

Universiteit
Leiden
The Netherlands

Using deep learning models in image processing

Xiong, Z.

Citation
Xiong, Z. (2026, May 19). *Using deep learning models in image processing*. Retrieved from https://hdl.handle.net/1887/4303646

Version:	Publisher's Version
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/4303646

Note: To cite this publication please use the final published version (if applicable).

Chapter 4

Weakly Supervised Semantic Segmentation via Contextual Cross-Attention

Based on the following publication:

Zhan Xiong and Fons J. Verbeek. Weakly Supervised Semantic Segmentation via Contextual Cross-Attention. In *9th International Conference on Frontiers of Signal Processing, ICFSP 2024, Paris, France, September 12-14, 2024*, pages 83–92. IEEE, 2024

Chapter Summary

Weakly supervised semantic segmentation (WSSS) using image-level labels faces two challenges: Class Activation Maps (CAMs) lack explicit background modeling and lose spatial information during pooling. However, post-processing for background inference often results in imprecise separation. Moreover, the absence of geometric regularization in standard loss functions hinders accurate boundary delineation. To address these limitations, we propose a novel end-to-end WSSS framework that integrates a conditional probabilistic model with a geometric prior. Our framework is initiated by explicitly modeling the background using Bernoulli distributions trained with contrastive self-supervised learning. We then introduce a variational total variation (TV) term to refine background regions and improve boundary fidelity. Foreground class probabilities are subsequently modeled with Categorical distributions conditioned on the refined background. This conditional foreground modeling is implemented as a novel Contextual Cross-Attention (CCA) module that captures long-range dependencies among foreground regions. Our hierarchical framework offers insights into WSSS representation learning and significantly improves segmentation accuracy on the PASCAL VOC dataset.

Keyword: Conditional Probabilistic Model Image-Level Classification Cross Attention Contrastive Learning Global-Local Prototype Transformation

4.1 Introduction

Semantic segmentation is essential, but its application is often constrained by the request for pixel-level annotations, particularly in fields like biology and medicine, where such annotations require domain expertise. Weakly Supervised Semantic Segmentation (WSSS), which uses weaker supervision such as image-level labels, offers a promising alternative. However, image-level WSSS faces inherent limitations, including **ambiguous background modeling** due to the lack of explicit indicators and **loss of spatial details** from pooling operations (e.g., max or global average pooling). These limitations complicate the discrimination between precise and imprecise partitions.

Existing WSSS approaches address background modeling through discriminative and generative paradigms. Discriminative approaches, which define the background as the complement of foreground classes, employ a one-against-all classification strategy to allocate pixels to the background by exclusion. Therefore, the background cannot be conceptualized as **simple** as other foreground classes, whose representation cannot be directly modeled in a finite feature space but only rationally defined as feature vectors outside the foreground space, depicted in Figure 4.1. Despite their effectiveness, existing models[734, 158, 100, 315, 556, 862] often rely on class activation maps (CAMs) [564] for foreground representation. However, these methods often resort to ad-hoc post-processing for background inference, leading to false background responses. In contrast, generative methods [356, 343] adopt a one-vs-one classification strategy and leverage adversarial learning to simplify segmentation into binary classifications for each foreground class. However, they may undermine spatial exclusivity among classes and slow inference in multi-class scenarios.

Despite emphasizing background modeling, image-level WSSS methods often overlook issues of spatial information loss. Current approaches fail to ensure that loss functions favor accurate partitions, especially when there are fewer foreground classes [802]. The crux comes from the pooling operations that map inaccurate partitions to feature vectors similar to the correct ones, making them indistinguishable to image-level loss functions. Consequently, there is a need for geometrical regularization to align partitions with smooth boundaries of foreground regions, depicted in Figure 4.2. However, existing methods mainly use post-processing techniques such as thresholding or dense conditional random fields (dense-CRF), which are dataset-specific and rely on parameter tuning.

To address these limitations within a discriminative framework, we propose a novel conditional probabilistic model coupled with a variational geometric prior, depicted in

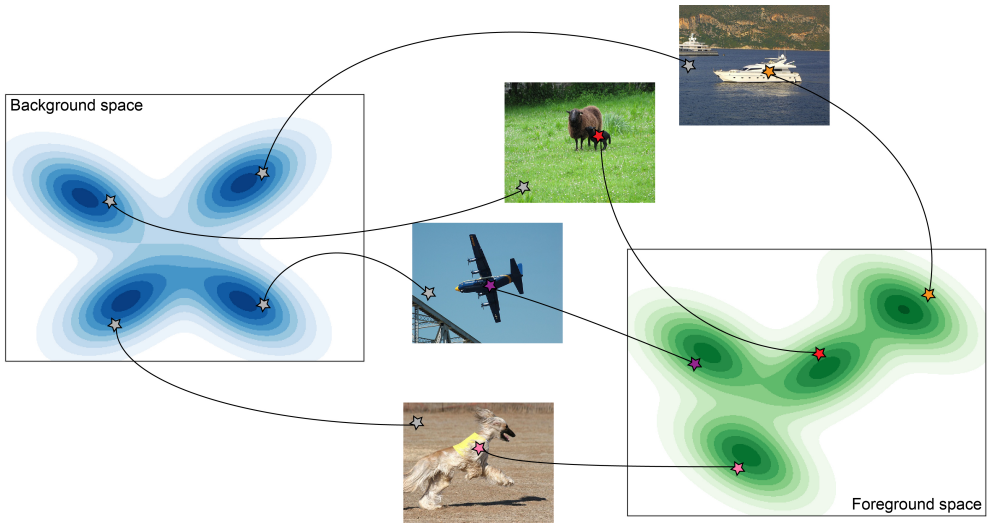


Figure 4.1: As the semantic information of foreground objects differs from that of backgrounds, the representative manifold is divided into two sub-spaces: foreground (green) and background (blue) spaces. All foreground objects share similar high-level representations and can be mapped to the same manifold. However, due to inter-class differences, they reside in different semantic clusters (color stars). All objects outside the foreground space belong to the background class. With different color/texture low-level features, they can also reside in different sub-groups (grey stars).

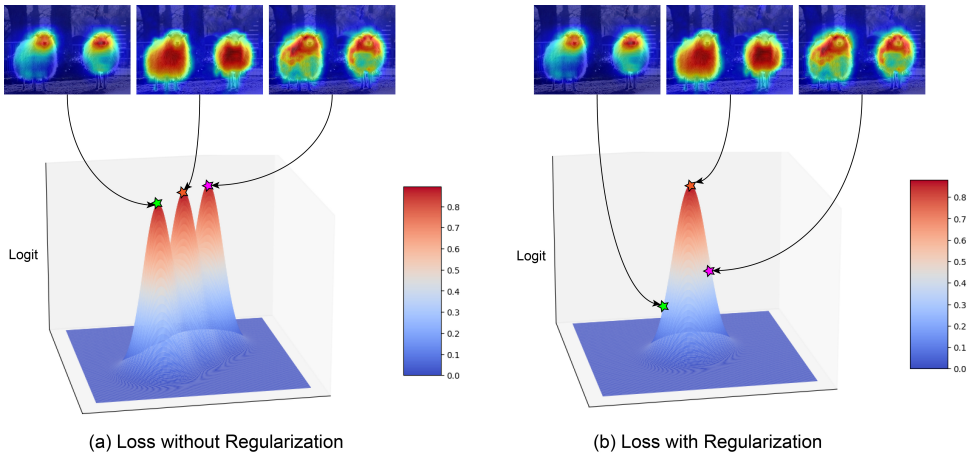


Figure 4.2: An illustration to show the role of geometrical regularization. Three partitions are qualified as bad (green star), good (orange star), and neutral (purple star) from left to right. (a) All partitions have similar logits in a standard image-level loss function. (b) With regularization, they are re-ranked with different logits according to their qualities.

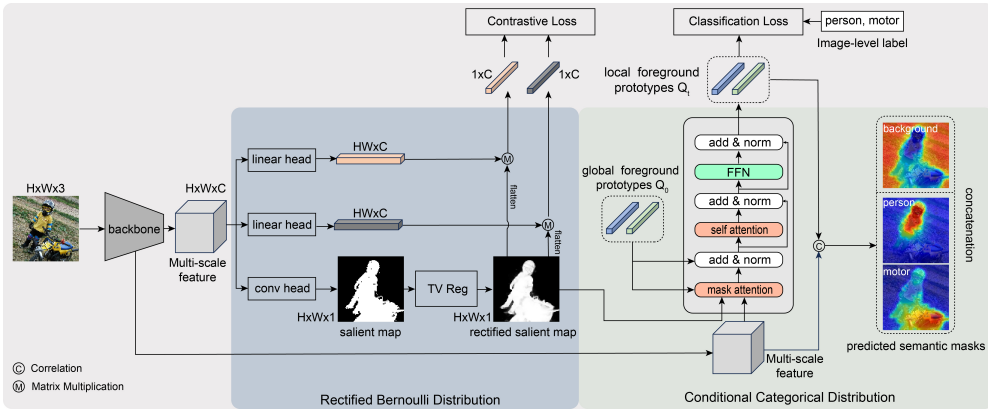


Figure 4.3: An Overview of the proposed framework. It consists of a rectified Bernoulli distribution and a conditional Categorical distribution. Global foreground prototype Q_0 is iteratively transformed into local foreground prototypes Q_t. Note that salient maps denote the complement of the background maps for better visualization. The final predicted semantic masks are obtained by concatenating a rectified background map and conditional foreground classification masks.

Figure 4.3. Our approach reformulates semantic segmentation as a hierarchical model. We first explicitly model the background classification as an independent Bernoulli distribution[1]. The Bernoulli distribution is implemented as a sub-network using the sigmoid activation function. To work around the missing background indication information, we introduce foreground-background contrastive self-supervised learning [87, 158, 714] for training instead. Contrastive learning can automatically mine complementary information of background representations across images without requesting image tags. However, the initial background regions generated from self-attention might not be smooth and compact. Therefore, we introduce a variational geometric regularization term [56] to promote segmentation boundary fidelity. We argue that it is empirically rational to only regulate between background and foreground because **each foreground object should have a regular geometrical shape**. Therefore, the union of these regions is still regularly continuous. This regularization encourages coherent region delineations with smooth boundaries, allowing regions to expand / contract to better cover objects. The assumption is empirically proven to be a simple yet efficient way to address spatial information loss with minimal computational overhead. The Bernoulli distribution with the geometrical regularization is called the rectified Bernoulli distribution. For foreground classification, we employ a conditional Categorical distribution within the identified foreground regions, enforcing soft pixel

[1]Strictly speaking, it is conditional independence when given previous layers.

assignments among foreground classes. Our conditional Categorical distribution can be formulated as a stack of masked cross-attention modules [105]. Each module captures long-range dependencies among foreground partitions only. We name the conditional Categorical distribution as contextual cross-attention (CCA). The CCA can extract global foreground prototypes from the dataset and gradually transform them into local prototypes based on the input content. Because the local prototypes are adapted to the features of the corresponding input, the soft label assignment can be computed as probabilities by normalizing the pairwise similarities between pixels and local prototypes, and it benefits from reduced variation compared to global ones. Our framework offers a new perspective on the structure of representative embedding spaces under the WSSS paradigm.

In summary, our main contributions are as follows:

- An end-to-end WSSS framework offering a new understanding of representative embedding spaces. Our model uses contrastive learning for explicit background modeling and a novel contextual cross-attention for foreground classification.
- A variational geometric regularization for smooth region delineation, efficiently addressing spatial information loss and improving segmentation masks with minimal computational overhead.
- Our method is a generic approach that can easily extend with other regularization terms and demonstrates superior performance on the benchmark dataset.

4.2 Related Work

Image-level Discriminative WSSS. Discriminative approaches to image-level WSSS often leverage CAMs [564] to identify foreground and background regions. However, CAMs can result in incomplete partitions and imprecise boundaries. Research has focused on expanding CAMs into more robust semantic masks to address this issue. One approach involves global-to-local prototype transformations [62, 856, 734, 83, 862, 100] to mitigate non-discriminative regions. Others integrate complementary information, such as semantic affinities [3], boundary information [550], uncertainty [394], or exploit information bottlenecks [355]. Some efforts mitigate challenges through data manipulation, such as introducing hard out-of-distribution data [357] or strong augmentations [597]. While these methods improve foreground modeling, they often rely on heuristic post-processing for background estimation, potentially limiting precise semantic seg-

mentation.

Contrastive Learning. The core idea of contrastive learning is to group feature vectors based on their similarities within a feature space, enforcing intra-group similarity and inter-group difference. Based on object class labels, this approach creates positive pairs from samples within the same class and negative pairs from those across different classes. Hence, contrastive learning [87] pulls the samples from positive pairs closer together and pushes the samples from negative pairs farther apart. Recent WSSS research uses contrastive learning to extract additional information across images, including pixel-to-prototype contrast [158], knowledge distillation [626], and biased object removal [315]. These complementary cues can assist in accurate semantic classification.

Cross-Attention. Cross-attention computes the similarities between pixels and queries. Some studies [773] implement cross-attention as k-means clustering, while others [663] use a K-NN clustering perspective. We can interpret image-level WSSS as a form of clustering, with object queries serving as cluster centers and their learnable embeddings. In this chapter, we adopt a k-means clustering interpretation and reformulate cross-attention as a conditional probability within our framework to predict pixel-level similarities. Therefore, cross-attention can be a scalable module that transforms cluster centers into different representations.

Regularization. Regularization prevents model underfitting or overfitting by introducing additional information. It is crucial for image-level WSSS because a classifier maps different partitions onto similar vectors. From a Bayesian perspective, one regularization term is a prior distribution reflecting *a priori* knowledge. Most recently, adversarial perturbations [356] have been applied to bridge the train-test error gap. However, existing regularization terms cannot directly guide image-level loss functions in evaluating partitions at the pixel level. To address this, we introduce a total variation (TV) regularization term to encourage partition alignment with object boundaries.

4.3 Method

In this section, we reformulate image-level WSSS as a hierarchical probabilistic framework in an end-to-end manner and provide the solutions of each component approximated by mean-field theory. Given a feature map $X \in \mathbb{R}^{C\times H\times W}$, where C, H and W denote the channel number and spatial dimensions, our model predicts a pixel-level semantic label assignment $S \in \mathbb{R}^{(K+1)\times H\times W}$, where K is the number of **foreground**

classes. We treat the background as an auxiliary class and split label vectors with two distinct parts: a scalar set $B = \{b_i | b_i \in \{0, 1\}\}$ for the background class, and a conditional 1-of-K vector set $F = \{(1 - b_i) \cdot \vec{f}_i | f_{i,k} \in \{0, 1\}, \sum_k f_{i,k} = 1\}$ for foreground classes. Here, $i \in \{H \times W\}$ indexes spatial locations, and $f_{i,k}$ is the label of the k-th foreground class at position i. To enhance segmentation boundary fidelity, we introduce an auxiliary term $\tilde{B} \in \mathbb{R}^{1 \times H \times W}$ as the initial binary background map, and the geometrical prior distribution uses $\tilde{B}$ to regulate the background regions.

4.3.1 Probabilistic Framework

We offer a novel perspective on semantic label assignment. In contrast to the conventional manner, we define it as a structural process and formulate it as a conditional probabilistic model:

$$P(S, \tilde{B}|X) = P(B, F, \tilde{B}|X) = P(F|B, X) \cdot P(B, \tilde{B}|X) \tag{4.1}$$

The $P(B, \tilde{B}|X)$ is defined as a rectified Bernoulli distribution to generate binary background maps with compact and smooth boundaries. Moreover, $P(F|B, X)$ denotes a conditional 1-of-K Categorical distribution to assign semantic labels within identified foreground regions.

Rectified Bernoulli Distribution

The rectified Bernoulli distribution is also a conditional probability:

$$P(B, \tilde{B}|X) = P(B|\tilde{B}) \cdot P(\tilde{B}|X) \tag{4.2}$$

$P(\tilde{B}|X)$ represents Bernoulli distributions that generate initial pixel-level background probability maps, and $P(B|\tilde{B})$ represents a geometrical prior that refines the initial predictions into more compact and smooth regions.

Denote a binary classifier as $h(\cdot)$, which inputs the feature map X. Then, the probability of being background is defined as a sigmoid function:

$$p_i = \text{Sigmoid}\left(h(X_i)\right) \tag{4.3}$$

Where i indexes the spatial location. Therefore, we parameterize Bernoulli distributions

as:

$$P(b_i|X) = \begin{cases} p_i, & b_i = 1, \\ 1 - p_i, & b_i = 0 \end{cases} \tag{4.4}$$

Therefore, the initial background probability map $\tilde{B}$ is denoted as the expectation of our Bernoulli distribution:

$$\tilde{B} = P(b_i = 1|X) = \text{Sigmoid}\left(h(X_i)\right) \tag{4.5}$$

That is $h(X)$ activated by the sigmoid function. The binary classifier $h(\cdot)$ can be a simple block consisting of a convolution layer followed by a normalization layer.

However, the missing spatial information in image-level labels leads to the derived $\tilde{B}$ suffering from blurred details, holes, and inaccurate boundaries. We need the geometrical regularization to produce compact and uniform background maps B.

We introduce the total variation (TV) [56] to refine the initial background maps. It removes value variations inside and preserves boundaries, making segments more compact and complete. We defined the geometrical prior as the Gibbs Distribution:

$$P(B|\tilde{B}) = \frac{1}{Z_g} e^{-E(B,\tilde{B})} \tag{4.6}$$

where Z_g is the normalization term and $E(B, \tilde{B})$ is the total variance:

$$E(B, \tilde{B}) = \frac{\lambda}{2}\|B - \tilde{B}\|_2^2 + \|\nabla B\|_2 \tag{4.7}$$

where λ is a scale factor, ∇ is a gradient operator and $\|\cdot\|_2$ is L_2-norm. According to the heat equation [194], it can be solved as an iterative time-dependent surface:

$$-\frac{\partial B}{\partial t} = \nabla\left(\frac{\nabla B}{\|\nabla B\|_2}\right) + \lambda(\tilde{B} - B) \tag{4.8}$$

Therefore, we can obtain rectified background maps B after t iterations:

$$B_{t+1} = B_t - \nabla t \cdot \left[\nabla\left(\frac{\nabla B_t}{\|\nabla B_t\|_2}\right) + \lambda(\tilde{B} - B_t)\right] \tag{4.9}$$

Conditional Categorical Distribution

Conditioned on the refined background regions B, we use image-level labels to compute soft foreground assignment. Assume $Q \in \mathbb{R}^{K \times C}$ is an embedding space holding the semantic information for K foreground classes. Because we only infer label assignments within foreground regions, i.e., $b_i = 0$, we can formulate the conditional foreground classification as:

$$\begin{aligned} P(F|B,X) &= \frac{1}{Z_f} \prod_{i=1}^{H\times W} (1-p_i) \cdot \prod_{k=1}^{K} \left(e^{X_i^{\mathrm{T}} \cdot Q_k}\right)^{f_{i,k}} \\ &= \frac{1}{Z_f} \prod_{i=1}^{H\times W} \prod_{k=1}^{K} \left(e^{\log(1-p_i) + X_i^{\mathrm{T}} \cdot Q_k}\right)^{f_{i,k}} \end{aligned} \tag{4.10}$$

where Z_f is the normalization term, T is a transpose operator. Since $\sum_k f_{i,k} = 1$, we can move $1 - p_i$ inside.

Notably, if we denote $\log(1 - p_i)$ as $\mathcal{M} \in (-\infty, 0)$, then the logarithm of the conditional probability is a generalized form of masked cross-attention [105]:

$$\log P(F|B,X) = \sum_{i=1}^{H\times W} \sum_{k=1}^{K} \mathrm{Softmax}(\mathcal{M}_i + X_i^{\mathrm{T}} \cdot Q_k) \cdot f_{i,k} \tag{4.11}$$

We approximate the foreground probability map F as the expectation of the conditional Categorical distribution and solve it as a process of soft k-means clustering. In order to adapt the iterative process into feedforward neural networks, we unfold the soft k-means clustering as follows:

$$Q_k^t = \mathrm{Softmax}(\mathcal{M} + X^{\mathrm{T}} \cdot Q_k^{t-1}) \cdot X + Q_k^{t-1} \tag{4.12}$$

Therefore, we linked masked cross-attention with soft k-means clustering to efficiently extract semantic information for foreground classes. Consequently, we define an embedding space where each Q_k^0 is a global prototype for the k-th foreground class over the entire dataset. Moreover, Q_k^t is a local prototype that is transformed t times based on the input data. To align with transformers, we extend Eq.(4.12) to a more generalized form, which allows soft k-mean clustering to work in a feedforward manner:

$$Q_t^k = \mathrm{Softmax}\left(\mathcal{M} + \varphi(X)^{\mathrm{T}} \cdot Q_{t-1}^k\right) \cdot \psi(X) + Q_{t-1}^k \tag{4.13}$$

where $\varphi(\cdot)$ and $\psi(\cdot)$ are linear projections respectively. In addition, we define our conditional foreground label assignment as the normalized similarities between pixels and local foreground prototypes:

$$F = \text{Softmax}\left(\mathcal{M} + \varphi(X)^{\mathrm{T}} \cdot Q_t^k\right) \tag{4.14}$$

Therefore, we named the iterative solution to the conditional Categorical distribution a stack of **Contextual Cross-Attention** modules that compute conditional foreground label assignments via a global-to-local prototype transformation. Thanks to smaller variations in local prototypes, our method can improve the quality of F.

4.3.2 Implementation

In this section, we elaborate on the details to implement three key components in the CCA model: (1) how to exploit the unsupervised contrastive learning to infer the missing background information. (2) how to integrate geometrical terms to enhance the loss function in WSSS. (3) how to convert our probabilistic framework into the well-established cross-attention mechanisms.

Unsupervised Contrastive Learning

It is necessary to train the binary background classification module to predict initial background maps. However, image-level labels provide no background information. Therefore, we adopt unsupervised contrastive learning to mine extra information across images within a dataset. The core idea is that each representation should have intra-class similarity and inter-class difference. Denote $\vec{v}_b$ and $\vec{v}_f$ as the feature representation of foreground and background respectively:

$$\vec{v}_b = \frac{\tilde{B} \otimes X^T}{\|\tilde{B}\|_1}, \qquad \vec{v}_f = \frac{(1-\tilde{B}) \otimes X^{\mathrm{T}}}{\|1-\tilde{B}\|_1} \tag{4.15}$$

Here, $\tilde{B} \in \mathbb{R}^{1\times HW}$ and $X \in \mathbb{R}^{C\times HW}$ are flattened. Operators T and $\otimes$ indicate the transpose and matrix multiplication. In addition $\|\cdot\|_1$ denote the L_1 norm. Therefore, $\vec{v}_b$ and $\vec{v}_f$ are mean vectors of each region.

In contrastive learning, we denote the pair of feature vectors from the same class (e.g., background-background or foreground-foreground combination) as the positive pair and vice versa. The core idea behind this is a fact of feature space. The feature representations across images from the same class should be similar to one another

but significantly different from those of other classes. Therefore, we can train a binary classification by pulling close a positive pair and pushing apart the negative pair[87]:

$$s_{i,j}^{neg} = \text{cosine}(v_i^b, v_j^f), \qquad s_{i,j}^{pos} = \text{cosine}(v_i^k, v_j^k) \tag{4.16}$$

where cosine$(\cdot)$ is the cosine similarity operator that outputs a real value between 0 and 1. Besides, $k \in \{b, f\}$ is either b or f. Finally, i and j are the indices of the image within a mini-batch. Consequently, we can explicitly model the background as a set of feature vectors that do not belong to any foreground class with contrastive learning loss:

$$\begin{aligned} L_{CST} = &- \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \log(1 - s_{i,j}^{neg}) \\ &- \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{1}_{[i \neq j]} \log(s_{i,j}^{pos}) \end{aligned} \tag{4.17}$$

where $\mathbf{1}_{[i \neq j]} \in \{0, 1\}$ is an indicator function and it equals 1 if $i \neq j$. In summary, L_{CST} considers the contrastive comparison both within image $(i = j)$ and cross-image $(i \neq j)$.

Geometrical Prior

We use TV regularization to refine the binary classification maps $\tilde{B}$ generated by contrastive learning, since background regions often exhibit holes or ragged boundaries. According to the implementation detail of level-set methods [270], we define an auxiliary surface $\Phi \in \mathbb{R}^{1 \times H \times W}$, and then the smooth surface $B = \frac{1}{1+e^{-\Phi}}$. In this chapter, we set $\nabla t = 1$ for simplicity, and implement $\nabla\left(\frac{\nabla B}{\|\nabla B\|}\right)$ as Gaussian curvature κ [365]. Therefore, the iterative update rule in Eq.(4.9) has become:

$$B_{t+1} = (1 + \lambda) \cdot B_t - \kappa_t - \lambda \cdot \tilde{B} \tag{4.18}$$

We can adjust the smoothness of the background map while maintaining data fidelity with the scale factor λ. We tried different values for λ and fixed it as $\lambda = 2$. After T iterations, we can obtain smooth background regions. In all experiments, we set $T = 3$. Notably, we chose not to include the geometrical prior in the neural network's objective because adding Eq.(4.7) as an area loss component leads to unstable training. Consequently, we derived Eq.(4.18) and used it iteratively. For correct backward

propagation, the TV regularization module is detached from other modules, and the rectified prediction is directly added to the original binary prediction for regularization.

Contextual Cross-Attention

To align with transformers, we integrate each iteration of contextual cross-attention with a self-attention and a Feed Forward Network (FFN) into a decoder layer [105]. To keep the iterative update rules in the k-means clustering algorithm, we set $T = 6$ decoder layers to capture global semantic information from multi-scale features and gradually transform them into image-specific local prototypes to capture non-discriminative foreground regions.

We use the cross-entropy function to compute classification loss between image label Y and image-specific local prototypes Q_T:

$$L_{CLS} = \text{CrossEntropy}(Y, W \cdot Q_T) \tag{4.19}$$

where W is a linear projection head for classification. Our predicted final semantic segmentation map S is defined as a concatenation of rectified binary background prediction B and conditional foreground label assignment F:

$$S = [B, (1 - B) \cdot F] \tag{4.20}$$

Noted that, $\sum_{k=0}^{K} S_{i,k} = 1$ satisfies the spatial exclusivity.

4.3.3 Loss Function

In this chapter, the loss function is defined as

$$L = L_{CLS} + \alpha \cdot L_{CST} \tag{4.21}$$

where α is a weight factor to adjust the importance of contrastive learning loss. Because the early background maps generated by contrastive learning are imprecise and noisy, this significantly destabilizes the training process of the entire model. Therefore, we first trained the contrastive learning head on the dataset, then used its weights as the pre-trained backbone, and finally jointly trained it with CCA modules. In all experiments, we set the $\alpha = 0.3$.

4.4 Experiments and Results

In this section, we conduct comprehensive experiments and ablation analyses of our CCA model and demonstrate its advantages in both interpretability and efficiency.

4.4.1 Experimental Setup

Datasets. We evaluate the proposed CCA on the PASCAL VOC 2012 [167]. The PASCAL VOC 2012 contains 20 foreground and 1 background categories. It consists of three splits: 1,464 images for training, 1,449 for validation, and 1,456 for testing. Following previous WSSS methods, we use the augmented 10,528 training images present in SBD [228] along with their image-level labels to train the network.

Evaluation Metrics. To evaluate the segmentation performance of all methods, we use the validation and test set of PASCAL VOC 2012. In addition, we use the PASCAL VOC 2012 training set to evaluate background map performance and conduct ablation studies. The mean Intersection-over-Union (mIoU) ratio is the evaluation metric in all experiments.

Training Details. In our experiments, we adopt ResNet50 [248] pre-trained on the ImageNet [138] as the backbone and use cascade CCA as in Mask2former [105]. We train all the reproduced models using the AdamW [445] optimizer with a polynomial learning scheduler. The augmentation strategy is the same as [83], including random flipping, random scaling, and cropping. All models take a batch size of 16 on 2 NVIDIA GeForce RTX 3090 GPUs.

Inference. To give a fair comparison, all models generate foreground cams and background maps, convert them into pseudo-labels, and adopt a dense-CRF method [338] as the post-process method for refinement. We further use these refined pseudo-labels to retrain a semantic segmentation network DeeplabV2 [78].

4.4.2 Results and Analysis

Improvements on background maps. We first evaluate background maps because they are crucial to locating the foreground regions for subsequent foreground prototype modeling. We compare the performance of generated background maps in correctly separating the background from other foreground classes. Table 4.1 compares CCA with other advanced methods on the PASCAL VOC 2012 training set. Among these

compared methods, WeakTr [862] achieves the best results with a mIoU of 58.1%. Our proposed explicit background modeling obtains a better performance of 58.8%. Furthermore, we report the performance with dense-CRF post-processing. Our performance with dense-CRF improves the mIoU to 65.7% and outperforms all other methods. We interpret the performance gain derived from two aspects: (1) contrastive learning can effectively mine supplementary information across different images to separate background-foreground regions; (2) TV regularization can adaptively enforce smooth boundaries of foreground regions at the pixel level. Therefore, the dense-CRF is less confusing when refining the binary background maps. Figure 4.4 visualizes the results of background maps on the PASCAL VOC 2012 train set. Our rectified background maps effectively capture all semantic regions across diverse scenes, including varying object scales, crowded scenes, and multiple categories. Notably, our model can correctly suppress false responses when there is no background and large-scale objects fully occupy the image. In contrast, other methods treat large-scale partial foreground objects as background, since background is assumed always to exist.

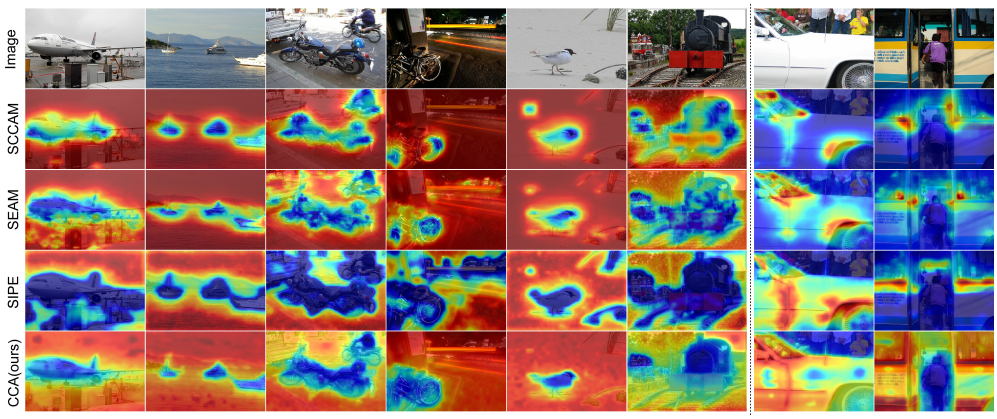


Figure 4.4: Visual comparison of CAMs. From top to bottom: original images, SCCAM, SEAM, SIPE, and the proposed CCA. All columns on the left side of the dotted line demonstrate the CAMs of the background. With geometric regularization, our CCA can effectively suppress false responses from background regions. The two images on the right side have no background, only large-scale foreground objects (i.e., a car and a bus). Other methods mistakenly classify them as background regions, whereas our CCA correctly classifies them as the car and the bus regions.

Improvements on initial pseudo-labels. To show the importance of background modeling, we focus on the performance of initial pseudo-labels generated from background and foreground activation maps. Table 4.2 shows that our method outperforms all other approaches. We can interpret this as the explicit background modeling

Method	Model	Regularization	Extra labels	Location Maps	+dense CRF
$\text{SEAM}_{\text{CVPR20}}$[683]	Thold	-	-	55.4	56.8
$\text{SCCAM}^{*}_{\text{CVPR20}}$[62]	Thold	-	-	50.9	55.3
$\text{EDAM}_{\text{CVPR21}}$[698]	Thold	-	-	52.8	58.2
$\text{AdvCAM}_{\text{CVPR21}}$[356]	Thold	AC	-	55.6	62.1
$\text{ECS}_{\text{CVPR21}}$[600]	Thold	-	-	56.6	58.6
$\text{OCCSE}^{*}_{\text{ICCV21}}$[342]	-	-	-	56.0	62.8
$\text{AEFT}_{\text{ECCV22}}$[768]	-	-	-	56.9	60.3
$\text{CLIMS}_{\text{CVPR22}}$[713]	-	AR	$\mathcal{S}$	57.1	64.2
$\text{SIPE}^{*}_{\text{CVPR22}}$[83]	Adapt	-	-	56.8	63.7
$\text{MCTformer}^{*}_{\text{CVPR22}}$[734]	Thold	-	-	57.2	63.3
$\text{WeakTr}^{*}_{\text{arXiv23}}$[862]	Thold	-	-	58.1	64.5
CCA(Ours)	Const	TV	-	**58.8**	**65.7**

Table 4.1: Evaluation of background maps on the VOC 2012 train set. The symbol * represents all reproduced methods. *Thold* - threshold. *Const* - contrastive learning. *Adapt* - adaption of the foreground. AC - adversarial climbing. TV - total variation. AR - area regularization. $\mathcal{S}$ - extra class-related background texts. '-' means not use. The best results are shown in bold.

Method	Backbone	Supervision	Val dataset	+dense CRF
$\text{SEAM}_{\text{CVPR20}}$[683]	ResNet38	$\mathcal{I}$	49.8	53.3
$\text{SCCAM}^{*}_{\text{CVPR20}}$[62]	ResNet101	$\mathcal{I}$	50.9	54.2
$\text{EDAM}_{\text{CVPR21}}$[698]	ResNet101	$\mathcal{I}$	51.5	54.4
$\text{AdvCAM}_{\text{CVPR21}}$[356]	ResNet101	$\mathcal{I}$	52.3	54.9
$\text{ECS}_{\text{CVPR21}}$[600]	ResNet38	$\mathcal{I}$	53.9	57.1
$\text{OCCSE}^{*}_{\text{ICCV21}}$[342]	ResNet38	$\mathcal{I}$	52.9	56.3
$\text{AEFT}_{\text{ECCV22}}$[768]	ResNet38	$\mathcal{I}$	54.9	57.3
$\text{CLIMS}_{\text{CVPR22}}$[713]	ResNet101	$\mathcal{I}+\mathcal{S}$	56.8	62.0
$\text{SIPE}^{*}_{\text{CVPR22}}$[83]	ResNet38	$\mathcal{I}$	56.3	61.1
$\text{MCTformer}^{*}_{\text{CVPR22}}$[734]	ResNet38	$\mathcal{I}$	55.3	60.7
$\text{WeakTr}^{*}_{\text{arXiv23}}$[862]	DeiT-S	$\mathcal{I}$	57.8	62.2
CCA(Ours)	ResNet50	$\mathcal{I}$	**59.2**	**63.4**

Table 4.2: Comparison of initial pseudo-labels on the VOC 2012 val set. The symbol * represents all reproduced methods. The best results are shown in bold. $\mathcal{I}$ - image-level label $\mathcal{S}$ - extra class-related background texts.

effectively removes irrelevant regions from the foreground and enforces a global-to-local prototype transformation within well-separated foreground regions. Therefore, the embedding feature space can capture more abstract information for global representations of foreground classes while generating more locally adaptive details for non-discriminative regions in each image.

Improvements on fully supervised segmentation. To further evaluate the performance of our methods, we train fully supervised models using refined pseudo-labels and compare the segmentation results on PASCAL VOC 2012 val and test sets. Following the common practice [356], IRN refined the pseudo-labels [3] and used them to train DeepLabV2 [78]. Table 4.3 presents the performance of semantic segmentation, where our CCA outperforms other methods with 74.3% on the validation set and 74.1% on the test set. Based on our CCA, DeepLabV2 shows greater robustness across various challenging scenarios, including different object scales, multiple objects, and multiple categories. In addition, Figure 4.8 shows the semantic segmentation predictions from our model trained with pseudo-labels.

Method	Backbone	Supervision	Val dataset	Test dataset
$\text{SEAM}_{\text{CVPR20}}$[683]	ResNet38	$\mathcal{I}$	64.5	65.7
$\text{SCCAM}^{*}_{\text{CVPR20}}$[62]	ResNet101	$\mathcal{I}$	66.1	65.9
$\text{EDAM}_{\text{CVPR21}}$[698]	ResNet101	$\mathcal{I}$	70.9	70.6
$\text{AdvCAM}_{\text{CVPR21}}$[356]	ResNet101	$\mathcal{I}$	68.1	68.0
$\text{ECS}_{\text{CVPR21}}$[600]	ResNet38	$\mathcal{I}$	66.6	67.6
$\text{OCCSE}^{*}_{\text{ICCV21}}$[342]	ResNet38	$\mathcal{I}$	68.4	68.2
$\text{AEFT}_{\text{ECCV22}}$[768]	ResNet38	$\mathcal{I}$	70.9	71.7
$\text{CLIMS}_{\text{CVPR22}}$[713]	ResNet101	$\mathcal{I}+\mathcal{S}$	70.4	70.0
$\text{SIPE}^{*}_{\text{CVPR22}}$[83]	ResNet38	$\mathcal{I}$	68.5	69.0
$\text{MCTformer}^{*}_{\text{CVPR22}}$[734]	ResNet38	$\mathcal{I}$	70.3	70.8
$\text{WeakTr}^{*}_{\text{arXiv23}}$[862]	DeiT-S	$\mathcal{I}$	72.2	72.3
CCA(Ours)	ResNet50	$\mathcal{I}$	**74.3**	**74.1**

Table 4.3: Comparison of WSSS on the VOC 2012 val and test sets. The symbol * represents all reproduced methods. $\mathcal{I}$ - image-level label. $\mathcal{S}$ - extra class-related background texts. The best results are shown in bold.

4.4.3 Ablation Study

Effects of explicit background modeling. We conduct an ablation study to verify the effect of explicit background modeling. We select 8 options and demonstrate their mIoUs in Table 4.4. Each option concerns a single background modeling method when using the TV term. We observe that the fixed-threshold method achieves the worst mIoU because no single threshold correctly separates background regions across all images. Likewise, estimating background directly from CAMs [83] is also difficult due to heavy reliance on the quality of foreground classifications. In addition, combining background information from a series of binary classifications (used in adversarial

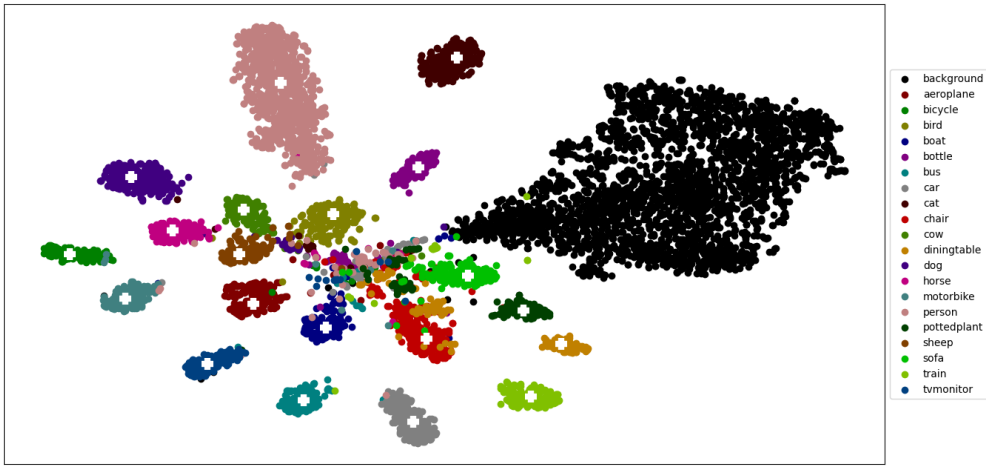


Figure 4.5: Visualization of 21 inter-class groups using t-SNE. Our CCA can capture abstract information to conceptualize each foreground class as a cluster in the feature space.

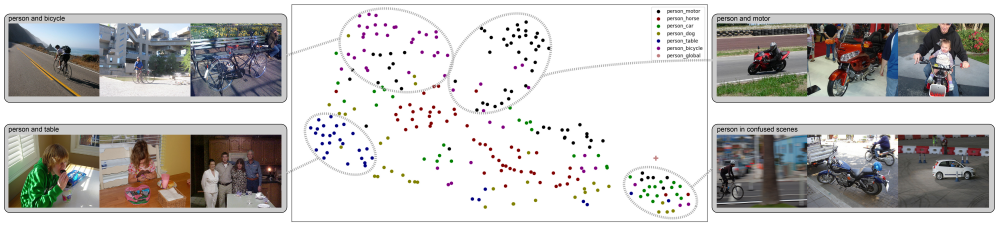


Figure 4.6: Visualization of intra-class deviation in the Person class using t-SNE. Grey dotted circles denote sub-groups. It illustrates the importance of global-to-local prototype transformation in the feature space. Our CCA can transform the abstract global prototype (red plus symbol) into image-specific local prototypes (broadly scattering dots) to capture non-discriminative foreground regions.

learning methods [342]) can achieve performance comparable to that of the contrastive learning method. However, they still face the challenge of spatial exclusivity across different classes, a fundamental requirement for semantic segmentation. Therefore, the presence of multiple responses in binary classifications ultimately degrades the mIoU scores. Our explicit background modeling, with or without the TV term, achieves top-2 scores. We can interpret this superiority as correctly recognizing background regions under different scenarios based on auxiliary cues mined from different images.

Effects of foreground prototype modeling. To understand how our CCA improves the foreground representation quality, we analyze the properties of inter-class separation and intra-class adaptation for our generated feature space via the t-SNE [638] technique.

	Fixed Threshold	Derived from Foreground	Adversary Learning	Contrastive Learning
w/o TV	51.8	52.1	54.3	55.6
w/ TV	55.3	56.4	58.2	59.2

Table 4.4: Ablation study of background modeling options. Each item represents one background modeling method with or without TV. The top 2 best results are in blue and red respectively.

Figure 4.5 visualizes the distribution of all local feature samples, where solid circles denote all local representations and white plus symbols denote global foreground representations. Although a small proportion of confusing samples exists at the center, our generated feature space can still group most samples into their correct categories, demonstrating the ability to distinguish between classes. The visualization supports three important observations: (1) the background class is hard to model within a definite feature space due to the large variation; (2) contrastive learning can extract crucial information to represent image-specific background representations; (3) intra-class deviation causes incomplete coverages using the global prototypes (i.e., large variations of local samples to the white plus symbols). Furthermore, we analyze the global-to-local transformation within the foreground class, Person, that interacts most with other classes. Figure 4.6. visualizes the distributions of local feature samples co-occurring with various classes (e.g., motor, bicycle, car). CCA can adapt a single global prototype (i.e., the plus symbol) into local prototypes to address intra-class deviation. Additionally, we observe that the transformed local prototypes can capture crucial image-specific low-level features and can be partitioned into sub-groups (grey dotted circles). Therefore, our CCA can capture abstract information to conceptualize each foreground semantic group, while transforming these representations into local details to adapt to small intra-class variations across images. Therefore, it can generate more precise partitions to cover non-discriminative foreground regions.

Effects of geometrical regularization. We perform ablation experiments on the PASCAL VOC2012 train set to further validate the effectiveness of geometrical regularization. Figure 4.7. shows that background maps from contrastive learning may have inaccurate separations. We observe that the geometrical regularization term TV can efficiently constrain the size of activated regions and generate smooth boundaries separating background and foreground regions. The first row shows that the TV prior can suppress falsely activated regions and contract the boundaries tightly along the foreground. In addition, the second row demonstrates that TV can expand background regions to cover the confusing parts. Therefore, it can exclude background

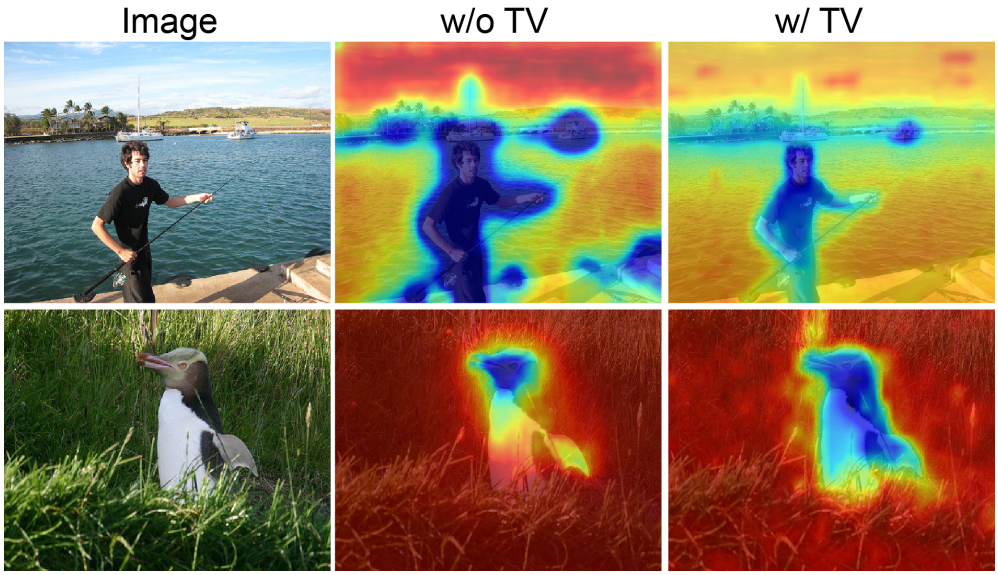


Figure 4.7: Ablation study on the effectiveness of TV term (see in Eq.4.17). TV term can suppress false response background regions (row 1) while expanding confusing regions (row 2). The original images are from the PASCAL VOC 2012 data.

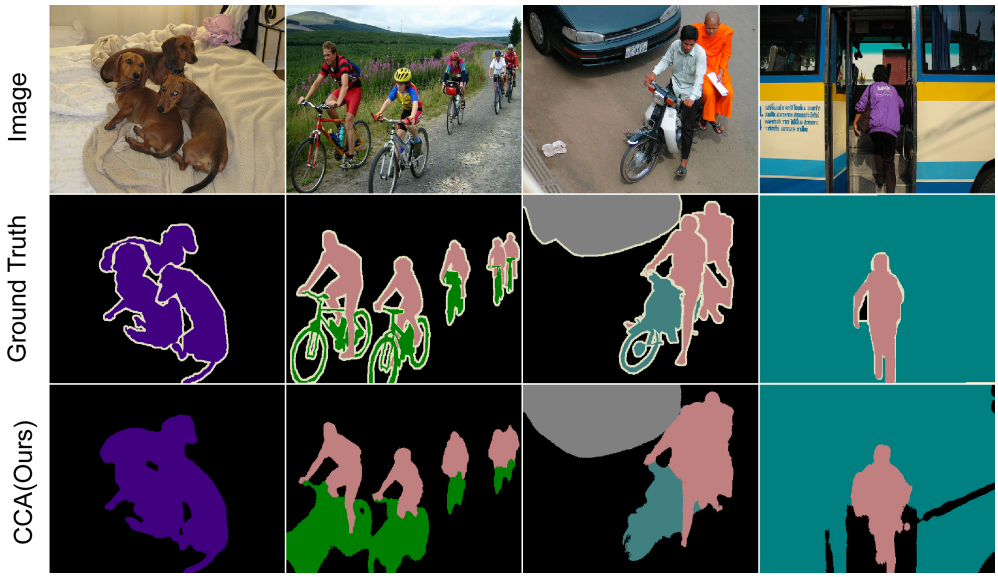


Figure 4.8: Qualitative visualization of semantic segmentation. The original images are from the PASCAL VOC 2012 data.

regions that are in close contact with the foregrounds. Table 4.4 shows quantitative

comparisons between different combinations. As shown, our geometrical regularization term consistently improves performance by suppressing falsely activated regions and expanding non-discriminative background regions.

4.5 Conclusion

This chapter presents a novel conditional probabilistic model with geometric regularization for WSSS. Our CCA approach provides a new understanding of representation learning within the WSSS paradigm. It separates semantic segmentation into two steps: first, extracting background regions, and then classifying foreground classes outside the background. We use Bernoulli distributions to model the background explicitly and train them with contrastive self-supervised learning. Since the initial background maps might be imprecise, we adopt a variational TV term to regulate their boundaries. Subsequently, we define conditional Categorical distributions within identified foreground regions to obtain conditional foreground classification and implement them as Contextual Cross-Attention to capture long-range dependencies. Comprehensive experiments demonstrate the superiority of our CCA over other WSSS methods. Future directions include exploring advanced geometric regularization techniques and extending our framework to fully-supervised and semi-supervised segmentation settings.