Weakly Supervised Scene Parsing with Point-based Distance Metric Learning

Rui Qian[1,3], Yunchao Wei[1*], Honghui Shi[2,1], Jiachen Li[1], Jiaying Liu[3] and Thomas Huang[1]
[1]IFP Group, Beckman Institute, UIUC, [2]IBM Research, [3]Peking University
{qianrui,liujiaying}@pku.edu.cn, {yunchao,hshi10,jiachenl,t-huang1}illinois.edu

Abstract

Semantic scene parsing is suffering from the fact that pixel-level annotations are hard to be collected. To tackle this issue, we propose a Point-based Distance Metric Learning (PDML) in this paper. PDML does not require dense annotated masks and only leverages several labeled points that are much easier to obtain to guide the training process. Concretely, we leverage semantic relationship among the annotated points by encouraging the feature representations of the intra- and inter-category points to keep consistent, *i.e.* points within the same category should have more similar feature representations compared to those from different categories. We formulate such a characteristic into a simple distance metric loss, which collaborates with the point-wise cross-entropy loss to optimize the deep neural networks. Furthermore, to fully exploit the limited annotations, distance metric learning is conducted across different training images instead of simply adopting an image-dependent manner. We conduct extensive experiments on two challenging scene parsing benchmarks of PASCAL-Context and ADE 20K to validate the effectiveness of our PDML, and competitive mIoU scores are achieved.

Introduction

Scene parsing is of critical importance in mid- and high-level vision. Various machine vision applications lie on the basis of detailed and accurate scene analysis and segmentation *i.e.* outdoor driving system (Ess et al. 2009), image retrieval (Wan et al. 2014) and editing (Tsai et al. 2016). To achieve the goal of successfully recognizing objects in common scenes, datasets with careful and comprehensive labeling such as the PASCAL-Context (Mottaghi et al. 2014) and ADE 20K (Zhou et al. 2017) are putting forward to the community at great expense. Compared with image semantic segmentation where many semantic regions are coarsely divided into a unified background class, pixel-wise annotations for scene parsing datasets require much more efforts as more fine-grained regions (*e.g.* wall and ground) need to be specified manually. Though current state-of-the-art methods such as PSPNet (Zhao et al. 2017) and Deeplab (Chen et al. 2018) have already obtained impressive performance on these datasets, the limitations of relying on densely-annotated data would be magnified as dozens of new circumstances *i.e.* autonomous driving in the wild, drone navigation coming into horizon. It is both time consuming and expensive to carefully annotate a brand new dataset for each specific application.

Attempts have been made to alleviate the annotation burden by introducing multiple weakly supervised formats *i.e.* image-level supervision (Pathak et al. 2014), box supervision (Dai, He, and Sun 2015) and scribble supervision (Lin et al. 2016). In the mean time, as suggested in (Bearman et al. 2016), to annotate a single pixel for each instance is a natural scheme for human reference and the cost could be greatly alleviated. Furthermore, Bearman's work tackles the easier semantic segmentation and focuses more on analyzing the effectiveness of the point-based scheme itself by analyzing the annotation quality and taking the interaction with annotators as an important factor in the loss function. We believe that the potential of point-based supervision has not been well explored and also, the more challenging task of point-based scene parsing has remained untouched.

Thus, this paper serves as an initial attempt to explore the possibility of point-guided weakly supervised scene parsing. Being given only one semantic annotated point per instance, we propose a novel point-based distance metric learning method (PDML) to tackle this challenging task. PDML leverages semantic relationship among the annotated points by encouraging the feature representations of the intra- and inter-category points to keep consistent. Points within the same category are optimized to share more similar feature representations and oppositely, features of points from different categories are optimized to be more distinct. We implement this optimizing procedure by utilizing a distance metric loss, which collaborates with the point-wise cross-entropy loss to optimize the whole deep neural network. More important, different from current weakly supervised methods whose solutions are constrained in a single image, we conduct distance metric learning across different training images, so that the limited human annotated points can be fully exploited. Extensive experiments are performed on two challenging scene parsing benchmarks: PASCAL-Context and ADE 20K. We achieve the mIoU score of 30.0% on PASCAL-Context, which is impressive compared to the result of 39.6% from fully supervised scheme by using only 7.2×10^{-5} the number of annotated pixels. And we achieve

[*] Corresponding author.

mIoU of 19.6% on ADE 20K, while the SegNet (Badrinarayanan, Kendall, and Cipolla 2017) produces roughly 21% under full annotation of this dataset.

In conclusion, the major contributions of this paper lie in the following aspects:

- We are the first to deal with the task of point-based weakly supervised scene parsing.
- We propose a novel deep metric learning method PDML which optimizes the intra- and inter-category embedding feature consistency among the annotated points.
- PDML is performed across different training images to fully exploit the limited annotations, which is very novel compared to traditional intra-image methods.
- Our method has competitive performance both qualitatively and quantitatively on PASCAL-Context and ADE 20K scene parsing dataset.

Related Work

Weakly Supervised Semantic Segmentation

Image-level Annotations Image-level annotations by just naming the objects and staff are easy and natural to obtain. (Pathak et al. 2014) explore multiple instance learning for semantic segmentation. (Papandreou et al. 2015) propose dynamic prediction of foreground and background by using Expectation-Maximization algorithm. (Kolesnikov and Lampert 2016) introduce the new loss function of seed, expand and constrain. (Wei et al. 2017b) utilize a simple to complex framework to further improve the performance. And object region mining and localization is experimented in (Wei et al. 2016), (Wei et al. 2017a), (Zhang et al. 2018b), (Zhang et al. 2018c) and (Hou et al. 2018). And (Wei et al. 2018) discuss the effect of dilated convolution in this task. However, the methods listed below all focus on object-based semantic segmentation while no attempts have been made to deal with the more difficult scene parsing. Recently (Zhang et al. 2018a) tackle description guided scene parsing but investigating on parsing a scene image into a structured configuration.

Region-specified Annotations Compared to image-level annotations where no explicit location related information are given, multiple attempts have been made to provide various region-specified semantic supervision. Annotated bounding boxes are utilized in (Dai, He, and Sun 2015) and (Papandreou et al. 2015). (Lin et al. 2016) use scribbles as the supervision information and graphic models are used in optimization. Furthermore, (Bearman et al. 2016) have a similar setup with us but target at semantic segmentation. Also, they focus more on analyzing point supervision regime itself by comparing the annotation time, error rate as well as quality with other supervision regimes. Their method takes the confidence of annotators as a parameter in the loss function and transfers image-level annotations to objectness priors by utilizing another model pretrained on non-overlapping datasets. In comparison, we put attention on the more difficult task of scene parsing. We do not use any additional data and focus on exploring the cross-image semantic relations to boost the scene parsing performance.

Deep Metric Learning

Deep metric learning on embedding features has been explored in various tasks such as image query-and-retrieval (Oh Song et al. 2016), face recognition (Schroff, Kalenichenko, and Philbin 2015) and verification (Ming et al. 2017). It has also been applied in semantic instance segmentation (Fathi et al. 2017) and grouping (Kong and Fowlkes 2018). More recently, (Liu et al. 2018) use metric learning as a fast and efficient way for training a semantic segmentation network.

Proposed Method

Motivation

Compared to full, box and scribble supervision regime, point-based supervision data is most natural for human reference and easiest to obtain. However, as illustrated in Table 1, the annotation number of pixels in single image is too tiny to train a neural network efficiently.

Annotation Method	Full	ScribbleSup	PointSup
Anno. pixel/image	170k	1817.48	**12.26**

Table 1: Comparison of the average annotated pixel number per image of different methods on PASCAL-Context training dataset.

Consider the limitation of supervision information and inspired by recent deep metric learning methods *e.g.* (Liu et al. 2018), we focus on exploring the relationship between feature representations of annotated pixels. While all current methods try to optimize embedding feature distances within a single image, we apply a novel method by forming triples from the embedding vectors of annotated pixels across images and optimizing the feature consistency within the triples by distance metric learning. In each triple, two embedding vectors belong to the same category and we name them as a positive pair. The other is from a different category and it forms a negative pair with one element in the positive pair. We minimize the distances between positive pairs and maximize those between negative pairs on inter-image level. There are at least two reasons for doing so:

- Objects and stuff which have similar feature representations before the classification module would be more likely to be specified into the same class. Oppositely, embedding features from different categories, if being distinct enough with each other, are more easier for the classifier to distinguish.
- Under the point-based regime, most annotated pixels in one image come from different categories. Simply optimizing distances between negative pairs would not help training. While extending to inter-image level, balanced number of positive and negative pairs can be obtained.

Furthermore, different from image-level weakly supervised methods relying heavily on the saliency maps generated by pretrained specific models, scribble and box guided methods depending on being optimized iteratively, our method does not require any additional data and can be learned in an end-to-end manner.

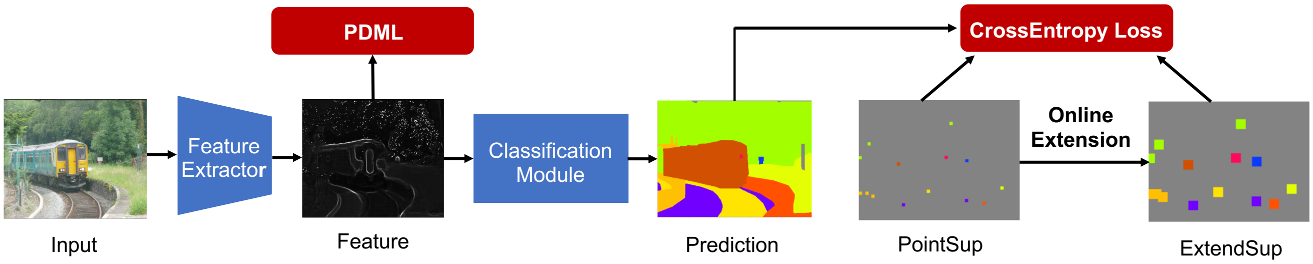


Figure 1: The pipeline of our proposed algorithm.

Overview

Figure 1 shows the pipeline of our algorithm. Similar to Deeplab (Chen et al. 2018), we utilize ResNet101 (He et al. 2016) pretrained on ImageNet as the backbone of our feature extractor. Atrous convolutions are adopted to increase the receptive field and reduce the degree of signal downsampling. Given a batch of input images, we first use the feature extractor to form deep embedding features. The features are then assigned to two streams. The first stream is feeding into the point-based distance metric learning module where the embedding feature vectors across different images are optimized towards learning representation consistency. The second stream is feeding into a fully-convolutional classification module to generate the pixel-wise prediction. At the mean time, online extension is performed to dynamically gather pixels with high classification confidence and tight spatial relationship to the original annotated pixels to form the extended label. The point-wise cross-entropy loss is calculated between the classification results and the two labels.

Point Supervision

Point supervision (PointSup) serves as the baseline of our method. Let $\mathcal{S}_p$ donate the set of training images with point-wise annotations. Then we have $\mathcal{S}_p = \bigcup_{w=1}^{N}\{(I_w, M_w)\}$ where I_w is the wth image in the training set. M_w is the corresponding pseudo annotated mask, where only several pixels are annotated with semantic labels. Let $g(f(I_w;\theta_f);\theta_g)$ donate our segmentation module where (f,θ_f), (g,θ_g) refer to the feature extraction module and its parameters, classification module and its parameters, respectively. The objective function is

$$\min_{\theta_f,\theta_g} \sum_{w=1}^{N} J_w(g(f(I_w;\theta_f);\theta_g)), \tag{1}$$

and then consider the class label set of the current image as $\mathcal{C}$, let the the conditional probability generated by the classification module of any label $c \in \mathcal{C}$ at the any location u as $g_{u,c}(f(I_w;\theta_f);\theta_g)$, then we get

$$J_w(g(f(I_w;\theta_f);\theta_g)) = -\frac{\sum_{c\in\mathcal{C}}\sum_{u\in M_w^c} \log g_{u,c}(f(I_w;\theta_f);\theta_g)}{\sum_{c\in\mathcal{C}} |M_w^c|}, \tag{2}$$

where $|*|$ refers to number of annotated pixels in $*$.

Distance Metric Learning

To optimize feature representations of annotated pixels to keep consist, we aim at minimizing distances between positive pairs and maximizing those between negative pairs. However, it is hard to find positive pairs within a single image. Only optimizing negative pairs would do no help but make the loss hard to converge. By extending to inter-image level, we could obtain balanced number of two kinds of pairs as illustrated in Figure 2. Let s be a subgroup of the training set $\mathcal{S}_p$, then $s = \{(I_a, M_a), (I_b, M_b), ...\}$. For each image *e.g.* I_a in the subgroup, we could define the embedding vector set $E_a = \bigcup_{i=1}^{|M_a|}\{P_{ai}\}$, where P_{ai} is the feature vector corresponding to the ith annotated pixel in image a. Suppose for three different feature vectors P_{ai}, P_{bj}, P_{bk}, where P_{ai} shares the same category with P_{bj} and different from P_{bk}, we apply the loss L_t as

$$L_t(P_{ai}, P_{bj}, P_{bk}) = \alpha L_p(P_{ai}, P_{bj}) + \beta L_n(P_{ai}, P_{bj}, P_{bk}), \tag{3}$$

where $L_p(P_{ai}, P_{bj})$, corresponding to the red dotted line in Figure 2(b) (*i.e.* (P_{a1}, P_{b1})), can be expressed as

$$\|P_{ai} - P_{bj}\|_2, \tag{4}$$

which aims at minizing the L2-norm distance between same-category embedding vectors. And $L_n(P_{ai}, P_{bj}, P_{bk})$, corresponding to the combination of one gray and one red dotted line in Figure 2(b) (*i.e.* (P_{a1}, P_{b1}) and (P_{a1}, P_{b2})), can be expressed as

$$\max(\|P_{ai} - P_{bj}\|_2 - \|P_{ai} - P_{bk}\|_2 + m, 0), \tag{5}$$

which aims at maximizing the gap between $\|P_{ai} - P_{bj}\|_2$ and $\|P_{ai} - P_{bk}\|_2$. m is a constant value and only when the gap is within this value would the the triple embedding vectors be optimized. Two hyper-parameters α and β are used to balance the the effect of L_p and L_n. We set $m = 20$, $\alpha = 0.8$, $\beta = 1$ in practice. And the algorithm of optimizing dense PDML is expressed in Algorithm 1.

Online Extension

To further improve the performance, we put attention on gathering more pixels during the training process. Previous works such as superpixel (Achanta et al. 2012) and K-Means clustering (Ray and Turi 1999) have explored the possibility of gathering pixels by measuring the similarity of low-level features. However, both methods have obvious drawbacks under the point-based scene parsing regime by gathering lots of wrong pixels. And from experiments we find

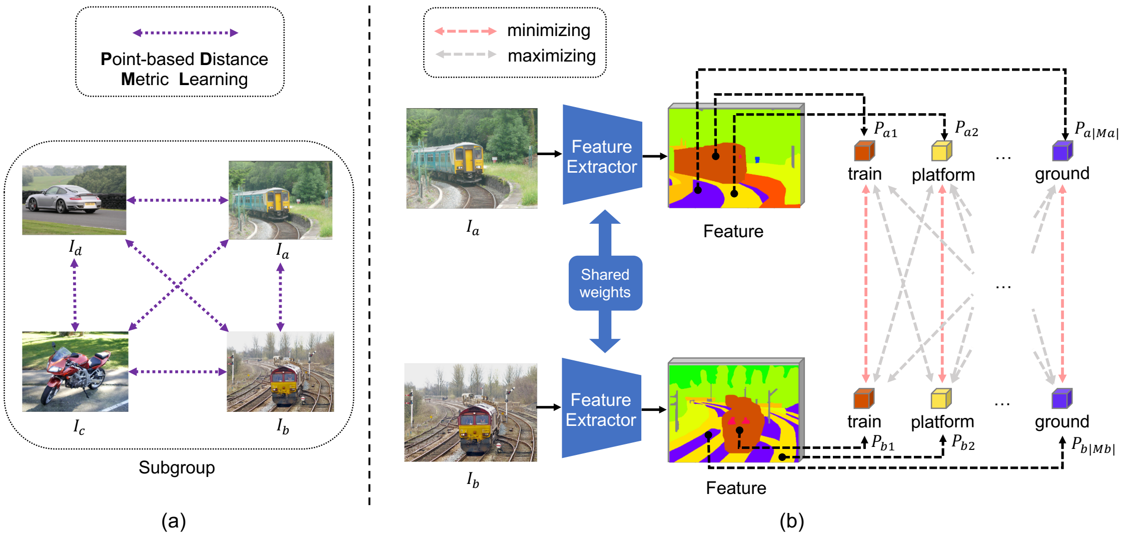


Figure 2: The architecture of our dense PDML. As shown in part (a), we first densely form image pairs in small subgroups divided from the training dataset. Then for each image pairs, as shown in part (b), we perform densely distance optimizing between the embedding vectors of annotated pixels. Best viewed in color.

Algorithm 1: Optimizing Procedure for PDML

```
while $|S_p| \geq 0$ do
    Extract supgroup $s$ from $S_p$;
    for $(I_n, M_n) \in s$ do
        for $(I_m, M_m) \in s - (I_n, M_n)$ do
            $Generate\ E_n, E_m$
            for $P_{ni} \in E_n$ do
                $E_{pos} = \{\}, E_{neg} = \{\}$
                for $P_{mj} \in E_m$ do
                    if $class(P_{mj}) = class(P_{ni})$ then
                        $E_{pos} = E_{pos} \cup \{P_{mj}\}$
                    else
                        $E_{neg} = E_{neg} \cup \{p_{mj}\}$
                    end
                end
                while $|E_{pos}| \geq 0\ and\ |E_{neg}| \geq 0$ do
                    $select\ P_{pos} \in E_{pos}, P_{neg} \in E_{neg}$
                    $\min L_t(P_{ni}, P_{pos}, P_{neg})$
                    $E_{pos} = E_{pos} - \{P_{pos}\}$
                    $E_{neg} = E_{neg} - \{P_{neg}\}$
                end
            end
        end
    end
    $S_P = S_P - s$
end
```

that wrong pixels would greatly degrade the performance of the network. More detailed analyses would be found in the section of discussion. To tackle this issue, we adopt a simple but accurate online extension method to collect more pixels with low false positive rate to extend the annotation data. Take $(I_a, M_a) \in S_p$ as an example, we now assume the current weights of feature extractor f and classification module g as θ_{f1} and θ_{g1}, then we could specify the new label candidate M_{a_score} by judging the pixel-wise classification score:

$$M_{a\_score} = \bigcup_{u} \{\max(g_{u,c \in C}(f(I_a;\theta_{f1});\theta_{g1})) > thr\}, \quad (6)$$

where thr is a threshold to filter the pixels. In other words, for every location u in the input image, if the max classification score of this pixel is greater than the threshold and the corresponding class is within the class label set of this image, it would be chosen for the extended label. From another perspective, we believe that pixels with close spatial distance to the annotated pixels are more likely from the same category. Thus we extend each annotated pixel in M_a to a 5×5 square to form the second label candidate M_{a_region}. Then we generate the final extended label M_{a_extend} by just selecting the candidates appearing in both schemes as:

$$M_{a\_extend} = M_{a\_score} \cap M_{a\_region}. \quad (7)$$

Experimental Results

Implementation Details

Dataset and Evaluation Metrics Our proposed model is trained and evaluated on two challenging scene parsing datasets: PASCAL-Context (Mottaghi et al. 2014) and ADE 20K (Zhou et al. 2017), as shown in Table 2. In PASCAL-Context, the most frequent 59 classes are used and others are divided into a unified background class. In ADE 20K, we adopt the annotation of 150 meaningful classes. We generate one pixel annotation for each instance in each image. The performances are quantitatively evaluated by pixel-wise accuracy and mean region intersection over union (mIoU).

Dataset	# Training	# Eval	Pixel/Image
PASCAL-Context	4998	5105	12.26
ADE20K	20210	2000	13.96

Table 2: Statistical results of two datasets. **Pixel/Image** illustrates the average number of pixels annotated in one image in the training dataset.

Training Setting We utilize ResNet101 (He et al. 2016) with the modification of atrous convolution as the backbone of our feature extractor. Weights pretrained on ImageNet are adopted to initialize. During training, we take a mini-batch of 16 images and randomly crop patches of the size of 321 × 321 from original images. We use the optimizer of SGD where momentum is set to 0.9 and the weight decay is 0.0005. The initial base learning rate is set to 0.00025 for parameters in the feature extraction layers and ten times for parameters in the classification module. Both learning rate will be decayed under the scheme of $base\_lr*(1-\frac{epoch}{max\_epoch})^{0.8}$. All the experiments are conducted on two NVIDIA V100 GPUs. Our code will be available publicly.

Quantitative and Qualitative Results

We evaluate different methods quantitatively by using pixel accuracy and mIoU which describes the the precision of prediction and the average performance among all classes, respectively. We run multiple experiments to determine the effects of the three parts of our proposed method: PointSup, PDML and online extension. To make a more comprehensive understanding of the effect of our method, we also train our network with fully-annotated label. The quantitative results are shown in Tables 3 and 6 (note we omit % of mIoU for simplicity). On PASCAL-Context, the full supervision setting could yield a 39.6% mIoU and our method, with only 7.2×10^{-5} the number of annotated pixels, could obtain the 30.0% mIoU performance. On ADE 20K, we achieve the mIoU of 19.6%, while the SegNet (Badrinarayanan, Kendall, and Cipolla 2017) achieves roughly 21% mIoU under full supervision. And the qualitative results are shown in Figures 5 and 6. Our final method combining point supervision, distance metric learning and online extension has the best scene parsing quality subjectively.

Discussion

Analysis of PDML

Loss function Recall that we apply the loss $L_t = \alpha L_p + \beta L_n$ to optimize the consistency of embedding vectors. We name the distance between the positive pairs as $dis(+)$ and that between negative pair as $dis(-)$. L_p aims at constraining $dis(+)$ and L_n aims at increasing the gap between $dis(+)$ and $dis(-)$. We argue that L_p is very import in the whole scheme, as shown in Figure 3. With only applying L_n, though the gap between $dis(+)$ and $dis(-)$ is optimized to be larger, the absolute values increase greatly at the mean time which leads to great performance drop. And by applying the constrain of L_p, the absolute values of distances remain at the normal scale and the gap is also optimized to make it easier for the classification module to distinguish.

Method				Metrics	
FullSup	PointSup	PDML	Online Ext.	mIOU	Pixel Acc
PASCAL-Context validation dataset					
√				39.6	78.6%
	√			27.9	55.3%
	√	√		29.7	57.5%
	√	√	√	**30.0**	**57.6%**
ADE 20K validation dataset					
√				33.9	75.8%
	√			17.7	58.0%
	√	√		19.0	59.0%
	√	√	√	**19.6**	**61.0%**

Table 3: Quantitative results on PASCAL-Context and ADE 20K validation dataset.

Also, we visualize the distribution of the pair numbers of each distance value in training procedure to demonstrate the effectiveness of L_t. At the first epoch, the distributions of distances of positive and negative pairs are almost symmetrical. During the training process, an obvious peak shift can be observed. The peak of the distances between positive pairs moves towards the origin of the coordinate axis and peak of the distances between negative pairs moves oppositely.

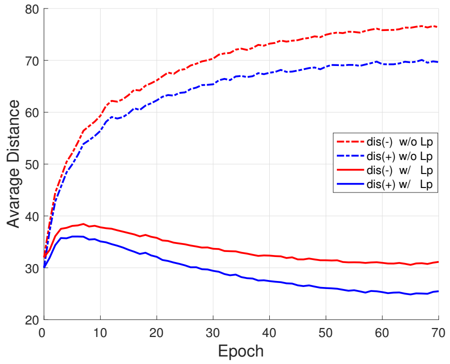
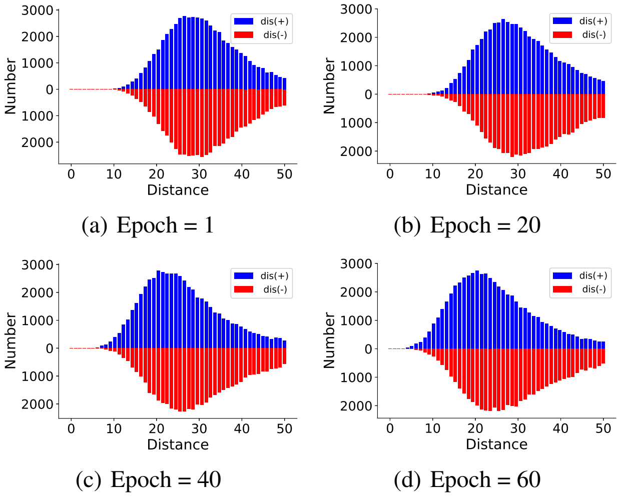


Figure 3: Comparison of the average value of $dis(+)$ and $dis(-)$ on whether using L_p constrain in the loss function.

Hyper-parameters We have three hyper-parameters in the loss function: m, α, β. Margin m is set for a mining purpose. A small margin would limit the number of optimizing embedding vector triples and make it less effective for classification. And a large margin size, bringing too many triples for optimization, would cause the training loss being hard to converge. A moderate margin value of 20 would produce the best performance. Loss weights α and β are used to balance the effect of L_p and L_n. We set β to 1 and adjust α correspondingly. A small α, similar to Figure 3, can not constrain the absolute value of $dis(+)$ and $dis(-)$ to remain at the normal scale and would lead to poor performances.



(a) Epoch = 1

(b) Epoch = 20

(c) Epoch = 40

(d) Epoch = 60

Figure 4: Summary of the distribution of the pair numbers of each distance value. An obvious peak shift can be observed.

While a large α would weaken the effect of maximizing the distance between $dis(+)$ and $dis(-)$ and make it hard to distinguish embedding vectors from different categories. A moderate value of 0.8 achieves the best performance.

Hyper-parameters			Metrics	
α	β	m	mIoU	Pixel Acc
0.8	1	8	29.2	57.2%
0.8	1	12	29.5	57.0%
0.8	1	16	29.4	57.1%
0.8	1	**20**	**29.7**	**57.5%**
0.8	1	24	29.4	56.9%
0.0	1	20	26.2	53.5%
0.4	1	20	28.6	55.8%
0.8	1	20	**29.7**	**57.5%**
1.2	1	20	29.5	57.1%
1.6	1	20	29.4	57.0%

Table 4: Comparisons of the performance of different hyper-parameter settings on PASCAL-Context validation dataset.

Analysis of Extension Method

We compare our online label extension method with other frequently used clustering method.

Superpixel Superpixel is used to cluster pixels by taking low-level feature similarity into consideration. We set the number of superpixels in the range of 50 to 200 per image depending on the number of annotated pixels. Each annotated pixel would be extended to the corresponding superpixel covering it.

K-Means We perform K-Means clustering on the feature representations of the input images. The annotated pixels are set to be the initial clustering centers and we set the maximum iteration time to be 300.

Score and Region Recall in the online extension part, we use two methods to generate new label candidates. The first is using a score thresholding and we name this extension method as **score**. And another is simply extending every pixel in the current label to a 5×5 square with the original one as the center. We name this method as **region**.

On the basis of the weight obtained by PDML, we implement different extension methods and their results are shown in Table 5. The pixel-wise extension accuracy is of critical importance in influencing the performance. Superpixel and region have better extension accuracies and have better performances correspondingly. We also test various thresholds for the score scheme. Our method taking both score with the threshold of 0.7 and region into consideration has an accuracy of 98.2% and the best testing performance.

Extension Method					Metrics	
Superpixel	K-Means	Score	Region	Extension Acc	mIoU	Pixel Acc
√				83.6%	26.7	52.9%
	√			12.1%	12.9	36.3%
		√(0.7)		56.9%	16.7	40.3%
			√	97.6%	29.7	57.5%
		√(0.5)	√	97.7%	29.7	57.5%
		√(0.6)	√	98.0%	29.8	57.4%
		√(0.7)	√	**98.2%**	**30.0**	**57.6%**
		√(0.8)	√	98.3%	29.4	57.1%
		√(0.9)	√	98.4%	29.4	56.9%

Table 5: Comparisons of different extension method on PASCAL-Context validation dataset.

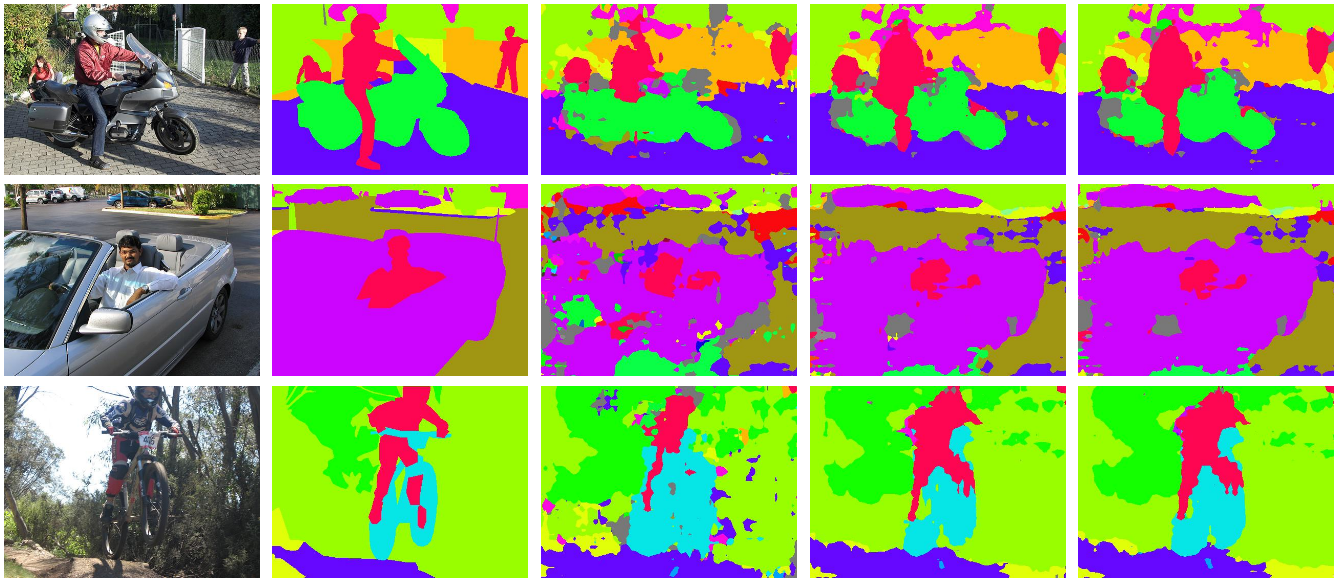
Conclusion

This paper is the first to tackle the task of point-based semantic scene parsing. We propose a novel deep metric learning method to leverage semantic relationship among the annotated points by encouraging the feature representation of the intra- and inter-category points to keep consistent. Points within the same category are optimized to share more similar feature representations and oppositely, those of points from different categories are optimized to be more distinct. Different from all current weakly supervised methods whose solutions are constrained in a single image, our proposed method focuses on optimizing the embedding vectors across different images in the training dataset to obtain sufficient balanced embedding vector pairs. The whole model can be trained in an end-to-end manner. Our method has competitive performance both qualitatively and quantitatively on PASCAL-Context and ADE 20K scene parsing datasets.

Acknowledgements. This work is supported by IBM-ILLINOIS Center for Cognitive Computing Systems Research (C3SR) - a research collaboration as part of the IBM AI Horizons Network, and by Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/ Interior Business Center (DOI/IBC) contract number D17PC00341. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/IBC, or the U.S. Government. Also, This work was supported by National Natural Science Foundation of China under contract No. 61772043 and Peking University Tencent Rhino Bird Innovation Fund.

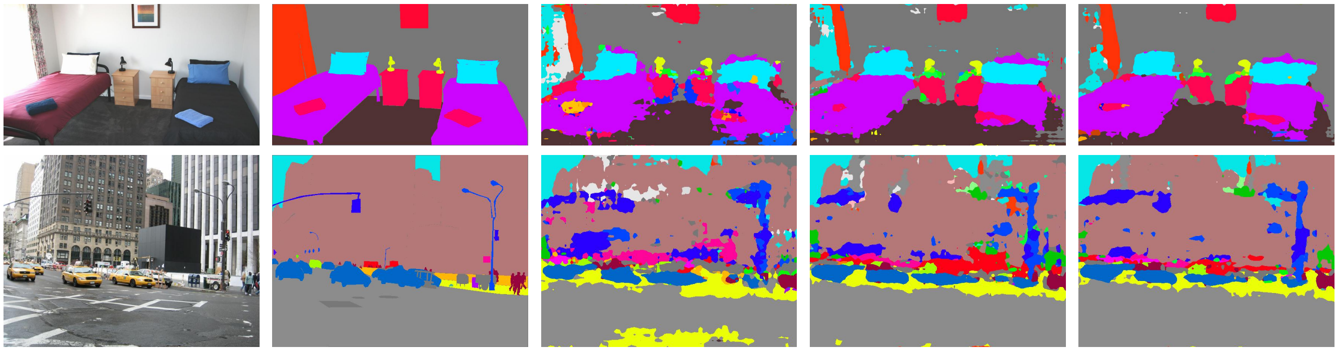
PointSup	**PDML**	**Onlin Ext.**	plane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow
√			41.8	39.8	52.7	32.6	39.3	50.3	48.9	64.5	17.3	**44.0**
√	√		**47.3**	**46.9**	55.1	34.0	46.6	50.7	**51.4**	67.1	**21.9**	40.1
√	√	√	45.7	44.5	**56.7**	**34.2**	**46.8**	**50.9**	50.2	**67.1**	21.3	41.8
table	dog	horse	motor	person	potplant	sheep	sofa	train	tv	book	building	cabinet
20.4	56.7	44.5	43.4	56.1	30.3	46.6	26.5	44.5	35.3	17.2	31.0	12.3
22.6	58.0	47.2	**46.6**	**59.1**	29.8	46.6	**30.0**	49.5	41.7	**18.8**	35.5	14.0
22.5	**58.3**	**47.5**	44.8	58.3	**31.5**	**47.7**	29.5	**49.6**	**43.5**	17.5	**35.6**	**14.4**
ceiling	cup	fence	floor	food	grass	ground	keyboard	light	mountain	mouse	curtain	platform
31.5	11.2	17.0	30.6	27.4	52.3	31.1	31.0	12.4	29.0	**13.9**	17.6	20.8
31.7	13.2	20.2	34.3	29.4	59.1	**33.3**	30.2	**16.2**	**29.3**	9.6	17.4	17.0
32.0	**13.7**	**20.2**	**34.4**	**30.4**	**59.7**	32.3	**31.3**	14.5	28.8	11.2	**18.9**	**20.9**
sign	plate	road	rock	sky	snow	bedclothes	track	tree	wall	water	window	**mIoU**
10.4	16.7	31.3	20.5	71.8	23.3	13.1	38.1	50.4	36.2	58.5	15.7	27.9
13.0	19.6	31.4	22.8	**74.1**	26.6	12.6	39.4	52.6	38.8	58.8	18.9	29.7
12.8	**20.1**	**33.7**	**22.8**	73.9	**27.1**	**13.2**	**41.4**	**52.6**	**39.4**	**59.0**	**19.4**	**30.0**

Table 6: Class-wise quantitative comparisons on PASCAL-Context validation dataset.



(a) Image (b) Ground Truth (c) PointSup (d) PointSup+PDML (e) Our method

Figure 5: Qualitative comparison of different methods on PASCAL-Context validation dataset.



(a) Image (b) Ground Truth (c) PointSup (d) PointSup+PDML (e) Our method

Figure 6: Qualitative comparison of different methods on ADE 20K validation dataset.

References

Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; Süsstrunk, S.; et al. 2012. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE TPAMI* 34(11).

Badrinarayanan, V.; Kendall, A.; and Cipolla, R. 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE TPAMI* 39(12):2481–2495.

Bearman, A.; Russakovsky, O.; Ferrari, V.; and Fei-Fei, L. 2016. Whats the point: Semantic segmentation with point supervision. In *ECCV*, 549–565. Springer.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2018. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE TPAMI* 40(4):834–848.

Dai, J.; He, K.; and Sun, J. 2015. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *IEEE ICCV*, 1635–1643.

Ess, A.; Müller, T.; Grabner, H.; and Van Gool, L. J. 2009. Segmentation-based urban traffic scene understanding. In *BMVC*, volume 1, 2. Citeseer.

Fathi, A.; Wojna, Z.; Rathod, V.; Wang, P.; Song, H. O.; Guadarrama, S.; and Murphy, K. P. 2017. Semantic instance segmentation via deep metric learning. *arXiv preprint arXiv:1703.10277*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *IEEE CVPR*, 770–778.

Hou, Q.; Jiang, P.-T.; Wei, Y.; and Cheng, M.-M. 2018. Self-erasing network for integral object attention. In *NIPS*.

Kolesnikov, A., and Lampert, C. H. 2016. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *ECCV*, 695–711. Springer.

Kong, S., and Fowlkes, C. 2018. Recurrent pixel embedding for instance grouping. In *IEEE CVPR*, 9018–9028.

Lin, D.; Dai, J.; Jia, J.; He, K.; and Sun, J. 2016. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *IEEE CVPR*, 3159–3167.

Liu, Y.; Jiang, P.-T.; Petrosyan, V.; Li, S.-J.; Bian, J.; Zhang, L.; and Cheng, M.-M. 2018. Del: Deep embedding learning for efficient image segmentation. In *IJCAI*, 864–870.

Ming, Z.; Chazalon, J.; Luqman, M. M.; Visani, M.; and Burie, J.-C. 2017. Simple triplet loss based on intra/inter-class metric learning for face verification. In *IEEE ICCVW*, 1656–1664.

Mottaghi, R.; Chen, X.; Liu, X.; Cho, N.-G.; Lee, S.-W.; Fidler, S.; Urtasun, R.; and Yuille, A. 2014. The role of context for object detection and semantic segmentation in the wild. In *IEEE CVPR*.

Oh Song, H.; Xiang, Y.; Jegelka, S.; and Savarese, S. 2016. Deep metric learning via lifted structured feature embedding. In *IEEE CVPR*, 4004–4012.

Papandreou, G.; Chen, L.-C.; Murphy, K.; and Yuille, A. L. 2015. Weakly-and semi-supervised learning of a dcnn for semantic image segmentation. *arXiv preprint arXiv:1502.02734*.

Pathak, D.; Shelhamer, E.; Long, J.; and Darrell, T. 2014. Fully convolutional multi-class multiple instance learning. *arXiv preprint arXiv:1412.7144*.

Ray, S., and Turi, R. H. 1999. Determination of number of clusters in k-means clustering and application in colour image segmentation. In *Proceedings of the 4th international conference on advances in pattern recognition and digital techniques*, 137–143.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *IEEE CVPR*, 815–823.

Tsai, Y.-H.; Shen, X.; Lin, Z.; Sunkavalli, K.; and Yang, M.-H. 2016. Sky is not the limit: semantic-aware sky replacement. *ACM Trans. Graph.* 35(4):149–1.

Wan, J.; Wang, D.; Hoi, S. C. H.; Wu, P.; Zhu, J.; Zhang, Y.; and Li, J. 2014. Deep learning for content-based image retrieval: A comprehensive study. In *ACM MM*.

Wei, Y.; Liang, X.; Chen, Y.; Jie, Z.; Xiao, Y.; Zhao, Y.; and Yan, S. 2016. Learning to segment with image-level annotations. *Pattern Recognition* 59:234–244.

Wei, Y.; Feng, J.; Liang, X.; Cheng, M.-M.; Zhao, Y.; and Yan, S. 2017a. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *IEEE CVPR*, volume 1, 3.

Wei, Y.; Liang, X.; Chen, Y.; Shen, X.; Cheng, M.-M.; Feng, J.; Zhao, Y.; and Yan, S. 2017b. Stc: A simple to complex framework for weakly-supervised semantic segmentation. *IEEE TPAMI* 39(11):2314–2320.

Wei, Y.; Xiao, H.; Shi, H.; Jie, Z.; Feng, J.; and Huang, T. S. 2018. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In *IEEE CVPR*, 7268–7277.

Zhang, R.; Lin, L.; Wang, G.; Wang, M.; and Zuo, W. 2018a. Hierarchical scene parsing by weakly supervised learning with image descriptions. *IEEE TPAMI* (1):1–1.

Zhang, X.; Wei, Y.; Feng, J.; Yang, Y.; and Huang, T. 2018b. Adversarial complementary learning for weakly supervised object localization. In *IEEE CVPR*.

Zhang, X.; Wei, Y.; Kang, G.; Yang, Y.; and Huang, T. 2018c. Self-produced guidance for weakly-supervised object localization. In *ECCV*.

Zhao, H.; Shi, J.; Qi, X.; Wang, X.; and Jia, J. 2017. Pyramid scene parsing network. In *IEEE CVPR*, 2881–2890.

Zhou, B.; Zhao, H.; Puig, X.; Fidler, S.; Barriuso, A.; and Torralba, A. 2017. Scene parsing through ade20k dataset. In *IEEE CVPR*.