UNIVERSAL WEAKLY SUPERVISED SEGMENTATION BY PIXEL-TO-SEGMENT CONTRASTIVE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

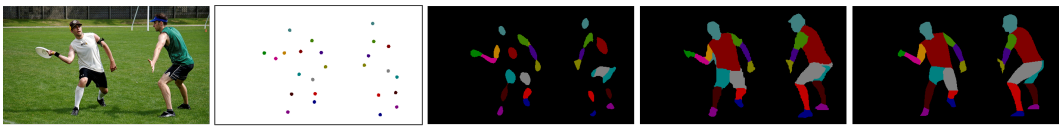
Weakly supervised segmentation is challenging as sparsely labeled pixels do not provide sufficient supervision: A semantic segment may contain multiple distinctive regions whereas adjacent segments may appear similar. Common approaches use the few labeled pixels in all training images to train a segmentation model, and then propagate labels within each image based on visual or feature similarity. Instead, we treat segmentation as a semi-supervised pixel-wise metric learning problem, where pixels in different segments are mapped to distinctive features. Naturally, our unlabeled pixels participate not only in data-driven grouping within each image, but also in discriminative feature learning within and *across* images. Our results on Pascal VOC and DensePose datasets demonstrate our substantial accuracy gain on various forms of weak supervision including image-level tags, bounding boxes, labeled points, and scribbles.

1 INTRODUCTION

Consider semantic segmentation given sparsely labeled pixels in Fig. 1: Each body part is labeled with a single seed pixel and the task is to segment out the entire person by individual body parts. This task is challenging, as not only a single body part could contain several visually distinctive areas, *e.g., head* consists of *eyes*, *nose*, *mouth*, *beard*), but two adjacent body parts could also have the same visual appearance, *e.g., upper arm*, *lower arm*, and *hand* have the same skin appearance.

Such a seeded segmentation problem belongs to a family of weakly supervised segmentation problems, where the task is to assign a semantic label to each pixel despite that only partial supervision is available during training. It addresses an extremely practical issue of acquiring as little human annotations as possible, instead of dense pixel-wise labeling for the sake of learning segmentation.

Such weak supervision takes many forms, from image-level tags (Kolesnikov & Lampert, 2016; Ahn & Kwak, 2018; Huang et al., 2018; Lee et al., 2019), bounding boxes (Dai et al., 2015; Khoreva et al., 2017; Song et al., 2019), keypoints (Bearman et al., 2016), to scribbles (Lin et al., 2016; Tang et al., 2018a;b), with an increasing level of weak supervision. Image tags only provide semantic information without any localization information, bounding boxes further outline the object sizes and locations, whereas keypoints and scribbles indicate the object foregrounds and rough outreach. They vary in the amount of and the spatial precision of partial annotations.



(a) Image (b) Annotation (c) Baseline (d) Ours (e) Ground Truth

Figure 1: The task of weakly supervised segmentation and result comparison. Given an image (a) and a single seed pixel labeling per body part (b), the goal is to segment out the entire body parts (e). The common baseline approach (c) is to propagate seed pixel labels within each image based on feature similarity. Our approach (d) treats segmentation as a semi-supervised pixel-wise metric learning problem and propagates labels by data-driven grouping and contrastive learning within and across images. Our segmentation is fuller and more accurate, approaching the ground-truth.

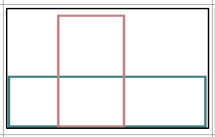
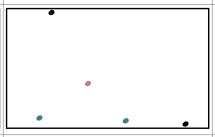
Image	Image Tags	Bounding Boxes	Labeled Points	Scribbles
	Person Motorbike			
SOTA Methods	CAM + iterative training	CAM of bounding boxes	CRF loss	CRF loss
Our Method	**single pixel-to-segment contrastive learning loss formulation**			
Relative Improvement	**+6.8%**	**+4.1%**	**+21.7%**	**+1.2%**

Figure 2: We propose a unified framework for weakly supervised semantic segmentation with different types of annotations. We demonstrate consistent performance gains compared to the state-of-the-art methods (Chang et al. (2020) for image tags, Song et al. (2019) for bounding boxes, and Tang et al. (2018b) for points and scribbles). For image tag and bounding box annotations, CAM (Zhou et al., 2016) is used to produce initial localized semantic cues as initial masks, followed by iterative training of segmentation networks. For labeled points and scribbles, CRF is used to propagate semantic labels to unlabeled regions based on low-level image similarity.

Weakly supervised semantic segmentation can be regarded as a semi-supervised pixel classification problem: Some pixels have labels, most don't, and the key is how to propagate annotations from coarsely and sparsely labeled pixels to unlabeled pixels.

Earlier works tackle this problem by generating pseudo dense supervision (Weston et al., 2012), and using image segmentation techniques such as graph cuts (Boykov et al., 2001), normalized cuts (Shi & Malik, 2000), and Conditional Random Fields (CRF) (Krähenbühl & Koltun, 2011). While such pseudo-supervision is imperfect, it effectively propagates labels from labeled pixels to unlabeled pixels within individual images, and a consensus from the entire image helps mitigate errors.

Recent works incorporate these ideas as an additional unsupervised loss on the feature learned for segmentation (Tang et al., 2018b): While seeded pixels receive labeling supervision, pixels in different segments shall have distinctive feature representations. Unlike traditional segmentation based on a fixed hand-crafted feature, with deep learning, seeded segmentation is no longer a seeded diffusion process in the pixel space, but a self-organizing discriminative learning in the feature space.

We propose a *Semi-supervised Pixel-wise Metric Learning* (SPML) model that can handle all these weak supervision varieties with a single pixel-to-segment contrastive learning formulation (Fig. 2). Instead of classifying pixels, our metric learning model learns a pixel-wise feature embedding based on common grouping relationships derived from any form of weak supervision.

Our key insight is to integrate unlabeled pixels into both supervised labeling and discriminative feature learning. They shall participate not only in data-driven grouping within each image, but also in discriminative feature learning *within* and more importantly *across* images. Intuitively, seed pixels are labeled not meant for themselves alone, but also for their surround pixels that share visual similarity. On the other hand, unlabeled pixels are not just passively brought into discriminative learning induced by sparsely labeled pixels, they themselves are organized based on bottom-up grouping cues (such as grouping by color similarity and separation by strong contours), and when they are examined *across* images, repeated patterns of frequent occurrences would also form a cluster that demand active discrimination from other patterns.

We capture the above insight in a single learning objective in the pixel-wise metric learning framework for segmentation, the goal of which is to map each pixel into a point in the feature space so that pixels in the same (different) semantic groups are close (far) in the feature space. Our model extends SegSort (Hwang et al., 2019) from its fully supervised and unsupervised segmentation settings to a universal weakly-supervised segmentation setting. With a single consistent feature learning criterion, such a model sorts pixels discriminatively within individual images and sorts segment clusters discriminatively across images, both steps minimizing the same feature discrimination loss.

Our experiments on Pascal VOC (Everingham et al., 2010) and DensePose (Alp Güler et al., 2018) demonstrate consistent gains over the state-of-the-art (SOTA), and the gain is substantial especially for the sparsest keypoint supervision.

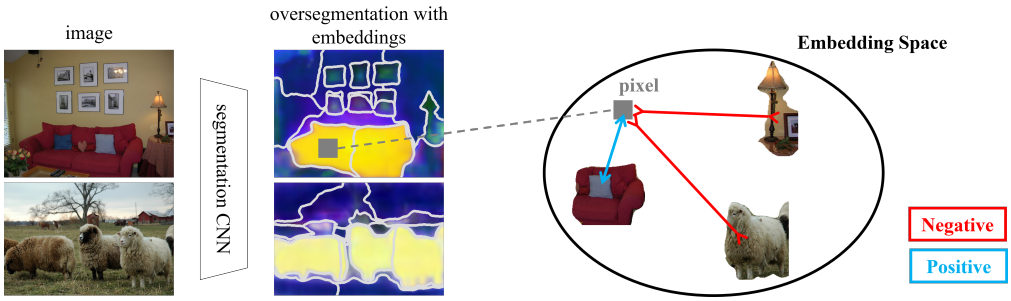


Figure 3: Overall method diagram. We develop pixel-wise embedding with contrastive learning between pixels and segments. We derive various forms of positive and negative segments for each pixel. Our goal is to attract (blue inward arrows) the pixel with positive segments, while repelling (red outward arrows) it from negative segments in the feature space.

2 Related Work

Semi-supervised learning. (Weston et al., 2012) treats it as joint learning with labeled and unlabeled data. The goal is to learn better feature representations by observing the common structures between labeled and unlabeled sets. Two lines of research have emerged. One is to leverage generative models (Kingma et al., 2014; Rasmus et al., 2015) for learning the underlying structure of unlabeled data. Such methods require good generative models and domain transfer methods to generate reliable training data. The other is to add regularization through consistency loss, *e.g.,* adversarial ensembling (Miyato et al., 2018), imitation learning and distillation (Tarvainen & Valpola, 2017), cross-view ensembling (Clark et al., 2018). Regularization approaches for weakly-supervised segmentation are most related to transductive learning (Joachims, 2003; Zhou et al., 2004; Fergus et al., 2009; Liu et al., 2019), where labels are propagated to unlabeled data via clustering in the pre-trained feature space. Our work does transductive learning in an adaptively learned feature space.

Weakly-supervised semantic segmentation. Partial annotations include scribbles (Lin et al., 2016; Tang et al., 2018a;b; Wang et al., 2019), bounding boxes (Dai et al., 2015; Khoreva et al., 2017; Song et al., 2019), points (Bearman et al., 2016), or image tags (Papandreou et al., 2015; Kolesnikov & Lampert, 2016; Ahn & Kwak, 2018; Huang et al., 2018; Li et al., 2018; Lee et al., 2019; Shimoda & Yanai, 2019; Zhang et al., 2019; Yao & Gong, 2020; Chang et al., 2020; Araslanov & Roth, 2020; Wang et al., 2020; Fan et al., 2020; Sun et al., 2020). Xu et al. (2015) formulates all types of weak supervision as linear constraints on a SVM. Recent works (Lin et al., 2016; Kolesnikov & Lampert, 2016; Pathak et al., 2015) typically use Class Activation Map (CAM) (Zhou et al., 2016) to obtain an initial dense mask and then train a model iteratively. GAIN (Li et al., 2018) utilizes image tag or bounding box annotations to refine these class-specific activation maps. Sun et al. (2020) considers within-image relationships and explores the idea of co-segmentation. Fan et al. (2020) estimates the foreground and background for each category, with which the network learns to generate more precise CAMs. Regularization is enforced at either the image level (Lin et al., 2016; Kolesnikov & Lampert, 2016; Pathak et al., 2015) or the feature level (Tang et al., 2018a;b) to produce better dense masks. We incorporate this concept into adaptive feature learning and train the model only once. It is worth noting that each of the annotations carries different assumptions, we propose to unify all these types of weak annotations in a single contrastive learning framework.

Non-parametric segmentation. Prior to deep learning, non-parametric models (Russell et al., 2009; Tighe & Lazebnik, 2010; Liu et al., 2011) usually use designed features with statistical or graphical models to segment images. Recently, inspired by non-parametric models for recognition (Wu et al., 2018b;a), SegSort (Hwang et al., 2019) captures pixel-to-segment relationships via a pixel-wise embedding and develops the first deep non-parametric semantic segmentation for supervised and unsupervised settings. Our work builds upon SegSort and thus have the flexibility of a non-parametric model at capturing data relationships and modeling subclusters within a category.

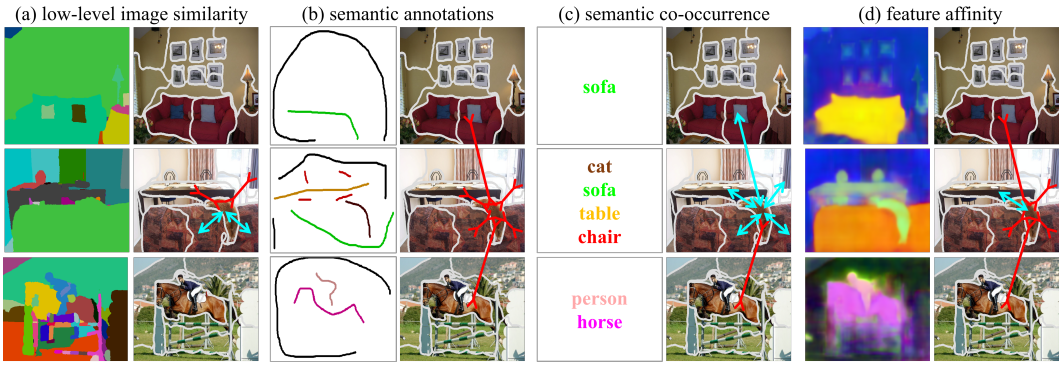


Figure 4: Four types of pixel-to-segment attraction and repulsion relationships. A pixel is attracted to (repelled by) segments: **a)** of similar (different) visual appearances such as color or texture, **b)** of the same (different) class labels, **c)** in images with common (distinctive) labels, **d)** of nearby (far-away) feature embeddings. They form different positive and negative sets.

3 Semi-Supervised Pixel-wise Metric Learning Method

Metric learning develops a feature representation based on data grouping and separation cues. Our method (Fig. 3) segments an image by learning a pixel-wise embedding with a contrastive loss between pixels and segments. We start from defining two disjoint sets – positive and negative segments (exemplars) with respect to pixel i. Our goal is to group i with positive segments while separating it from negative ones. Given latent feature $\phi(i)$ at pixel i and a specific distance metric, the objective is to decrease (increase) the distance between $\phi(i)$ and its positive (negative) segments.

Under the supervised setting, we can define the positive and negative sets as the same- and different-category pixels with respect to pixel i, which are denoted as $\mathcal{C}^+$ and $\mathcal{C}^-$. However, this idea is not applicable to weakly- or un-supervised setting, when the label at x_i is unknown. Particularly for weakly-supervised segmentation, we only have sparsely labeled pixels in the image, resulting in much smaller sets of $\mathcal{C}^+$ and $\mathcal{C}^-$ and degraded learning efficiency.

Our basic idea is to enlarge the sets of $\mathcal{C}^+$ and $\mathcal{C}^-$ to improve the learning efficacy. By exploring different relationships and assumptions in the image data, we are able to generate abundant positive and negative segments for any pixel at the same time, providing more supervision in the latent feature space. We use four types of pixel-to-segment semantic relationships (Fig. 4):

1. **Low-level image similarity**: Positive and negative segments can be generated by selecting pixel i's own segment or all the other segments, respectively.
2. **Semantic annotation**: Pseudo-labeled segments can be generated by their majority pixels' annotations. Here, the annotations can be dense masks, scribbles, points, or the localization cues generated by Class Activation Maps (Zhou et al., 2016).
3. **Semantic co-occurrence**: Positive (negative) segments can be generated by using other images that include (exclude) the same category as in pixel i's image.
4. **Feature affinity**: Pseudo-labeled segments can be generated by propagating labels from known pixels to their nearest neighbor segments in the feature space, assuming that semantic clusters emerge during training and their territories are thus expanded.

3.1 Pixel-to-Segment Semantic Relationships

We consider propagating semantic labels from labeled data $\mathcal{C}$ to unlabeled data $\mathcal{U}$ by exploiting the aforementioned relationships. We denote $\mathcal{C}$ and $\mathcal{U}$ as the sets of segment indices. Our label propagation is driven by grouping and separating data in a learned feature space. We now describe each pixel-to-segment semantic relationship for augmenting the sets of positive/negative segments using both labeled and unlabeled pixels.

Low-level image similarity. We propagate the labels within visually coherent regions, as visual similarity often goes with semantic similarity. To generate an over-segmentation, we follow Hwang et al. (2019) by using the HED contour detector (Xie & Tu, 2015) (pre-trained on BSDS500 dataset (Arbelaez et al., 2010)) and the procedure in gPb-owt-ucm (Arbelaez et al., 2010). Such bottom-up segmentation techniques consider both local and global appearance affinity without semantic information. Some of the segments could contain pixels from different categories. For learning pixel embedding at pixel i, we define positive and negative segments w.r.t. i as i's own segment and all the other segments, denoted as $\mathcal{V}^+$ and $\mathcal{V}^-$, respectively.

Such a formulation helps learn embeddings that respect low-level visual cues. In our implementation, we only consider segments in the same image as i's image as positive and negative samples. We follow SegSort Hwang et al. (2019) to align the contour-based over-segmentations with segmentations generated by K-Means clustering in SegSort.

Semantic annotation. We contrast the pixel embedding based on semantic annotations. For box and image-level tag annotations, which do not provide precise localization, we instead produce semantic cues with CAM (Zhou et al., 2016). We align these semantic annotations with oversegmentations induced by the SegSort feature. We then define segments carrying same-category and different-category labels as positive and negative samples. To reduce noise, we ignore all the unlabeled segments. We denote $\mathcal{C}^+$ and $\mathcal{C}^-$ as all the same-category and different-category regions.

Semantic co-occurrence. Semantic context characterizes the co-occurrences of different objects and can be used to group and separate pixels. We define semantic context as the union of object classes in each image. Even without the location of labels, we can leverage semantic context to impose global regularization in the latent feature space, *i.e.,* the pixel-wise feature embedding should be separated from all the other pixels from images without overlapping object categories.

We denote $\mathcal{O}^+$ as the set of segments among images with overlapping object categories. For example, if pixel i is labeled as sofa and another image also contains sofa, all the segments from that image are included in $\mathcal{O}^+$; otherwise they are considered negative segments $\mathcal{O}^-$.

This semantic context relationship does not require localization annotations or cues yet enhances the global and higher-level regularization on the pixel embedding.

Feature affinity. We consider propagating the labels from an annotated set to an unlabeled set by nearest neighbor search in the feature space. We assume that semantic clusters emerge during training with sparse supervision, reinforced by aforementioned pixel-to-segment relationships. By propagating labels in the feature space, we reinforce the learning of semantic clusters.

To be specific, we assign a semantic label for each unlabeled segment by finding its nearest labeled segment in the feature space. We denote this new labeled set (together with the original labeled set) as $\hat{\mathcal{C}}$, where $\hat{\mathcal{C}}^+$ denotes all the same-category segments other than segment s in $\hat{\mathcal{C}}$ and pixel i belongs to the segment s. Such a formulation is based on three assumptions: 1) The size of the original labeled set is large enough to cover the feature space, 2) the labeled segments are distributed uniformly in the feature space, and 3) the embedding already encodes certain semantic information. Therefore, we only apply this relationship to propagate the keypoint annotations in the DensePose dataset, where each body part is annotated by a point.

3.2 Pixel-wise Metric Learning Loss

SegSort (Hwang et al., 2019) is an end-to-end image segmentation model that generates a pixel-wise embedding and a consequent over-segmentation. It assumes that each segment has an independent normal distribution on a hypersphere. Spherical K-Means clustering (Banerjee et al., 2005) is used for segmenting an image and learning the discriminative feature clustering jointly. It is worth noting that, such an assumption indicates homogeneous representations within each segment. To learn the embedding, SegSort formulates a maximum likelihood loss that maximizes the discrimination between segments. In addition, soft neighborhood assignment (Goldberger et al., 2005) is incorporated to enforce grouping of semantically similar segments. During inference, the segment labels are predicted by K-Nearest Neighbor retrievals.

Spherical K-Means clustering (Banerjee et al., 2005) is an EM framework, which generates segmentations from a set of pixel embeddings, or $\mathcal{E} = \{\boldsymbol{e}_i\} = \{\phi(i)\}$ in an image. During M-step,

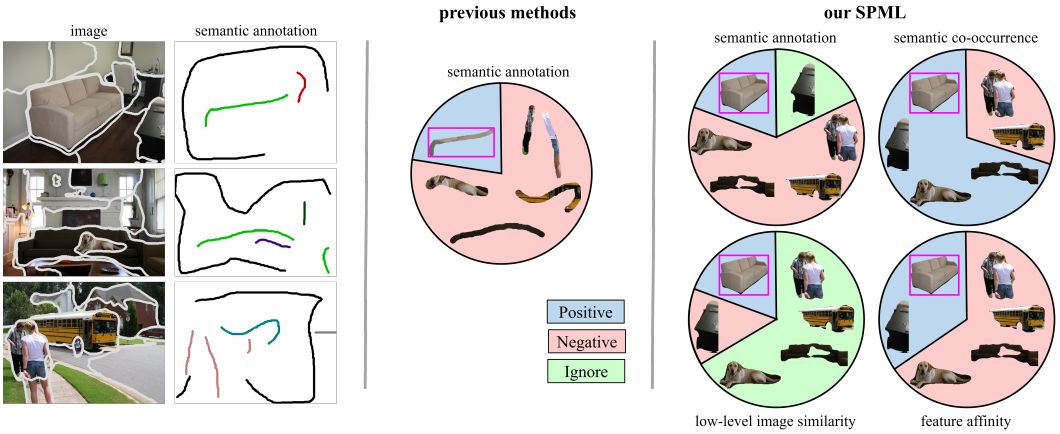


Figure 5: Given images and their labeled scribbles (left), previous methods train a pixel-wise classifier using only labeled pixels (middle) or propagate labels within each image. Instead, we (right) leverage four types of pixel-to-segment semantic relationships to augment the labeled sets.

the mean embedding, or the *prototype*, $\boldsymbol{p}_s$ is calculated for segment s, where $\boldsymbol{p}_s = \frac{\sum_{i \in \boldsymbol{R}_s} \boldsymbol{e}_i}{\sum_{i \in \boldsymbol{R}_s} \|\boldsymbol{e}_i\|}$ and $\boldsymbol{R}_s$ denotes the region of segment s. E-step assigns pixel embedding $\boldsymbol{e}_i$ to the segment of nearest prototype $\boldsymbol{p}_s$, where $z_i = \arg\max_s \boldsymbol{p}_s^\top \boldsymbol{e}_i$ is the segment index of pixel i. Therefore, the posterior probability of pixel i belonging to segment s can be derived as $p(z_i = s|\boldsymbol{e}_i, \Theta) = \frac{\exp(\kappa \boldsymbol{p}_s^\top \boldsymbol{e}_i)}{\sum_{l \in \Omega} \exp(\kappa \boldsymbol{p}_l^\top \boldsymbol{e}_i)}$, where κ denotes the concentration hyper-parameter and Ω represents the set of all segments.

The SegSort objective is to minimize the negative log-likelihood loss:

$$L^i_{\text{SegSort}} = -\log p(z_i = s|\boldsymbol{e}_i, \Theta) = -\log \frac{\exp(\kappa \boldsymbol{p}_s^\top \boldsymbol{e}_i)}{\sum_{l \in \Omega} \exp(\kappa \boldsymbol{p}_l^\top \boldsymbol{e}_i)}. \tag{1}$$

To strengthen the grouping of same-category segments, Hwang et al. (2019) further adopt the soft neighborhood assignments (Goldberger et al., 2005):

$$p(z_i = c^+|\boldsymbol{v}_i, \Theta) = \frac{\exp(\kappa \boldsymbol{p}_{c^+}^\top \boldsymbol{e}_i)}{\sum_{l \neq s, l \in \Omega} \exp(\kappa \boldsymbol{p}_l^\top \boldsymbol{e}_i)}; \qquad p(z_i = s|\boldsymbol{e}_i, \Theta) = 0. \tag{2}$$

where c^+ is the segment index of all the other same-category segments as s. We also denote c^- as the index of all the different-category segments from s. By letting $\mathcal{C}^+ = \{c^+\}$ and $\mathcal{C}^- = \{c^-\}$ w.r.t pixel i, the corresponding loss is formulated:

$$L_{\text{SegSort}^+}(i, \mathcal{C}^+, \mathcal{C}^-) = -\log \sum_{l \in \mathcal{C}^+} p(z_i = l|\boldsymbol{e}_i, \Theta) = -\log \frac{\sum_{l \in \mathcal{C}^+} \exp(\kappa \boldsymbol{p}_l^\top \boldsymbol{e}_i)}{\sum_{l \in \mathcal{C}^+ \cup \mathcal{C}^- \setminus \{s\}} \exp(\kappa \boldsymbol{p}_l^\top \boldsymbol{e}_i)}. \tag{3}$$

For our weakly supervised segmentation, the total pixel-to-segment contrastive loss for pixel i consists of 4 terms, one for each of the 4 pixel-to-segment attraction and repulsion relationships:

$$\begin{aligned} L(i) = {} & \lambda_I L_{\text{SegSort}^+}(i, \mathcal{V}^+, \mathcal{V}^-) + \lambda_C L_{\text{SegSort}^+}(i, \mathcal{C}^+, \mathcal{C}^-) \\ & + \lambda_O L_{\text{SegSort}^+}(i, \mathcal{O}^+, \mathcal{O}^-) + \lambda_A L_{\text{SegSort}^+}(i, \hat{\mathcal{C}}^+, \hat{\mathcal{C}}^-). \end{aligned} \tag{4}$$

Fig. 5 shows how our metric learning differs from previous classification methods at utilizing labeled and unlabeled pixels and segments. By easily integrating a full range of pixel-to-segment attraction and repulsion relationships from low-level image similarity to mid-level feature affinity, and to high-level semantic co-occurrence, we go far beyond the direct supervision from semantic annotations.

4 Experiments

Dataset: Pascal VOC 2012. Pascal VOC 2012 dataset (Everingham et al., 2010) includes 20 object categories and one background class. Following Chen et al. (2017), we use the augmented train-

Dataset	Annotation	λ_I	λ_C	λ_O	λ_A
Pascal VOC 2012	scribbles	0.1	1.0	0.5	0.0
	boxes	0.3	1.0	1.0	0.0
	points	1.0	1.0	1.0	0.0
	image tags	0.3	1.0	1.0	0.0
DensePose	points	0.1	1.0	0.0	0.5

Table 1: Hyper-parameters for different types of annotations on Pascal and DensePose datasets.

ing set with 10,582 images and validation set with 1,449 images. We use the scribble annotations provided by Lin et al. (2016) for training.

Dataset: DensePose. DensePose (Alp Güler et al., 2018) is a human pose parsing dataset based on MSCOCO (Lin et al., 2014). The dataset is annotated with 14 body part classes. We extract the keypoints from the center of each part segmentation. The training set includes 26,437 images and we use minival2014 set for testing, which includes 1,508 images. See Appendix for more details.

Architecture, training and testing. For all the experiments on VOC, we base our architecture as DeepLab (Chen et al., 2017) with ResNet101 (He et al., 2016) as backbone network. For the experiments on DensePose dataset, we adopt PSPNet (Zhao et al., 2017) as backbone network. We only use models pre-trained on ImageNet (Deng et al., 2009) dataset. For training our models, we set $\lambda_I, \lambda_C, \lambda_O$ and λ_A according to different types of annotations and datasets, which is shown in table 1. For inference, we follow SegSort (Hwang et al., 2019) to perform k-nearest neighbor retrievals. See Appendix for more detail of hyper-parameters and setting for training and testing.

Table 1 summarizes the four relationships we experiment with for different annotation types and datasets. They can be formulated jointly in a single pixel-wise metric learning framework (Fig. 3).

Pascal: Image tag annotations. On Pascal VOC dataset, our method outperforms others by a large margin. Table 2 shows that, without additional saliency labels, our method still achieves SOTA. Compared to (Chang et al., 2020), we improves mIoU by a sizable 4.5%.

Pascal: Box annotations. Table 2 shows that, with the same DeepLab/ResNet101 backbone network, our method outperforms others: On the validation set, our performance is 2.9% higher.

Pascal: Image tags	Saliency	*val*	*test*
Huang et al. (2018)	✓	61.4	63.2
Lee et al. (2019)	✓	64.9	65.3
Zhang et al. (2019)	-	66.3	66.5
Yao & Gong (2020)	✓	67.1	67.2
Chang et al. (2020)	-	66.1	-
Our SPML	-	**70.6**	**71.5**

Pascal: Bounding boxes	*val*	*test*
Khoreva et al. (2017)	69.4	-
Song et al. (2019)	70.2	-
Our SPML	**73.1**	**74.6**

Table 2: Pascal VOC 2012 dataset with image tag (left) and bounding box (right) annotations.

Pascal: Scribble annotations. Table 3 shows that, without CRF post-processing, we get 74.1% mIoU, achieving 97.6% of full supervision performance; with CRF post-processing, we reach new SOTA: We get 75.9% mIoU, achieving 98.6% of full supervision performance.

Pascal: Varying sparsity of scribble and point annotations. Exploiting metric learning with different relationships in the data frees us from the classification framework, which requires a huge number of annotations. We investigate whether we can retain the segmentation performance with even sparser annotations. Table 3 shows that, as we shorten the length of scribbles from $100\%, 80\%, 50\%, 30\%$ to 0% (points), we reach $98.6\%, 98.1\%, 96.5\%, 95.1\%$ and 92.5% of full supervision performance. Compared to the scribble annotations, our performance only drops by 2.6% with using point labels, which is significantly better than the baseline.

Pascal: Learning from different weak annotations.

DensePose: Point annotations. We train our baseline using the code released by Tang et al. (2018b). Table 4 shows that, without CRF post-processing, we are able to

DensePose: Points	mIoU	WvF
Tang et al. (2018b)	31.3	51.9
Our SPML	**46.9**	**81.8**

Table 4: DensePose minival 2014 set.

Pascal: Scribbles	CRF	Full	Weak	WvF
Tang et al. (2018a)		75.6	72.8	96.3
Tang et al. (2018a)	✓	76.8	74.5	97.0
Tang et al. (2018b)		75.6	73.0	96.6
Tang et al. (2018b)	✓	76.8	75.0	97.7
Wang et al. (2019)		75.6	73.2	96.8
Wang et al. (2019)	✓	76.8	76.0	99.0
Our SPML		75.9	**74.1**	**97.6**
Our SPML	✓	77.0	75.9	98.6
Our SPML	✓	77.3	76.4	98.8

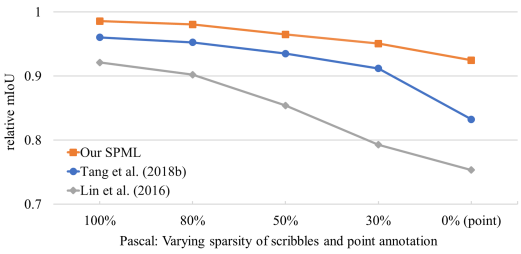
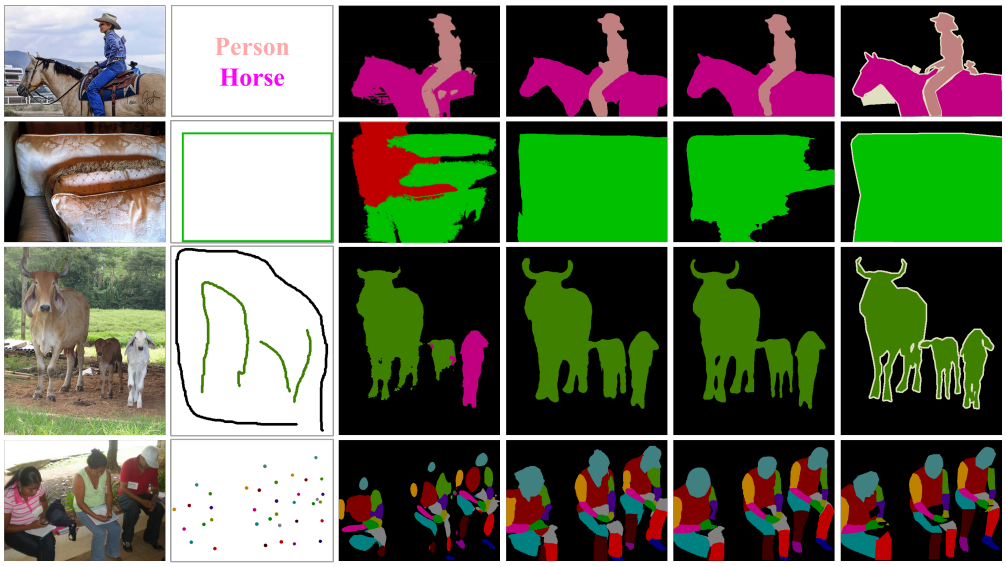


Table 3: Pascal VOC 2012 dataset using scribble annotations. **Left**: mIoU on validataion (white) and test (gray) set. **WvF** denotes relative mIoU w.r.t full supervision. **Right**: Relative mIoU performance w.r.t full supervision on different lengths of scribbles.



(a) Image (b) Annotation (c) Baseline (d) Our SPML (e) Full Supervision (f) Ground Truth

Figure 6: Visual quality comparison on Pascal and DensePose dataset with different types of annotations. Our results are well aligned with image boundary, and visually comparable to full supervision.

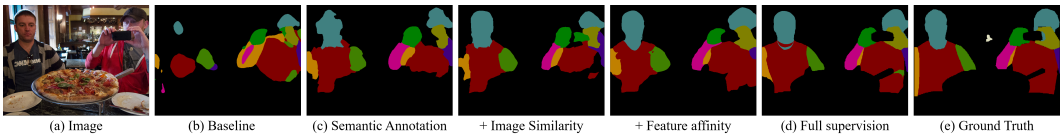
reach 81.8% of full supervision performance. Note that our method does not require a two-stage training procedure as in the baseline.

Visual quality comparison and Ablation study. Fig. 6 shows that our results are well aligned with image boundary and visually close to fully-supervised counterpart. Fig. 7 shows significant improvement of our results by additively introducing different relationships for more regularization.

Please refer to the Appendix for more details and ablation studies.

5 Summary

We propose a novel universal weakly-supervised semantic segmentation method via Semi-supervised Pixel-wise Metric Learning. Four common types of pixel-to-segment attraction and repulsion relationships can be derived, whether the partial annotation is coarse image tags and bounding boxes, or sparse keypoints and scribbles. Our results on PASCAL VOC and DensePose show consistent and substantial gains over SOTA, especially for the sparsest keypoint supervision.



(a) Image (b) Baseline (c) Semantic Annotation + Image Similarity + Feature affinity (d) Full supervision (e) Ground Truth

Figure 7: Visual comparison of combinations of different regularizations. Our results (c) are improved significantly by introducing more relationships as regularizations.

REFERENCES

Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4981–4990, 2018.

Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7297–7306, 2018.

Nikita Araslanov and Stefan Roth. Single-stage semantic segmentation from image labels. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2020.

Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):898–916, 2010.

Arindam Banerjee, Inderjit S Dhillon, Joydeep Ghosh, and Suvrit Sra. Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 2005.

Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What's the point: Semantic segmentation with point supervision. In *European conference on computer vision*, pp. 549–565. Springer, 2016.

Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on pattern analysis and machine intelligence*, 23(11):1222–1239, 2001.

Yu-Ting Chang, Qiaosong Wang, Wei-Chih Hung, Robinson Piramuthu, Yi-Hsuan Tsai, and Ming-Hsuan Yang. Weakly-supervised semantic segmentation via sub-category exploration. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2020.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4): 834–848, 2017.

Kevin Clark, Thang Luong, and Quoc V Le. Cross-view training for semi-supervised learning. *arXiv preprint arXiv:1809.08370*, 2018.

Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1635–1643, 2015.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010.

Junsong Fan, Zhaoxiang Zhang, Chunfeng Song, and Tieniu Tan. Learning integral objects with intra-class discriminator for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Rob Fergus, Yair Weiss, and Antonio Torralba. Semi-supervised learning in gigantic image collections. In *Advances in neural information processing systems*, pp. 522–530, 2009.

Jacob Goldberger, Geoffrey E Hinton, Sam T Roweis, and Ruslan R Salakhutdinov. Neighbourhood components analysis. In *NIPS*, 2005.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7014–7023, 2018.

Jyh-Jing Hwang, Stella X Yu, Jianbo Shi, Maxwell D Collins, Tien-Ju Yang, Xiao Zhang, and Liang-Chieh Chen. Segsort: Segmentation by discriminative sorting of segments. In *ICCV*, 2019.

Thorsten Joachims. Transductive learning via spectral graph partitioning. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pp. 290–297, 2003.

Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 876–885, 2017.

Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Advances in neural information processing systems*, pp. 3581–3589, 2014.

Alexander Kolesnikov and Christoph H Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *European Conference on Computer Vision*, pp. 695–711. Springer, 2016.

Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in neural information processing systems*, pp. 109–117, 2011.

Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5267–5276, 2019.

Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Tell me where to look: Guided attention inference network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9215–9223, 2018.

Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3159–3167, 2016.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Bin Liu, Zhirong Wu, Han Hu, and Stephen Lin. Deep metric transfer for label propagation with limited annotated data. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 0–0, 2019.

Ce Liu, Jenny Yuen, and Antonio Torralba. Nonparametric scene parsing via label transfer. *PAMI*, 2011.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018.

George Papandreou, Liang-Chieh Chen, Kevin P Murphy, and Alan L Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *Proceedings of the IEEE international conference on computer vision*, pp. 1742–1750, 2015.

Deepak Pathak, Philipp Krahenbuhl, and Trevor Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *Proceedings of the IEEE international conference on computer vision*, pp. 1796–1804, 2015.

Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. Semi-supervised learning with ladder networks. In *Advances in neural information processing systems*, pp. 3546–3554, 2015.

Bryan Russell, Alyosha Efros, Josef Sivic, Bill Freeman, and Andrew Zisserman. Segmenting scenes by matching image composites. In *NIPS*, 2009.

Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000.

Wataru Shimoda and Keiji Yanai. Self-supervised difference detection for weakly-supervised semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5208–5217, 2019.

Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3136–3145, 2019.

Guolei Sun, Wenguan Wang, Jifeng Dai, and Luc Van Gool. Mining cross-image semantics for weakly supervised semantic segmentation. In *ECCV*, 2020.

Meng Tang, Abdelaziz Djelouah, Federico Perazzi, Yuri Boykov, and Christopher Schroers. Normalized cut loss for weakly-supervised cnn segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1818–1827, 2018a.

Meng Tang, Federico Perazzi, Abdelaziz Djelouah, Ismail Ben Ayed, Christopher Schroers, and Yuri Boykov. On regularized losses for weakly-supervised cnn segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 507–522, 2018b.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in neural information processing systems*, pp. 1195–1204, 2017.

Joseph Tighe and Svetlana Lazebnik. Superparsing: scalable nonparametric image parsing with superpixels. In *ECCV*, 2010.

Bin Wang, Guojun Qi, Sheng Tang, Tianzhu Zhang, Yunchao Wei, Linghui Li, and Yongdong Zhang. Boundary perception guidance: A scribble-supervised semantic segmentation approach. *International Joint Conference on Artificial Intelligence*, 2019.

Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. *Proceedings of the IEEE International Conference on Computer Vision*, 2020.

Jason Weston, Frédéric Ratle, Hossein Mobahi, and Ronan Collobert. Deep learning via semi-supervised embedding. In *Neural networks: Tricks of the trade*, pp. 639–655. Springer, 2012.

Zhirong Wu, Alexei A Efros, and Stella X Yu. Improving generalization via scalable neighborhood component analysis. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 685–701, 2018a.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3733–3742, 2018b.

Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 1395–1403, 2015.

Jia Xu, Alexander G Schwing, and Raquel Urtasun. Learning to segment under various forms of weak supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3781–3790, 2015.

Qi Yao and Xiaojin Gong. Saliency guided self-attention network for weakly and semi-supervised semantic segmentation. *IEEE Access*, 8:14413–14423, 2020.

Bingfeng Zhang, Jimin Xiao, Yunchao Wei, Mingjie Sun, and Kaizhu Huang. Reliability does matter: An end-to-end weakly supervised semantic segmentation approach. *arXiv preprint arXiv:1911.08039*, 2019.

Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2881–2890, 2017.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, 2016.

Dengyong Zhou, Olivier Bousquet, Thomas N Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In *Advances in neural information processing systems*, pp. 321–328, 2004.

A APPENDIX

We propose a single pixel-to-segment contrastive learning loss formulation for weakly supervised semantic segmentation. We explore different types of visual relationships to group and separate pixels within and *across* images. We demonstrate state-of-the-art performance using our proposed method with different types of annotations. Here, we include more details on the following aspects:

- We present the visual results of our method in A.1.
- We showcase the semantic cues generated by CAM for image tag and bounding box annotations in A.2.
- We illustrate the data pre-processing used for DensePose dataset in A.3.
- We describe the details of our experimental settings and hyper-parameters in A.4.
- We present the ablation study regarding hyper-parameters in A.5.
- We present mIoU performance with varying sparsity of scribble annotations on Pascal dataset in A.6.
- We present per-category results with scribble and point annotations on Pascal and DensePose dataset in A.7.
- We provide detailed experimental results using image tag, bounding box and scribble annotations on Pascal dataset in table 13, 12 and 11, respectively.

A.1 VISUALIZATION

We present the visual results on VOC (with image tags, bounding boxes and scribbles) and DensePose (with keypoints) dataset in figure 8. We observe that our segmentation results are especially well aligned with image boundary. When visual evidence is prominent, our weakly-supervised results are even better than the fully-supervised counterpart.

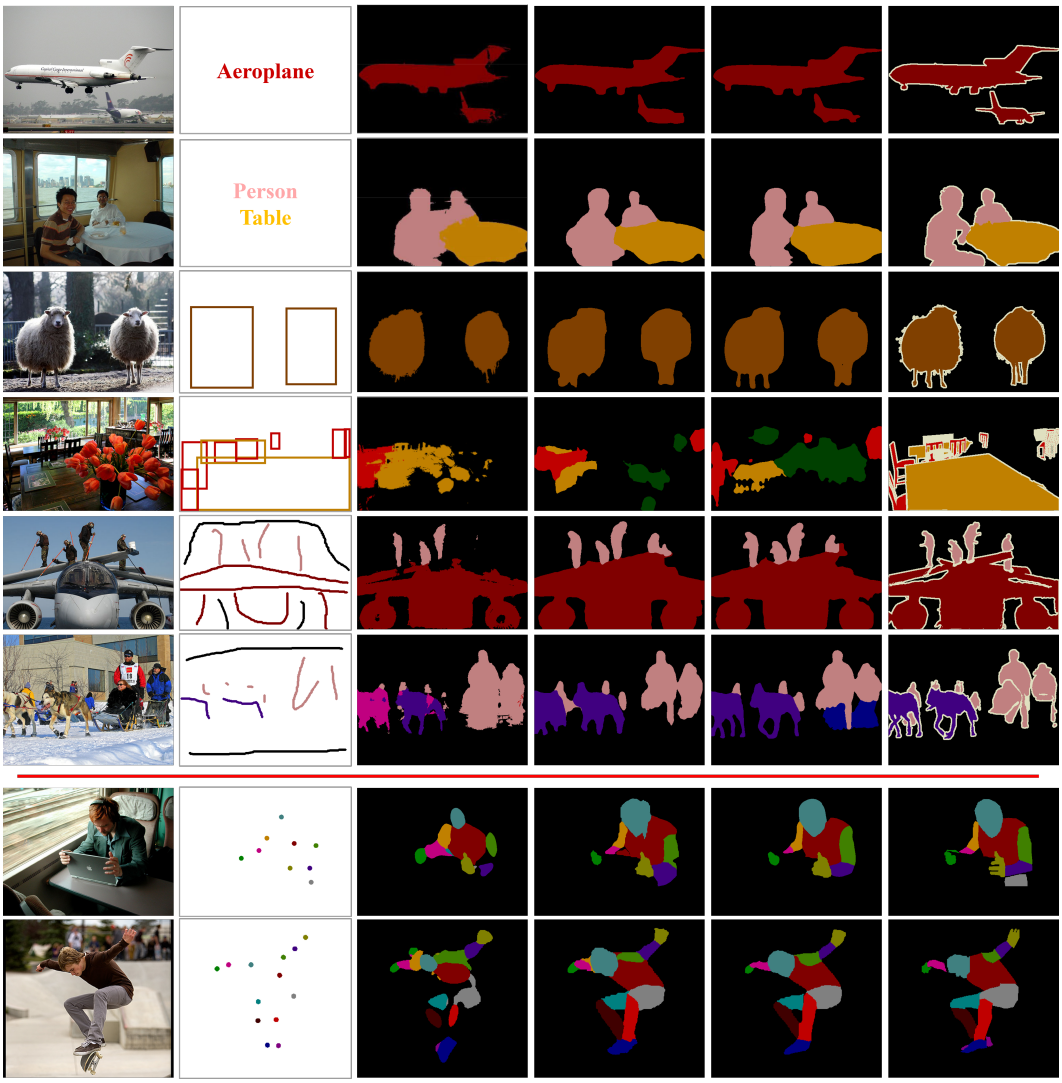
We then demonstrate the efficacy of each visual relationship in figure 9. By adding **semantic annotation**, **low-level image similarity** and **feature affinity** progressively, we observe consistent improvement of our results. The predicted segmentation becomes more coherent and better aligned with image boundary. We lastly showcase that our method implicitly encodes semantic contexts. In figure 10, We observe that retrieved segments appear in the similar semantic context as the query segments. For examples, given a bottle next to a desktop, our model retrieves bottles also next to a desktop; a set of sofas in a living room can be retrieved using one sofa query example; screens of a desktop can also be retrieved likewise.

A.2 SEMANTIC ANNOTATIONS

Since image tag and bounding box annotation do not provide any of precisely localized semantic information, we adopt CAM (Zhou et al., 2016) to produce localized semantic cues. Without using additional saliency labels, we use the classifier trained by Ahn & Kwak (2018) to generate CAM. Let $M_c(I)$ be the activation map of class c in image I. Proposed in Ahn & Kwak (2018), $M_c(I)$ is normalized to the range between 0 and 1, where $M_c(I) = \frac{M_c(I)}{\max M_c(I)}$. The background confidence $M_{bg}(I)$ can then be estimated by $M_{bg}(I) = \{1 - \max_{c \in C} M_c(I)\}^{\alpha}$, where α is the hyper-parameter adjusting background confidence and C is the set of all object categories. In our experiments, we set α to 32 and confidence threshold to 0.3 and 0.2 for image tag and bounding box annotations. For bounding box annotations, we further restrict all the regions outside bounding boxes as "background". See figure 11 for more visual examples.

A.3 DATA PRE-PROCESSING FOR DENSEPOSE DATASET

We next illustrate our pre-processing to generate training labels given keypoint annotations in DensePose dataset. As shown in figure 12, we first assume a Gaussian heat map from every keypoint. By thresholding, we derive 3 regions from every Gaussian blob: labelled, unknown and background region. In labelled region, pixels are annotated as each body part. We then propagate labels, including



(a) Image (b) Annotation (c) Baseline (d) Our SPML (e) Full Supervision (f) Ground Truth

Figure 8: Visual comparison of baseline method (c), our SPML (d) and fully-supervised SegSort (e) on Pascal VOC 2012 and DensePose dataset. On Pascal dataset (top 6 rows), our baseline method is based on Lee et al. (2019); Song et al. (2019); Tang et al. (2018b) for image tag, bounding box and scribble annotations, respectively. On DensePose dataset (bottom 2 rows), our baseline is Tang et al. (2018b). The results from our weakly-supervised model is visually very close to its fully-supervised counterpart, or even better when visual cues are prominent.

background class, to pixels in the unknown region. The std of Gaussian heat map is estimated from instance size, and we use ground-truth information in our paper.

A.4 Hyper-parameters and Experimental Setup

Architecture and training. For all the experiments on VOC, we base our architecture as DeepLab (Chen et al., 2017) with ResNet101 (He et al., 2016) as backbone network. For the experiments on DensePose dataset, we adopt PSPNet (Zhao et al., 2017) as backbone network. We only use models pre-trained on ImageNet (Deng et al., 2009) dataset.

We next describe the hyper-parameters used for each experiment. For all the experiments, we set "batchsize" to 8, "cropsize" to 336×336 and 480×480 for VOC and DensePose dataset. Following (Chen et al., 2017), we adopt poly learning rate policy by multiplying base learning rate by $1 - (\frac{iter}{max_iter})^{0.9}$. We set initial learning rate to 0.003, momentum to 0.9. For the hyper-parameters

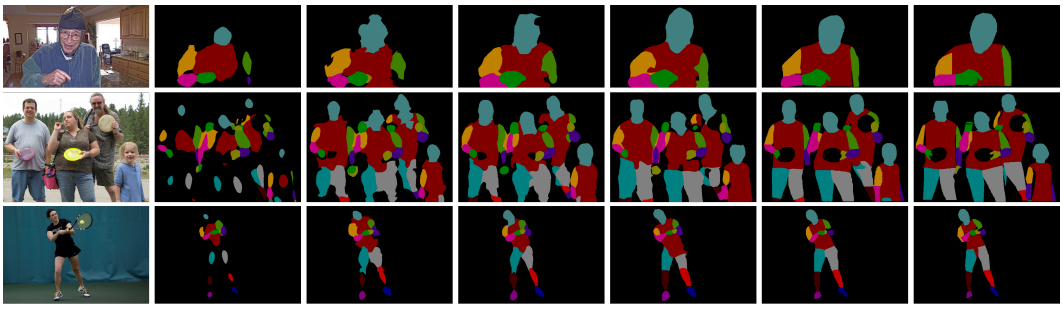
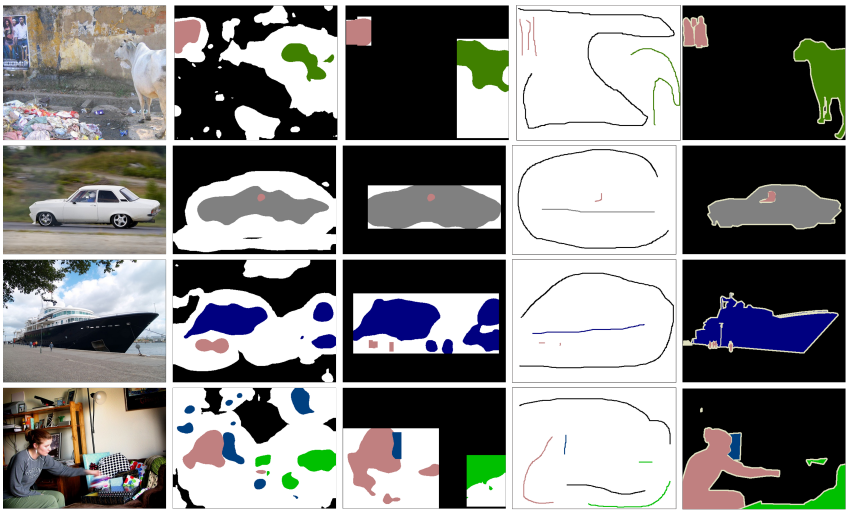


Figure 9: Visual quality comparison of different regularizations. Our method improves significantly after more relationships are introduced as regularization.

in SegSort framework, we use unit-length normalized embeddings of dimension 64. We iterate K-Means clustering for 10 iterations and generate 25 and 144 clusters on VOC and DensePose dataset. The concentration parameter κ is set to $6, 16, 12, 12$ for **semantic annotation**, **low-level image similarity**, **semantic co-occurrence** and **feature affinity**, respectively. Shown in table 1, we also set $\lambda_I, \lambda_C, \lambda_O$ and λ_A according to different types of annotations and datasets. We train for $30k$ and $45k$ iterations on VOC and DensePose dataset for all the experiments. We use additional memory banks to cache up previous 2 batches.

Inference and testing. Following Hwang et al. (2019), we conduct k-nearest neighbor retrievals for inference. We produce oversegmentations and calculate the prototypes in the training set. For Pascal VOC and DensePose dataset, We partition each image into 144 and 576 clusters, respectively. We set $k = 20$ for nearest neighbor retrievals, and then, we infer the semantic label of each queried segment with majority label of retrievals. Notably, we do not use left-right flipped and multi-scale inputs as data augmentation for inference. For fair comparison, we also use CRF post-processing (if specified) to improve our performance on Pascal VOC dataset under different settings of supervision.

Figure 10: Visual examples of nearest neighbor segment retrievals. We observe that retrieved segments (right) appear in the similar semantic context as the query segments (left). For examples, given a bottle next to a desktop, our model retrieves bottles also next to a desktop.



(a) Image (b) CAM (label) (c) CAM (box) (d) Scribble (e) Ground Truth

Figure 11: Visual examples of semantic annotations used on Pascal VOC 2012 dataset. For image tag and bounding box annotation, we use the classifier trained by Ahn & Kwak (2018) to infer CAM as semantic annotation. These semantic annotations are noisy, which do not precisely localize on the objects.

λ_I	λ_O	mIoU
0.3	0.5	73.7
0.1	0.5	74.1
0.05	0.5	73.5
0	0.5	71.7

λ_I	λ_O	mIoU
0.1	1.0	74.1
0.1	0.5	74.1
0.1	0.1	74.1
0.1	0	72.8

λ_C	λ_I	λ_O	λ_A	mIoU
1	0	0	0	71.2
1	0.1	0	0	72.8
1	0.1	0.5	0	74.1
1	0.1	0.5	0.1	73.8

Table 5: Ablation study of different weighting parameters for each objective function on Pascal VOC validation dataset.

A.5 Ablation study of hyper-parameters

We conduct ablation study over different regularizations on Pascal VOC and DensePose dataset. As shown in table 5, we achieve the most optimal performance on Pascal VOC dataset with $\lambda_I = 0.1$ and $\lambda_O = 0.5$. We also observe performance drops 0.3 of mIoU by adding **feature affinity** regularization. We argue that scribble/box/point annotations are not uniformly distributed across object instance and background, and results in noisy label propagation. In table 6, we also demonstration

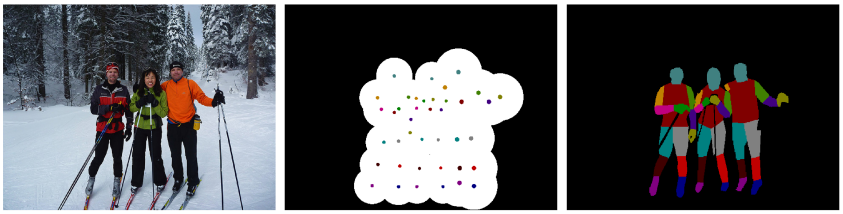


Figure 12: Preparing training labels on DensePose dataset. From left to right are input image, our training labels and ground-truth mask. For each keypoint, a Gaussian heat map is applied to determine labelled, unknown and background region. The white region denotes unknown pixels, to which we propagate labels from annotated or background region.

λ_C	λ_I	λ_A	λ_O	mIoU
1	0	0	0	42.9
1	0.1	0	0	44.1
1	0.1	0.5	0	46.9
1	0.1	0.5	0.5	46.5

Table 6: Ablation study of different weighting parameters for each objective function on DensePose minival 2014 set.

the efficacy of each loss. We show that regularization on semantic annotation, low-level image similarity and feature affinity improve the performance significantly, whereas, our model doesn't benefit from semantic co-occurrence regularization.

Method	Backbone	CRF	Full	100%	80%	50%	30%	0%
Lin et al. (2016)	DeepLab-MSc-LargeFOV	✓	68.5	63.1	61.8	58.5	54.3	51.6
Tang et al. (2018b)	DeepLab-MSc-LargeFOV	✓	68.7	66.0	65.5	64.2	62.7	57.2
Our SPML	DeepLab/ResNet101		75.9	74.1	73.7	72.6	71.5	69.6
Our SPML	DeepLab/ResNet101	✓	77.0	75.9	75.5	74.3	73.2	71.2
Our SPML	PSPNet/ResNet101		78.3	76.7	75.9	75.1	73.6	71.3
Our SPML	PSPNet/ResNet101	✓	79.0	78.3	77.2	76.6	75.3	73.0

Table 7: mIoU performance on Pascal VOC 2012 validation set on different lengths of scribble.

Backbone	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mIoU
Tang et al. (2018b)	83.2	35.8	82.8	66.8	75.1	90.9	83.9	89.2	35.8	82.5	53.7	83.4	83.2	79.5	82.2	57.6	81.9	41.6	81.1	73.5	73.2
Our SPML	**85.6**	**39.1**	**86.1**	**69.1**	**80.2**	**92.5**	**84.1**	**90.3**	35.2	75.5	**58.1**	**85.1**	74.0	**82.1**	**84.7**	54.9	**82.2**	**49.8**	**86.2**	67.9	**74.1**
Tang et al. (2018b)	86.2	37.3	85.5	69.4	77.8	91.7	85.1	91.2	38.8	85.1	55.5	85.6	85.8	81.7	84.1	61.4	84.3	43.1	81.4	74.2	75.2
Our SPML	**88.3**	**40.7**	**88.9**	**71.8**	**81.6**	**93.2**	**85.4**	**91.4**	37.0	77.3	**59.6**	**86.9**	76.2	**83.9**	**86.4**	56.1	**85.5**	**52.1**	**87.8**	69.3	**75.9**

Table 8: Per-class results on Pascal VOC 2012 validation set. White- and gray-colored background denotes using without- and with- CRF post-processing for inference.

Backbone	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mIoU
DeepLab/ResNet101	91.5	43.5	83.0	67.9	81.7	89.8	88.7	94.6	37.5	81.6	68.7	88.8	82.4	88.6	87.6	64.1	87.6	52.7	76.5	71.4	77.3
PSPNet/ResNet101	92.5	43.4	83.9	68.9	79.8	94.6	90.2	95.5	39.8	86.8	69.4	90.2	85.4	89.7	87.0	73.9	91.3	59.4	84.8	79.3	80.0
DeepLab/ResNet101	84.2	41.7	92.5	65.8	73.9	88.0	85.8	91.7	38.1	81.1	68.4	88.6	86.0	84.3	86.7	67.8	84.4	60.3	72.1	68.4	76.4
PSPNet/ResNet101	91.2	39.9	93.1	65.6	71.5	92.1	84.9	92.0	46.9	86.6	68.6	88.0	86.4	82.5	86.5	70.6	90.4	55.2	82.0	73.0	77.7

Table 9: Per-class results on Pascal VOC 2012 testing set. White- and gray-colored background denotes using full and scribble annotations. CRF post-processing is used for inference.

A.6 MEAN IOU PERFORMANCE WITH VARYING SPARSITY OF SCRIBBLES.

We also report absolute mIoU performance by varying sparsity of scribbles on Pascal VOC 2012 validation set. The results are summarized in table 7. Our results are higher than baseline by larger margin when the annotation is sparser.

A.7 PER-CATEGORY MIOU ON PASCAL VOC AND DENSEPOSE DATASET.

We next present per-category results on Pascal VOC and Denspose dataset. In table 8, we compare with Tang et al. (2018b) on VOC validation set. Without- and with CRF post-processing, our method outperform the baseline method among most categories by large margin. We further conduct

Method	bg.	torso	RHand	LHand	LFoot	RFoot	RThigh	LThigh	RLeg	LLeg	LArm	RArm	LFarm	RFarm	Heaad	mIoU	WvF
Softmax	96.2	73.7	61.1	57.2	37.2	37.8	56.8	54.8	49.7	49.5	62.0	63.8	58.3	61.5	84.6	60.3	-
SegSort	95.8	71.9	57.4	53.0	33.4	33.4	54.0	51.8	46.4	46.9	59.2	61.1	54.4	57.9	83.2	57.3	-
Tang et al. (2018b)	87.2	28.3	37.5	36.0	18.9	19.5	21.2	20.8	16.1	16.6	33.9	35.3	35.6	37.6	25.2	31.3	51.9
Our SPML	93.9	60.6	49.0	43.4	26.3	25.0	41.2	40.2	33.6	34.7	42.9	46.6	44.3	46.9	74.9	46.9	81.8

Table 10: Per-class results on DensePose minival 2014 set with keypoint annotations. White- and gray-colored background indicates using full and point supervision.

Method	Backbone	CRF	Full	Weak	WvF
Tang et al. (2018a)	DeepLab/ResNet101		75.6	72.8	96.3
Tang et al. (2018a)	DeepLab/ResNet101	✓	76.8	74.5	97.0
Tang et al. (2018b)	DeepLab/ResNet101		75.6	73.0	96.6
Tang et al. (2018b)	DeepLab/ResNet101	✓	76.8	75.0	97.7
Wang et al. (2019)	DeepLab/ResNet101		75.6	73.2	96.8
Wang et al. (2019)	DeepLab/ResNet101	✓	76.8	76.0	99.0
Our SPML	DeepLab/ResNet101		75.9	**74.1**	**97.6**
Our SPML	DeepLab/ResNet101	✓	77.0	**75.9**	**98.6**
Our SPML	PSPNet/ResNet101		78.3	**76.7**	**98.1**
Our SPML	PSPNet/ResNet101	✓	79.0	**78.3**	**99.1**
Our SPML	DeepLab/ResNet101	✓	77.3	**76.4**	**98.8**
Our SPML	PSPNet/ResNet101	✓	80.0	**77.7**	**97.1**

Table 11: Pascal VOC 2012 validataion and test set with scribble annotations. **WvF** denotes relative mIoU w.r.t full supervision. White- and gray-colored background denotes validation and test set.

Method	Backbone	CRF	*val*	*test*
Dai et al. (2015)	DeepLab-LargeFOV	✓	62.0	64.6
Khoreva et al. (2017)	DeepLab/ResNet101	✓	69.4	-
Song et al. (2019)	DeepLab/ResNet101	✓	70.2	-
Our SPML	DeepLab/ResNet101	-	**71.2**	**72.8**
Our SPML	DeepLab/ResNet101	✓	**73.1**	**74.6**
Our SPML	PSPNet/ResNet101	-	**73.6**	**73.5**
Our SPML	PSPNet/ResNet101	✓	**74.9**	**74.9**

Table 12: Pascal VOC 2012 dataset with bounding box annotation.

experiments on VOC testing set, using DeepLab and PSPNet as backbone network. In table 9, we can retrieve most performance w.r.t full supervision. We also compare per-category results on DensePose dataset in table 10. We train our baseline method using the code released by Tang et al. (2018b). We outperform the baseline method by large margin in every category.

Method	Backbone	Saliency	CRF	*val*	*test*
Huang et al. (2018)	DeepLab/ResNet101	✓	✓	61.4	63.2
Ahn & Kwak (2018)	WideResNet38	-	✓	61.7	63.7
Araslanov & Roth (2020)	WideResNet38	-	✓	62.7	64.3
Wang et al. (2020)	WideResNet38	-	✓	64.5	65.7
Lee et al. (2019)	DeepLab/ResNet101	✓	✓	64.9	65.3
Shimoda & Yanai (2019)	WideResNet38	-	✓	64.9	65.5
Zhang et al. (2019)	DeepLab/ResNet101	-	✓	66.3	66.5
Yao & Gong (2020)	DeepLab/ResNet101	✓	✓	67.1	67.2
Chang et al. (2020)	DeepLab/ResNet101	-	-	64.8	-
Chang et al. (2020)	DeepLab/ResNet101	-	✓	66.1	-
Our SPML	DeepLab/ResNet101	-	-	**68.6**	**69.5**
Our SPML	DeepLab/ResNet101	-	✓	**70.6**	**71.5**
Our SPML	PSPNet/ResNet101	-	-	**70.5**	**70.0**
Our SPML	PSPNet/ResNet101	-	✓	**72.5**	**72.0**

Table 13: Pascal VOC 2012 dataset with image tag annotation.