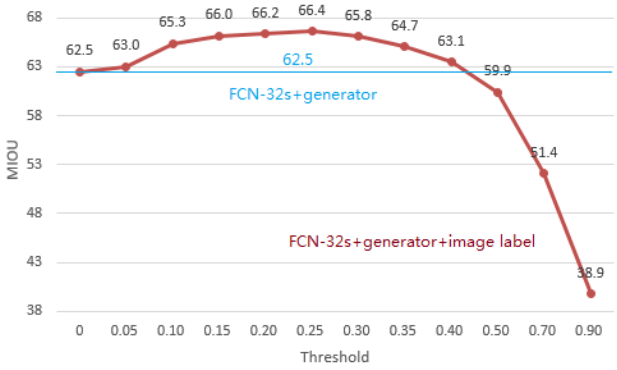

Figure 4: Analysis of the effects of different thresholds on PASCAL VOC 2012 testing set.

pixel labeling performance.

First, a lower threshold may incur more outlier categories that do not exist in the image indeed, which belongs to Case 1 in Assumption II in Section 3.2. In this case, the amount of invalid information that the image labels can help to decrease for segmentation model is limited, so the improvement is little. In the extreme case of which the threshold $\tau = 0$, the image labels contain all categories, implying that $m = N - n$ and $d_m = 0$ as proved in Case 1 in Assumption II. In this situation, the image label does not decrease any invalid information, and the result equals to that of the baseline ($i.e.$, FCN-32s+generator) as shown in Figure 4.

If the threshold is set too high, some correct categories may be missed in image labels, which is consistent with Case 2 in Assumption II. Since some correct categories are excluded, certain amount of valid information needed is lost, which even make the performance worse than that of the baseline methods ($e.g.$, $\tau = 0.7$). Especially, the performance drops quickly when the threshold approaches 1 ($i.e.$, 100%) which resulted from that almost all corrected categories are eliminated, as shown in Figure 4.

In general, a relatively low threshold is more appealing in practice, and MIOU gains could be produced as long as τ is below 0.43. This situation is consistent with Case 3 in Assumption II, where some desired correct categories are maintained together with some undesired outlier ones. The results in Figure 4 show that $\tau = 0.25$ is a good tradeoff.

4.2 Performance Analysis

Analysis on the Auxiliary Loss

Besides the generative model, we can also introduce a discriminator which is similar in GAN [Goodfellow *et al.*, 2014] to provide auxiliary loss to boost the model training. The discriminator here is composed of five convolutional layers, and the segmentation results obtained with generator and with discriminator respectively are compared in Table 2. Apparently, the comparison shows the generator is superior over the discriminator to boost the proposed segmentation model as an auxiliary means.

Analysis on the Network Fusion

We tested different operations to fuse the information extracted by the image-labeling network and the pixel-labeling network. As shown in Table 3, dot product is a better choice than addition for network fusion. The reason is that the dot product can correct the probabilities of outlier categories immediately (by setting them to zero), while addition can only increase the right probabilities to some extent.

Overall Performance

On PASCAL VOC 2012 testing set, as shown in Table 4, by incorporating the proposed framework, the typical FCN segmentation model FCN-32s and FCN-8s were boosted considerably with the overall gains of 7.1% and 6.0% in MIOU, respectively. One point needs to be clarified is that the MIOU values of FCN-32s+generator in Table 4 and Figure 4 are inconsistent, because the image label did not participate in training in the experiments of Table 4 but did in Figure 4. Also, the improvement can be visualized in Figure 3, where some segmented examples are given for comparison.

5 Conclusions

In this paper, we addressed the semantic segmentation problem with the assistance of the image-level classification labels. Theoretical studies showed that good image-level labels can reduce the uncertainty in pixel-wise labeling to boost the segmentation accuracy. We also proposed a deep model by fusing the image-level and pixel-wise labeling networks. Various experimental results demonstrated that typical segmentation networks can be improved considerably on benchmark dataset with the proposed architecture.

Acknowledgements

This work was supported by the NSFC Grant No.61573222, Shenzhen Future Industry Special Fund

JCYJ20160331174228600, Major Research Program of Shandong Province 2015ZDXX0801A02, National Key Research and Development Plan of China under Grant 2017YFB1300205 and Fundamental Research Funds of Shandong University 2016JC014.

References

[Chen *et al.*, 2016] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L Yuille. Attention to scale: Scale-aware semantic image segmentation. In *CVPR*, 2016.

[Everingham *et al.*, 2010] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010.

[Fulkerson *et al.*, 2009] Brian Fulkerson, Andrea Vedaldi, and Stefano Soatto. Class segmentation and object localization with superpixel neighborhoods. In *ICCV*, 2009.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014.

[Hariharan *et al.*, 2011] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *ICCV*, 2011.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[Jie *et al.*, 2016] Zequn Jie, Xiaodan Liang, Jiashi Feng, Xiaojie Jin, Wen Feng Lu, and Shuicheng Yan. Tree-structured reinforcement learning for sequential object localization. In *NIPS*, 2016.

[Jin *et al.*, 2017] Xiaojie Jin, Yunpeng Chen, Zequn Jie, Jiashi Feng, and Shuicheng Yan. Multi-path feedback recurrent neural networks for scene parsing. In *AAAI*, 2017.

[Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[Lee *et al.*, 2015] Chen-Yu Lee, Saining Xie, Patrick W. Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *AISTATS*, 2015.

[Li *et al.*, 2017] Xin Li, Zequn Jie, Wei Wang, Changsong Liu, Jimei Yang, Xiaohui Shen, Zhe Lin, Qiang Chen, Shuicheng Yan, and Jiashi Feng. Foveanet: Perspective-aware urban scene parsing. In *ICCV*, 2017.

[Lin *et al.*, 2016] Guosheng Lin, Chunhua Shen, Anton van den Hengel, and Ian Reid. Efficient piecewise training of deep structured models for semantic segmentation. In *CVPR*, 2016.

[Liu *et al.*, 2015] Wei Liu, Andrew Rabinovich, and Alexander C Berg. Parsenet: Looking wider to see better. *arXiv:1506.04579*, 2015.

[Long *et al.*, 2015] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[Luo *et al.*, 2017] Ping Luo, Guangrun Wang, Liang Lin, and Xiaogang Wang. Deep dual learning for semantic image segmentation. In *CVPR*, 2017.

[Ma *et al.*, 2016] Lin Ma, Zhengdong Lu, and Hang Li. Learning to answer questions from image using convolutional neural network. In *AAAI*, 2016.

[Pan *et al.*, 2017] Tianxiang Pan, Bin Wang, Guiguang Ding, and Jun-Hai Yong. Fully convolutional neural networks with full-scale-features for semantic segmentation. In *AAAI*, 2017.

[Papandreou *et al.*, 2015] George Papandreou, Liang-Chieh Chen, Kevin Murphy, and Alan L Yuille. Weakly-and semi-supervised learning of a dcnn for semantic image segmentation. *arXiv:1502.02734*, 2015.

[Rahman and Wang, 2016] Md Atiqur Rahman and Yang Wang. Optimizing intersection-over-union in deep neural networks for image segmentation. In *ISVC*, 2016.

[Shotton *et al.*, 2009] Jamie Shotton, John Winn, Carsten Rother, and Antonio Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *IJCV*, 81(1):2–23, 2009.

[Szegedy *et al.*, 2015] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

[Wang *et al.*, 2017] Panqu Wang, Pengfei Chen, Ye Yuan, Ding Liu, Zehua Huang, Xiaodi Hou, and Garrison Cottrell. Understanding convolution for semantic segmentation. *arXiv:1702.08502*, 2017.

[Wei *et al.*, 2014] Yunchao Wei, Wei Xia, Junshi Huang, Bingbing Ni, Jian Dong, Yao Zhao, and Shuicheng Yan. Cnn: Single-label to multi-label. *arXiv:1406.5726*, 2014.

[Yi *et al.*, 2017] Zili Yi, Hao Zhang, Ping Tan Gong, et al. Dualgan: Unsupervised dual learning for image-to-image translation. *arXiv:1704.02510*, 2017.

[Zhang *et al.*, 2017] Wei Zhang, Xiaodong Yu, and Xuanyu He. Learning bidirectional temporal cues for video-based person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 2017.

[Zhang *et al.*, 2018] Wei Zhang, Qi Chen, Weidong Zhang, and Xuanyu He. Long-range terrain perception using convolutional neural networks. *Neurocomputing*, 275:781–787, 2018.

[Zhao *et al.*, 2016] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. *arXiv:1612.01105*, 2016.

[Zhu *et al.*, 2017] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv:1703.10593*, 2017.