CIAN: Cross-Image Affinity Net for Weakly Supervised Semantic Segmentation

Junsong Fan,[1,2,4] **Zhaoxiang Zhang,**[1,2,3,4*] **Tieniu Tan,**[1,2,3,4]
Chunfeng Song,[1,2,4] **Jun Xiao**[4*]
[1]Center for Research on Intelligent Perception and Computing, CASIA
[2]National Laboratory of Pattern Recognition, CASIA
[3]Center for Excellence in Brain Science and Intelligence Technology, CAS
[4]University of Chinese Academy of Sciences
{fanjunsong2016, zhaoxiang.zhang}@ia.ac.cn, {tnt, chunfeng.song}@nlpr.ia.ac.cn, xiaojun@ucas.ac.cn

Abstract

Weakly supervised semantic segmentation with only image-level labels saves large human effort to annotate pixel-level labels. Cutting-edge approaches rely on various innovative constraints and heuristic rules to generate the masks for every single image. Although great progress has been achieved by these methods, they treat each image independently and do not take account of the relationships across different images. In this paper, however, we argue that the cross-image relationship is vital for weakly supervised segmentation. Because it connects related regions across images, where supplementary representations can be propagated to obtain more consistent and integral regions. To leverage this information, we propose an end-to-end cross-image affinity module, which exploits pixel-level cross-image relationships with only image-level labels. By means of this, our approach achieves 64.3% and 65.3% mIoU on Pascal VOC 2012 validation and test set respectively, which is a new state-of-the-art result by only using image-level labels for weakly supervised semantic segmentation, demonstrating the superiority of our approach.

Introduction

Semantic segmentation provides per pixel predictions for a given image. Recently, fully convolutional network (FCN) based methods (Long, Shelhamer, and Darrell 2015; Chen et al. 2018; 2017) have achieved impressive performance. However, these deep methods need large scale datasets with precise pixel-level annotations for training (Everingham et al. 2010; Lin et al. 2014), which is quite expensive to obtain. To alleviate the difficulty of collecting data for training, weakly supervised learning (WSL) (Zhou 2017) is proposed for semantic segmentation. It makes use of weak annotations for training, e.g. bounding boxes (Dai, He, and Sun 2015; Khoreva et al. 2017), sparse scribbles (Lin et al. 2016; Vernaza and Chandraker 2017), and image-level class labels (Kolesnikov and Lampert 2016; Wei et al. 2017a; 2018; Ahn and Kwak 2018; Huang et al. 2018). In this paper, we focus on the most challenging problem by only using image-level labels for semantic segmentation.

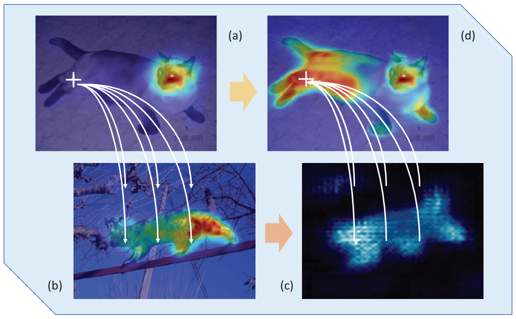

Figure 1: Illustration of the cross-image affinity. (a) the raw activation of a given image. (b) the activation of a reference image of the same class. (c) the affinity map from the marked query pixel in a) to reference b). (d) the activation after retrieving supplementary information according to the affinity.

The main difficulty of weakly supervised semantic segmentation is to recover the precise spatial information from only image tags. To this end, existing works usually rely on attention mechanisms, e.g. CAM (Zhou et al. 2016). However, these attention maps are derived from classification networks, which focus on the classification precision instead of the targets' integrity, thus only the most discriminative regions are obtained, which are sparse and incomplete.

To tackle this problem, Wei et al. (Wei et al. 2017a) adopt an iterative erasing strategy to mine complementary seeds. In the following works (Wei et al. 2018), they also apply multi-dilation convolution blocks to expand the seeds. Ahn and Kwak (Ahn and Kwak 2018) train additional pixel-level affinity net to complete the seeds. Huang et al. (Huang et al. 2018) dynamically fill in the undefined regions by seed region growing algorithm. Kolesnikov et al. (Kolesnikov and Lampert 2016) and Briq et al. (Briq, Moeller, and Gall 2018) take advantage of additional constraints to regularize the predictions. A common characteristic of these methods is that the images are treated independently of each other.

Contrastingly, we propose that the cross-image relationship is also vital for mining complete regions for weakly supervised segmentation. To intuitively understand this concept, see Fig. 1, there are two images of the same class, and

[*]Corresponding author

only partial regions are activated for each of them. We leverage the pixel-wise affinities and retrieve complementary information from one image to another, then more integral regions can be obtained. More formally, there are three major benefits of introducing the cross-image relationships:

Firstly, the cross-image relationship helps to provide supplementary information for identifying the pixels. Based on the supplementary components, related features can be refined and amplified to address some of the ambiguity and/or false predictions. Secondly, this explicit relationship helps the network to learn more consistent representations across the whole dataset. This is because the affinity module propagates related representations across different images, and a consistent one will be approached finally. Thirdly, through the cross-image relationship, the labels can be directly shared across a group of images, making better use of the valuable weak supervision.

Based on the motivation of building cross-image relationships, we propose an end-to-end cross-image affinity module, which can be directly plugged into existing segmentation networks. We name it cross-image affinity net (CIAN). CIAN explicitly models the pixel-level relationships among different images, efficiently leverages the relationships to refine the original representations and obtains more integral regions for segmentation. We conduct thorough experiments to demonstrate the effectiveness of the proposed approach. Our approach achieves new state-of-the-art results of weakly supervised semantic segmentation by only using image-level labels, with 64.3% mIoU on Pascal VOC 2012 (Everingham et al. 2010) validation set, and 65.3% on the test set. In summary, the main contributions are as follows:

- We firstly propose to explicitly model the cross-image relationship for weakly supervised semantic segmentation. An end-to-end cross-image affinity module is proposed to provide supplementary information from related images. By means of this, more integral regions can be obtained for weakly supervised segmentation.
- Extensive experiments demonstrate the usefulness of modeling cross-image relationships. Besides, we show that our approach is orthogonal to the quality of the seeds, which continually improves the training with even better seeds. Thus it can be potentially combined with future works that generate better seeds to further boost the performance.
- With the naive seeds generated by CAM, our CIAN achieves 65.3% mIoU on the VOC 2012 test set, which is a new state-of-the-art result by only using image-level labels for semantic segmentation, demonstrating the superiority of the approach.

Related Work

Weakly Supervised Semantic Segmentation. In this paper, we focus on the image-level label based weakly supervised semantic segmentation. State-of-the-art approaches follow a pipeline of first generating pseudo-masks (seeds) then training segmentation networks. The CAM (Zhou et al. 2016) is widely adopted as the cornerstone for generating seeds. However, CAM only activates the most discriminative regions, which is incomplete for segmentation.

To alleviate this problem, Wei et al. (Wei et al. 2017a) propose to adopt an iterative erasing strategy. MDC (Wei et al. 2018) proposes to merge multiple CAMs with different dilation rates. DSRG (Huang et al. 2018) proposes to dynamically fill in the sparse seeds by region growing. Wang et al. (Wang et al. 2018) propose to alternately train a superpixel based classification network and the segmentation network. Other works (Briq, Moeller, and Gall 2018; Kolesnikov and Lampert 2016) propose some heuristic constraints. The concurrent work FickleNet (Lee et al. 2019) randomly drops connections in each sliding window. Although these methods are effective, they ignore the rich relationships across different images, while we prove that the cross-image relationship is effective for obtaining consistent and integral regions for weakly supervised segmentation.

Co-segmentation. Co-segmentation aims to predict the common objects' masks for a given group of images (Chen, Huang, and Nakayama ; Li, Jafari, and Rother 2018; Li et al. 2018; Hsu, Lin, and Chuang 2018; Hsu et al. 2018). This task is related to ours since we also operate on a group of images to learn and utilize the cross-image relationship. The main difference between co-segmentation and weakly supervised segmentation is that co-segmentation focuses on finding class-agnostic masks for universal objects. When testing, it takes as input a group of images, so that common object can be defined. While our weakly supervised segmentation infers single images for known classes. Besides, many co-segmentation approaches are trained by strong pixel-level masks (Chen, Huang, and Nakayama ; Li et al. 2018). There is also weakly supervised segmentation by adopting co-segmentation to generate seeds (Shen et al. 2017), but it is different from ours that our affinity module is an end-to-end component for segmentation networks instead of the seeds.

Pixel-Level Affinity. Recent work AffinityNet (Ahn and Kwak 2018) samples sparse points by CAM seeds and trains an additional affinity net by metric learning. It is different from ours that our affinity module is an end-to-end component for dynamically sharing information across different images. In form, our method is closely related to the Non-local approaches (Wang et al. 2017; Fu et al. 2019; Yuan and Wang 2018), since pixel-level affinity is incorporated. However, Non-local networks focus on the long-range intra-image contexts to discover the hidden structures of a single image, while our approach aims at highlighting common objects and share complementary information across different images to combat the weak label problem.

Our Approach

In this section, we elaborate on all the components of the proposed approach. Following common practice, (Kolesnikov and Lampert 2016; Wei et al. 2018; Huang et al. 2018; Hou et al. 2018), we firstly generate initial seeds from image labels, then use them to train the segmentation network which is equipped with our proposed CIAN module. The framework is illustrated in Fig. 2.

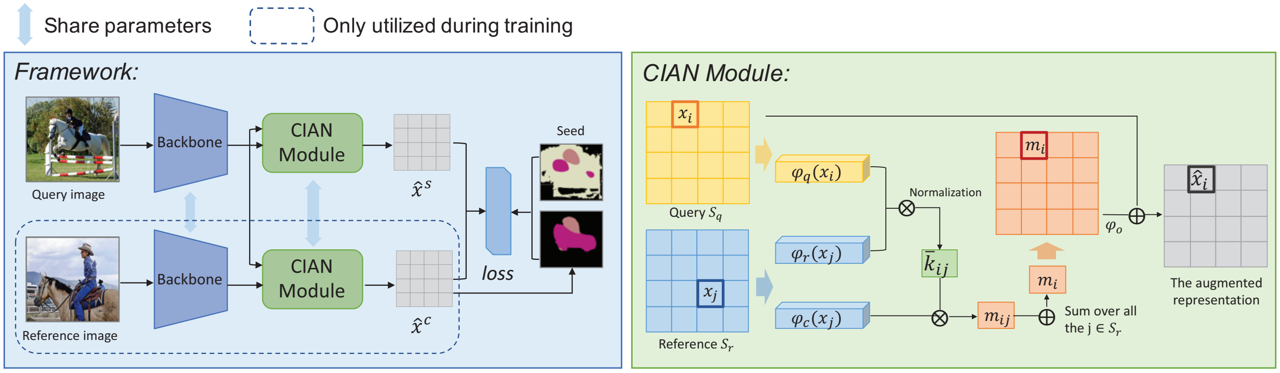

Figure 2: **Left**: the framework of CIAN. For simplicity, only one pair is drawn. Embedded features are obtained by a siamese backbone for both query and reference images, following with the cross-image affinity module to augment the features. For testing, reference images are unavailable, thus the query image only pairs to itself. **Right**: the proposed cross-image affinity module (CIAN module). The affinity is derived from the query and the reference features, then the query retrieves supplementary information from the reference accordingly.

Initial Seeds

Following the common practice (Kolesnikov and Lampert 2016; Wei et al. 2018; Huang et al. 2018; Hou et al. 2018), we adopt on-the-shelf CAM (Zhou et al. 2016) approach to generate initial seeds. It trains a classification network, which has a global-pooling layer right before the last classification layer. After training, it removes the global-average pooling layer and directly applies the classification layer to every pixel in the feature map to obtain the score map. However, because the classification tasks only focus on the most discriminate regions, only sparse and incomplete regions are highlighted. Our proposed CIAN can effectively alleviate the influence of the incomplete seeds with the help of cross-image relationships.

The Cross-Image Affinity Module

The affinity module models pixel-wise relationships between two images upon high-level representations. For simplicity, we use the term *pixels* to refer spatial-wise vectors of the feature map. Let S_q and S_r denote the sets of pixel indices of query image q and reference r, respectively. For any pair of pixels $\{(x_i, x_j)|i \in S_q, j \in S_r\}$ from the two images, the affinity score k_{ij} is modeled by:

$$k_{ij} = \exp\left(\varphi_q(x_i) \cdot \varphi_r(x_j)\right) \tag{1}$$

where φ_q and φ_r are learnable functions implemented by neural network layers, which can be seen as generalized kernel functions to enhance the encoding flexibility.

The next step is to retrieve supplementary information from reference x_j to query x_i. To this end, we first compress useful information by function φ_c, and weight it by the corresponding affinity score to get the *message* m_{ij}:

$$m_{ij} = \bar{k}_{ij}\varphi_c(x_j) \tag{2}$$

where, $\bar{k}_{ij}$ is the normalized version of affinity k_{ij}, to ensure that the sum of all the weights of the reference pixels is a unit, $\bar{k}_{ij} = k_{ij} / \sum_{j \in S_r} k_{ij}$.

Finally, all the messages from different reference pixels to the query x_i are summed together, normalized by function φ_o, and merged into the original representation x_i:

$$m_i = \sum_{j \in S_r} m_{ij} \tag{3}$$

$$\hat{x}_i = x_i + \varphi_o(m_i) \tag{4}$$

The above $\hat{x}_i$ is the so-called cross-affinity augmented representation. This process is repeated for all the pixels in the query image. The final classification layer takes as input augmented representations and outputs the segmentation results. The whole process is illustrated in Fig. 2.

On the one hand, meaningful affinity is forced to be learned to fit the existing seed supervision. On the other hand, the learned affinity bridges different images together to provide supplementary and/or complementary information. By means of this, the augmented representation and the affinity prompt each other and can be learned simultaneously.

Multiple Pairs

The above affinity module operates on a pair of two images. It can be easily extended to formulate relationships among multiple reference images and a single query image.

Given the query image q and its N reference partners $\{r^{(h)}|h = 1, ..., N\}$, we compute all the messages from the reference images to q according to Eq. 3, which are denoted as $\{m_i^{(h)}|i \in S_q; h = 1, ..., N\}$. Then, before adding them into the corresponding raw representation x_i, we merge all the messages from the multiple pairs. For example, it can be implemented by maximum function:

$$m_i = \max_{h \in \{1,...,N\}} m_i^{(h)} \tag{5}$$

Finally, m_i is normalized by φ_o and added to the original x_i, as in Eq. 4. Other merging functions, e.g. average function, are also available.

Training Loss

Cross-Entropy Loss. Following the common practice of semantic segmentation (Chen et al. 2014; 2018; Long, Shelhamer, and Darrell 2015), we adopt the pixel-wise cross-entropy loss to train the segmentation network. Use f_c to denote the final softmax classification layer of the segmentation network, the class probability of pixel $\hat{x}_i$ is then obtained by $f_c(\hat{x}_i) \in \mathbb{R}^C$, where C is the number of classes. For image $q : \hat{x} = \{\hat{x}_i | i \in S_q\}$, the cross entropy loss is defined as:

$$L_{ce}(\hat{x}) = -\frac{1}{|S'_q|} \sum_{i \in S'_q} y_i^T \log f_c(\hat{x}_i) \qquad (6)$$

where y_i is the onehot pseudo-label for pixel x_i obtained from the initial seeds. We use S'_q to denote the set of valid pixels for training in image q, since there are may pixels not assigned with labels by the seeds. The unassigned pixels are just ignored during training.

Completion Loss. As discussed above, the cross-affinity module augmented representations can provide more complete object estimations. Thus, it can be utilized to online generate pseudo-masks to compensate the sparsity of the initial seeds. To this end, we generate the online pseudo-label $\hat{y}_i \in \mathbb{R}^C$ from the prediction $f_c(\hat{x}_i)$:

$$\hat{y}_{i,l} = \mathbb{I}\left[(\arg\max f_c(\hat{x}_i) = l)\ \textbf{AND}\ (l \in L_q)\right] \qquad (7)$$

where, L_q is the set of image-level labels for image q, $\mathbb{I}[\cdot]$ is the index function and equals 1 if the statement is true. The meaning of Eq. 7 is that the l-th element of $\hat{y}_i$ is 1 iff it matches the prediction $f_c(\hat{x}_i)$ and the image-level labels.

Finally, we use $\hat{y}$ to optimize all the pixels in the query image q, and the whole procedure is named completion loss:

$$L_{cp}(\hat{x}) = -\frac{1}{|S''_q|} \sum_{i \in S''_q} \hat{y}_i^T \log f_c(\hat{x}_i) \qquad (8)$$

where, S''_q is the set of effective pixels according to $\hat{y}$.

The Overall Loss. To ensure there do exist usable information across images, we need common-class pairs for training. However, it is hard to sample pairs with identically the same classes, since there can be multiple classes in a single image thus faces a combinatorial explosion problem. Therefore, we relax to sample images with at least one common class as pairs. To reduce the influence of possibly unmatched classes, we utilize the *self-affinity*, which simply adopts the above affinity module to the image itself, i.e. $x_q = x_r$. We denote the representations augmented by the cross-affinity and the self-affinity as $\hat{x}^c$ and $\hat{x}^s$, respectively. The overall loss is computed as:

$$L = L_{ce}(\hat{x}^c) + L_{ce}(\hat{x}^s) + L_{cp}(\hat{x}^c) + L_{cp}(\hat{x}^s) \qquad (9)$$

Another important reason to use the self-affinity is that during testing we cannot make pairs since the deployed model should be able to address single images. Directly removing the affinity residual incurs heavy distribution change of representations, thus is unfeasible. Instead, we address this problem by augmenting the representation with the self-affinity during testing. To this end, $L_{ce}(\hat{x}^s)$ and $L_{cp}(\hat{x}^s)$ are minimized to further reduce the training and testing gap.

Experiments and Analysis

Datasets

We evaluate our proposed method on Pascal VOC 2012 segmentation benchmark (Everingham et al. 2010). This is the standard dataset for weakly supervised semantic segmentation. It has 20 foreground classes and one background class. A single image may contain multiple classes. Following the common practice (Wei et al. 2017a; 2018; Huang et al. 2018), we use the expanded set collected by Hariharan et al. (Hariharan et al. 2011), i.e., there are 10582 training images, 1449 validation images, and 1456 testing images. In our experiments, only the image class labels are used for training. The performance is evaluated by mean intersection over union (mIoU) of all the 21 classes.

Implementation Details

Initial Seeds. As aforementioned, we adopt CAM to generate initial seeds. Specifically, it uses ImageNet pre-trained VGG16. To obtain larger maps, we replace the last two pooling layers with stride 1 and use dilation rate 2 in the Conv5 block. We train the CAM with the multi-class sigmoid loss with learning rate $1e^{-3}$. We normalize the CAM into range $[0, 1]$ and generate foreground regions by threshold 0.3. Following related works (Kolesnikov and Lampert 2016; Wei et al. 2018; Huang et al. 2018; Hou et al. 2018), we use an off-the-shelf saliency model (Jiang et al. 2013) to generate background seeds by threshold 0.06. Finally, all the remaining unassigned pixels and conflictual assignments are abandoned and marked as ignored.

CIAN Module. We choose Deeplab-V2-Largefov (Chen et al. 2018) framework for segmentation. The CIAN module takes as input the feature maps right before the classification layer. φ_q, φ_r and φ_c are all implemented by single 1×1 convolution layers. To speedup the computation, φ_q and φ_r halve the feature dimensions, $\varphi_r(x)$ and $\varphi_c(x)$ are spatially downsampled by max-pooling with stride 2. During training, we randomly sample reference images for each query image. We experimentally find that a single reference image for each query is adequate for learning cross-image relationships, more pairs bring negligible improvement. This may because the pairs are randomly sampled and all the potential combinations can be visited along the training process. To stabilize the training, the image to itself is also adopted as a pair and merged by Eq. 5. Since the affinity is unreliable during the initial training stage, a zero-initialized batch normalization layer is attached to φ_o before adding it to original representations. Following (Huang et al. 2018), we also adopt the retraining strategy, i.e. generate predictions by currently trained network, and take these predictions as pseudo-labels to train the network with another round.

	# Params	# FLOPs
Vanilla Deeplab	42.4 M	72.5 G
Baseline	50.8 M	88.2 G
Ours	50.8 M	89.7 G

Table 1: Comparison of the number of parameters and computation complexity. The values are based on ResNet101.

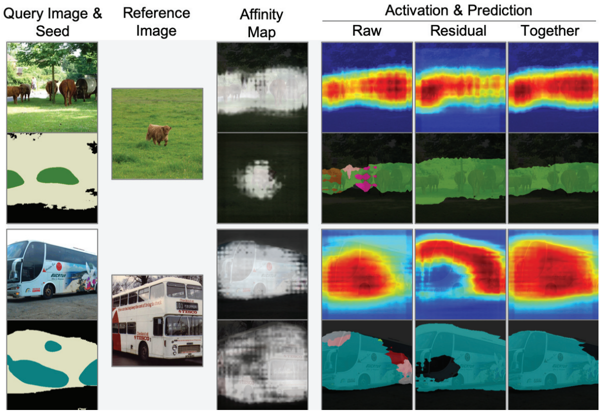

Figure 3: Visualization of the affinity. Two typical pairs are illustrated. The first two columns are the query and reference respectively. The third column is the average affinity map for the self- and the cross-affinity respectively. The last three columns show the activation and predictions of the raw, residual, and augmented representations, respectively.

Reproducibility. All the backbones are pre-trained by ImageNet, and the newly added layers are initialized by random Normal with the standard deviation 0.01. We adopt the SGD optimizer with an initial learning rate $5e^{-4}$ and momentum 0.9, which is poly-decayed by power 0.9. We use batch size 16 to train 20 epochs with randomly cropped images of size 321. Standard data augmentation, i.e., random cropping, scaling, and horizontal flipping are adopted. CRF (Krähenbühl and Koltun 2011) with default parameters for post-processing. Codes are implemented with MXNet (Chen et al. 2015), and are available at: https://github.com/js-fan/CIAN.

Computation Complexity

Our CIAN module works at the top layer with relatively small spatial size, thus the overhead is marginal compared with the standard segmentation networks, as shown in Table 1. Besides, the computation complexity of our CIAN is the same as the self-affinity augmented baseline during testing, and the overhead only happens during training.

An Intuitive View of The Cross-Image Affinity

To help the reader further understand how the cross-image affinity helps, we visualize the learned maps in Fig. 3. Since the affinity component is an addable residual to the raw representation, we visualize the activation of the raw representation, the cross-affinity residual, and the final augmented representation, respectively, as shown in the last three columns in Fig. 3.

We can see that the affinity maps (in the third column) focus on correlated target object regions. The first example shows that the spatial distribution of the retrieved residual's activation is similar to the raw. After merging, the representation is further strengthened, thus some of the false predictions can be avoided. The second example shows that the cross-image residual retrieves complementary activation, and thus makes more integral final predictions. This case illustrates that different images may be activated with different parts, and the learned cross-image affinity can help to share complementary information among images to achieve more complete and consistent estimations.

Ablation Studies

We conduct thorough experiments to demonstrate the advantage of the CIAN module. By default, ResNet101 is adopted as the backbone, and results are evaluated on Pascal VOC 2012 validation set.

The Effect of The CIAN Module. We first prove that the learned cross-image relationship benefits the weakly supervised segmentation task. For fair comparisons, the *baseline* model is also augmented by the affinity module, with the limitation that no cross-image pairs are available, i.e., only self-affinity is applied for the baseline. By means of this, the capacity of the baseline model is the same as our CIAN, and we can conclude that the improvement actually comes from the cross-image relationships exploited during training.

The ablation results are shown in Table 2. Starting with the baseline, with the naive cross-entropy loss for the cross-image representations (+CE), the model achieves 0.8% improvement. By further adopting the completion loss with the cross-image representations (+CP), the model achieves another 3.4% improvement. This improvement is much more than the former because the cross-affinity augmented representations provide more complete pseudo-masks than the initial seeds. Even though the cross-affinity can directly refine the representations, with only sparse regions from the initial seeds, much useful information will be abandoned and does not participate in the optimization due to the enormous number of ignored pixels. Thus, the completion loss is necessary for mining cross-image relationships and fully utilizing the retrieved complementary information. Finally, the retraining strategy recovers boundaries by the CRF refinement when generating predictions, thus the result is further improved by 1.8% (+RT) and achieves the cutting-edge performance, as shown in Table 5.

The Query-Reference Pairs In our CIAN, the query-reference pairs are sampled that there is at least one common class. If not, e.g. assume that we randomly sample reference images to the query without any class label constraint, reliable relationships would not be learned. From this point of view, we can also demonstrate the usefulness of valid cross-image relationships.

Method	bkg	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mIoU
Baseline	85.0	73.7	33.5	74.6	49.1	63.0	77.5	68.6	65.9	22.7	63.0	**40.2**	58.4	64.5	69.9	57.1	35.9	74.3	**34.5**	61.1	50.9	58.3
+ CE	84.9	75.6	**34.1**	72.4	45.1	65.8	79.1	67.6	68.7	21.1	66.9	39.7	62.4	72.5	71.4	58.7	34.4	75.8	32.5	61.2	52.0	59.1
+ CP	87.1	76.0	**34.1**	73.1	49.7	69.5	83.5	**73.9**	77.5	27.2	70.9	37.1	72.4	74.6	73.6	62.5	42.7	76.6	34.2	**63.6**	**53.5**	62.5
+ RT	**88.2**	**79.5**	32.6	**75.7**	**56.8**	**72.1**	**85.3**	72.9	**81.7**	**27.6**	**73.3**	39.8	**76.4**	**77.0**	**74.9**	**66.8**	**46.6**	**81.0**	29.1	60.4	53.3	**64.3**

Table 2: Comparison on the VOC 2012 val set for ablation study. The baseline is only augmented by self-affinity. (+ CE) denotes to adopt cross-image affinity with the cross-entropy loss. (+ CP) denotes to further adopt the completion loss. (+ RT) denotes to adopt the retraining strategy.

Method	Common-class	Random-class
+ CE	**59.1**	58.5
+ CP	**62.5**	59.7

Table 3: Comparison on the VOC 2012 val set between the random-class pairs and our common-class pairs. In random-class pairs, images may not have any common classes, while in our common-class pairs they have at least one common class.

Ratio		ResNet101			ResNet50		
		baseline	ours	Delta	baseline	ours	Delta
CRF	0	58.3	62.5	+ 4.2	57.5	60.4	+ 2.9
	0.05	61.4	65.8	+ 4.4	60.2	63.5	+ 3.3
	0.10	62.2	67.3	+ 5.1	61.6	65.7	+ 4.1
no CRF	0	53.8	58.1	+ 4.3	52.7	56.0	+ 3.3
	0.05	57.7	61.2	+ 3.5	56.4	59.4	+ 3.0
	0.10	60.3	64.1	+ 3.8	58.4	62.1	+ 3.7

Table 4: Comparison on the VOC 2012 val set with different ratios of strong supervision. The larger ratio corresponds to better seeds. Our approach continually improves over the baseline with different seeds. Results both with or without CRF post-processing are given.

To this end, in contrast to the proposed common-class sampling strategy, we train the CIAN model with random-class pairs. The results are summarized in Table 3. It shows that with random-class pairs, the performance is in line with the baseline model, which is much worse than our common-class counterpart.

We notice that with random-class pairs and completion loss, there is still 1.4% improvement compared with the baseline. This is because although there are no valid cross-image pairs, the network's online predictions still provide better pseudo-masks than initial sparse seeds. However, without valid cross-image relationships, the completion is limited and inferior. As a comparison, our CIAN with reliable common-class pairs outperforms the baseline by 4.2%, which is much better than the random-class pair's 1.4%. This result also reveals that only adopting the online completions (as in Eq.8) without available cross-image relationships does not fully utilize the information.

Orthogonal to The Seed Quality We prove that our approach does not rely on occasionally generated initial seeds. Indeed, the CIAN module consistently improves over the baseline with even stronger seeds. Therefore, our approach is orthogonal to those state-of-the-art approaches (Wei et al. 2018; 2017a; Hou et al. 2018), which generate better initial seeds. To quantitatively assess the orthogonality, we imitate a group of better seeds by randomly substituting a portion of the seeds with the ground truth, which is similar to the setting of semi-supervised learning. Specifically, 5% or 10% of the 10582 training seeds are substituted, respectively.

As shown in Table 4, with 5% and 10% of the seeds substituted, our CIAN outperforms the baseline by 4.4% and 5.1% respectively. Similar improvements are achieved with the ResNet 50 backbone. Our approach consistently brings significant improvement with better seeds. Therefore, this approach can be potentially fused with those works generating stronger seeds.

Comparison with State-of-The-Art

It should be careful to make comparison with other approaches, because they may leverage additional supervision, different pre-trainings, or adopt different backbones. We summarize the state-of-the-art results and list their difference in Table 5.

Our approach with ResNet101 achieves mIoU score of 64.3% and 65.3% on VOC12 val and test set respectively, outperforming all of the previous results by only using image-level labels. It is surprising that our result outperforms some early fully supervised works like FCN (Long, Shelhamer, and Darrell 2015). In spite of the different backbones, our result is comparable with some works with stronger supervisions, e.g. box-supervised (Dai, He, and Sun 2015) and scribble-supervised (Lin et al. 2016) approaches. With the same backbone and training set, our approach outperforms AISI (Fan et al. 2018), which relies on a well-trained instance saliency network. Note that instance saliency is trained by pixel-level annotated instance masks, which is quite costly to obtain. Besides, our result with ResNet50 is comparable with the state-of-the-arts with ResNet101 (Wang et al. 2018; Huang et al. 2018; Hou et al. 2018), demonstrating the advantage of leveraging cross-image relationships.

Qualitative Results

To help better understand the final effect of the cross-image affinity module to the predictions, we visualize some of the typical predictions of both the baseline and our CIAN, as shown in Fig. 4. The first two rows show that by the CIAN module that borrows information from other images, some of the missing parts can be completed. The next two rows

Methods	Sup.	val	test
Fully supervised			
FCN†(Long, Shelhamer, and Darrell 2015)	*F.*	-	62.2
Deeplab†(Chen et al. 2014)	*F.*	67.6	70.3
Weakly supervised			
BoxSup†(Dai, He, and Sun 2015)	*L.+B.*	62.0	64.6
ScribbleSup†(Lin et al. 2016)	*L.+S.*	63.1	-
AISI (Fan et al. 2018)	*L.+I.*	63.6	64.5
CCNN†(Pathak, Krähenbühl, and Darrell 2015)	*L.*	35.3	35.6
EM-Adapt†(Papandreou et al. 2015)	*L.*	38.2	39.6
STC†(Wei et al. 2017b)	*L.*	49.8	51.2
SEC†(Kolesnikov and Lampert 2016)	*L.*	50.7	51.7
AugFeed†(Qi et al. 2016)	*L.*	54.3	55.5
AE-PSL†(Wei et al. 2017a)	*L.*	55.0	55.7
GuidedSeg†(Oh et al. 2017)	*L.*	55.7	56.7
DCSP (Chaudhry, Dokania, and Torr 2017)	*L.*	60.8	61.9
AFFNet†(Ahn and Kwak 2018)	*L.*	58.4	60.5
MDC†(Wei et al. 2018)	*L.*	60.4	60.8
MCOF (Wang et al. 2018)	*L.*	60.3	61.2
DSRG (Huang et al. 2018)	*L.*	61.4	63.2
SeeNet (Hou et al. 2018)	*L.*	63.1	62.8
Ours:			
CIAN (res50)	*L.*	62.4	**63.8**
CIAN (res101)	*L.*	**64.3**	**65.3**

Table 5: Comparison of state-of-the-arts on VOC 2012. Methods marked by † use VGG16, the others use ResNet101. The supervision (Sup.) includes: image-level label (*L.*), instance saliency (*I.*), bounding box (*B.*), scribble (*S.*) and full supervision (*F.*).

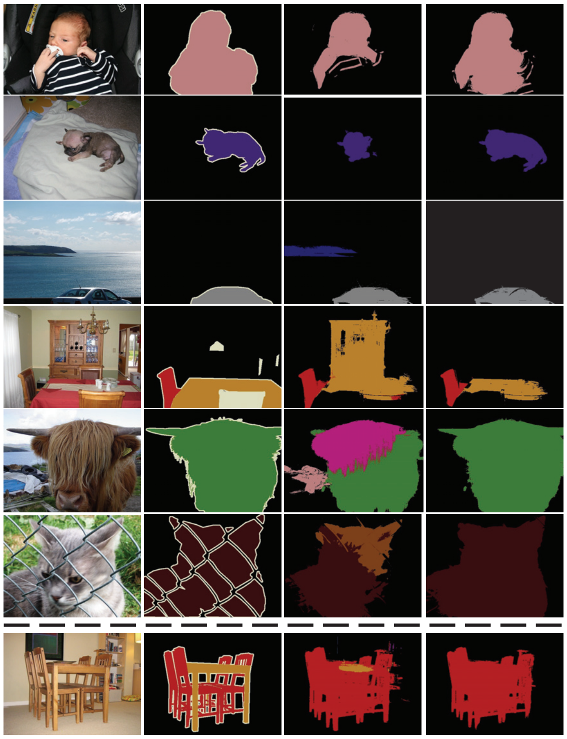

Figure 4: Visualization of the predicted results on VOC 2012 val set. The first six rows show the typical cases to demonstrate how the CIAN outperforms the baseline. The last row shows a typical failure case where multiple classes are of similar appearance. Best viewed in color.

show that some false positive predictions can be inhibited because they are never matched in any reference images. The following two rows demonstrate that the module helps to exclude clutter. This is because cross-image relationships help the network to learn more consistent representations across the whole dataset, and the related representations can be strengthened by each other, thus reduces the noise. The last row shows a typical failure case that interweaving objects with similar appearance and small spatial scale are confused and wrongly optimized. We leave it for future studies to address this problem.

Conclusion

In this paper, we propose to leverage cross-image relationships for weakly supervised semantic segmentation. We propose an end-to-end CIAN module to build pixel-level affinities across different images, which can be directly plugged into existing segmentation networks. With the help of cross-image relationships, incomplete regions can retrieve supplementary information from other images to obtain more integral object region estimations and rectify the ambiguity. Extensive experiments demonstrate the advantage of utilizing cross-image relationships. Besides, our approach achieves state-of-the-art performance on VOC 2012 semantic segmentation task with only image-level labels.

Acknowledgements

This work was supported in part by the National Key R&D Program of China (No. 2018YFB1402600, No. 2018YFB1402605), the National Natural Science Foundation of China (No. 61836014, No. 61761146004, No. 61773375, No. 61721004), the Beijing Municipal Natural Science Foundation (No. Z181100008918010) and CAS-AIR.

References

Ahn, J., and Kwak, S. 2018. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. *arXiv:1803.10464*.

Briq, R.; Moeller, M.; and Gall, J. 2018. Convolutional simplex projection network (cspn) for weakly supervised semantic segmentation. *arXiv:1807.09169*.

Chaudhry, A.; Dokania, P. K.; and Torr, P. H. 2017. Discovering class-specific pixels for weakly-supervised semantic segmentation. *arXiv:1707.05821*.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2014. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv:1412.7062*.

Chen, T.; Li, M.; Li, Y.; Lin, M.; Wang, N.; Wang, M.; Xiao, T.; Xu, B.; Zhang, C.; and Zhang, Z. 2015. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. *arXiv:1512.01274*.

Chen, L.-C.; Papandreou, G.; Schroff, F.; and Adam, H.

2017. Rethinking atrous convolution for semantic image segmentation. *arXiv:1706.05587*.

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. *TPAMI* 40(4):834–848.

Chen, H.; Huang, Y.; and Nakayama, H. Semantic aware attention based deep object co-segmentation. *arXiv:1810.06859*.

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

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *IJCV*.

Fan, R.; Hou, Q.; Cheng, M.-M.; Yu, G.; Martin, R. R.; and Hu, S.-M. 2018. Associating inter-image salient instances for weakly supervised semantic segmentation. In *ECCV*.

Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; and Lu, H. 2019. Dual attention network for scene segmentation. In *CVPR*.

Hariharan, B.; Arbeláez, P.; Bourdev, L.; et al. 2011. Semantic contours from inverse detectors. In *ICCV*.

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

Hsu, K.-J.; Tsai, C.-C.; Lin, Y.-Y.; Qian, X.; and Chuang, Y.-Y. 2018. Unsupervised cnn-based co-saliency detection with graphical optimization. *ECCV*.

Hsu, K.-J.; Lin, Y.-Y.; and Chuang, Y.-Y. 2018. Co-attention cnns for unsupervised object co-segmentation. *IJCAI*.

Huang, Z.; Wang, X.; Wang, J.; Liu, W.; and Wang, J. 2018. Weakly-supervised semantic segmentation network with deep seeded region growing. In *CVPR*.

Jiang, H.; Wang, J.; Yuan, Z.; Wu, Y.; Zheng, N.; and Li, S. 2013. Salient object detection: A discriminative regional feature integration approach. In *CVPR*.

Khoreva, A.; Benenson, R.; Hosang, J.; Hein, M.; and Schiele, B. 2017. Simple does it: Weakly supervised instance and semantic segmentation. In *CVPR*.

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

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

Lee, J.; Kim, E.; Lee, S.; Lee, J.; and Yoon, S. 2019. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In *CVPR*.

Li, M.; Dong, S.; Zhang, K.; Gao, Z.; Wu, X.; Zhang, H.; Yang, G.; and Li, S. 2018. Deep learning intra-image and inter-images features for co-saliency detection. *BMVC*.

Li, W.; Jafari, O. H.; and Rother, C. 2018. Deep dbject co-segmentation. In *Asian Conference on Computer Vision*, 638–653. Springer.

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

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

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

Oh, S. J.; Benenson, R.; Khoreva, A.; Akata, Z.; Fritz, M.; and Schiele, B. 2017. Exploiting saliency for object segmentation from image level labels. In *CVPR*.

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:1502.02734*.

Pathak, D.; Krähenbühl, P.; and Darrell, T. 2015. Constrained convolutional neural networks for weakly supervised segmentation. In *ICCV*, 1796–1804.

Qi, X.; Liu, Z.; Shi, J.; Zhao, H.; and Jia, J. 2016. Augmented feedback in semantic segmentation under image level supervision. In *ECCV16*.

Shen, T.; Lin, G.; Liu, L.; Shen, C.; and Reid, I. 2017. Weakly supervised semantic segmentation based on web image co-segmentation. *arXiv:1705.09052*.

Vernaza, P., and Chandraker, M. 2017. Learning random-walk label propagation for weakly-supervised semantic segmentation. In *CVPR*.

Wang, X.; Girshick, R.; Gupta, A.; and He, K. 2017. Non-local neural networks. *arXiv:1711.07971* 10.

Wang, X.; You, S.; Li, X.; and Ma, H. 2018. Weakly-supervised semantic segmentation by iteratively mining common object features. In *CVPR*, 1354–1362.

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 *CVPR*.

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. *TPAMI*.

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 *CVPR*.

Yuan, Y., and Wang, J. 2018. Ocnet: Object context network for scene parsing. *arXiv:1809.00916*.

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

Zhou, Z.-H. 2017. A brief introduction to weakly supervised learning. *National Science Review* 5(1):44–53.