

# Study of an Expansion Method Based on an Image-Specific Classifier and Multi-Features for Weakly Supervised Semantic Segmentation

Zhengyang Lyu<sup>a</sup>, Pierre Beausery<sup>b</sup> and Alexandre Baussard<sup>c</sup>  
*Université de Technologie de Troyes, LIST3N, 12 Rue Marie Curie, 10300 Troyes, France*

**Keywords:** Semantic Segmentation, Weak Supervision, Convolutional Neural Network, Support Vector Machine.

**Abstract:** In this paper, we propose a study of an expansion method based on an image-specific classifier and multi-features for Weakly Supervised Semantic Segmentation (WSSS) with only image-level labels. Recent WSSS methods focus mainly on enhancing the pseudo masks to improve the segmentation performance by obtaining improved Class Activation Maps (CAM) or by applying post-process methods that combine expansion and refinement. Most of these methods either lack of consideration for the balance between resolution and semantics in the used features, or are carried out globally for the whole data set, without taking into account potential additional improvements based on the specific content of the image. Previously, we proposed an image-specific expansion method using multi-features to alleviate these limitations. This new study aims firstly at determining the upper performance limit of the proposed method using the ground truth masks, and secondly at analysing this performance limit in relation with the features chosen. Experiments show that our expansion method can achieve promising results, when used with the ground truth (upper performance) and the features that strike a balance between semantics and resolution.

## 1 INTRODUCTION


Semantic segmentation is a popular task in computer vision, with wide applications in various fields such as autonomous driving, medical imaging, or remote sensing imaging. However, training a Fully Supervised Semantic Segmentation (FSSS) model requires laborious pixel-level annotations.


Weakly Supervised Semantic Segmentation (WSSS) has been proposed to reduce the annotation burden. The weak supervision can be based on points (Amy et al., 2016), scribbles (Vernaza and Chandraker, 2017), bounding-boxes (Dai et al., 2015) or image-level labels (Chen et al., 2022). The latter, which is considered in this paper, is the most prevalent in research since it is the easiest and cheapest annotations to obtain.


Figure 1 illustrates a comparison between the pipelines of FSSS and WSSS with only image-level labels. The goal of both tasks is to get a segmentation model capable of making pixel-level prediction for a given image. Generally, segmentation models based on Convolutional Neural Network (CNN)

(Chen et al., 2017; Chen et al., 2018) or more recently, transformers (Strudel et al., 2021) are commonly used. In contrast to FSSS, where the ground truth is available during training, WSSS methods must first generate pseudo masks, which are used as ground truth during the second step to train the segmentation model. To generate the pseudo masks, we start by training a classification model with image-level labels and then, generating Class Activation Maps (CAM) by processing the class-wise deep features from the trained network for each image in the training set (Zhou et al., 2016). Next, an expansion method, which includes interpolation and argmax operations, is used to get seed. Finally a refinement process (Krähenbühl and Koltun, 2011; Ahn et al., 2019) allows to provide an improved pseudo mask, by recovering details using characteristics given by image color features. Of course in WSSS approaches the quality of the pseudo masks directly influences the accuracy of the segmentation results, since they are used to train, in the second step, a fully supervised segmentation model. That is why, recent WSSS methods focus on the generation of pseudo masks whose quality approaches that of the ground truth.

Figure 2 shows the mean Intersection-over-Union (mIoU), obtained by several WSSS methods: (Zhang

<sup>a</sup>  <https://orcid.org/0009-0001-8838-4170>

<sup>b</sup>  <https://orcid.org/0000-0002-2883-1303>

<sup>c</sup>  <https://orcid.org/0000-0002-6693-4282>

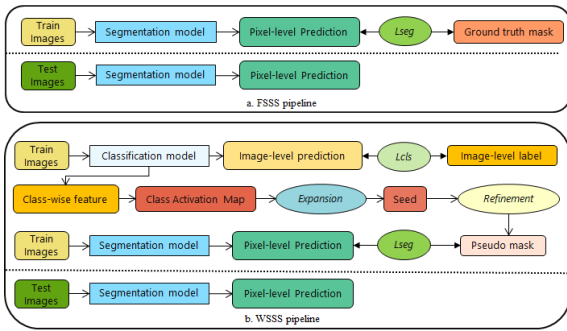


Figure 1: Pipelines of fully supervised semantic segmentation (FSSS) and weakly supervised semantic segmentation (WSSS) with image-level labels.  $L_{seg}$  is the segmentation loss function.  $L_{cls}$  is the classification loss function.

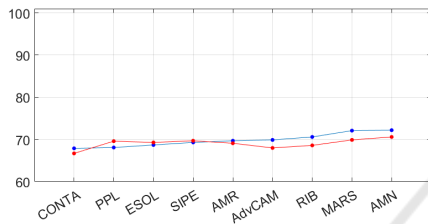


Figure 2: mIoU (%) of the pseudo mask on train set (blue line), mIoU of segmentation (red line), trained from the pseudo masks, on test set of PASCAL VOC 2012 for several WSSS methods.

et al., 2020; Li et al., 2022b; Li et al., 2022a; Chen et al., 2022; Qin et al., 2022; Lee et al., 2022a; Lee et al., 2021a; Lee et al., 2021b; Jo et al., 2023; Lee et al., 2022b), for the pseudo masks (blue line) and the final segmentation (red line). All the considered methods use a classic expansion process and common used refinements processes, such as dCRF (Krähenbühl and Koltun, 2011) and IRNet (Ahn et al., 2019), to generate the pseudo masks. DeepLabV2 (Chen et al., 2017), with ResNet101 backbone, is used as the fully supervised segmentation model. The quality of the pseudo-mask is still far from perfect, and segmentation performance is expected to be better.

Figure 3 provides some visual comparisons between the outputs of the segmentation models trained using ground truth in the FSSS task, or the pseudo mask in the WSSS task. The model trained using pseudo masks tends to generate predictions with imprecise and blurred boundaries even if those pseudo masks seem relatively accurate. However, this effect is mitigated with full supervision (as can be seen line 3 in figure 3). This suggests that the features learned by the WSSS model are adversely affected by the labelling errors presented in the pseudo-masks. When pseudo masks are deduced from a global model, the labelling errors observed on different pseudo masks can reinforce each other or have common causes

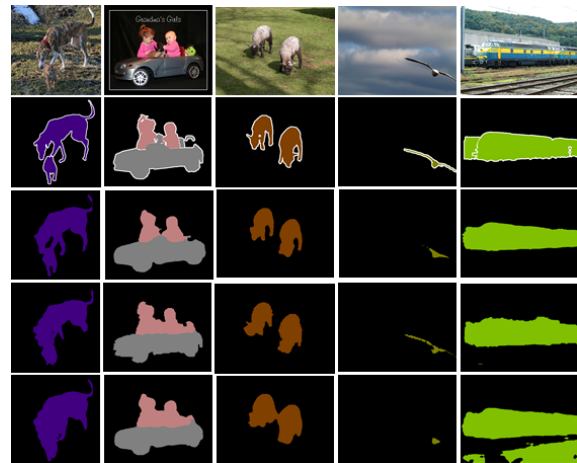


Figure 3: Visual comparison between the the output of the segmentation models trained by ground truth and pseudo mask. From top to bottom: image, ground truth, output of the fully supervised segmentation network, pseudo mask, output of segmentation network trained with pseudo mask.

and contribute to degrading the overall quality of the pseudo masks used to create the segmentation network training set (i.e., rail tracks with train). To reduce this effect, it is necessary to make better use of the individual properties of each image and to improve the pseudo masks.

According to those observations, we proposed in a previous work an image-specific expansion method using multi-features to alleviate the limitations of expansion methods which ignore the balance of resolution and semantics in the features used or miss consideration of image’s specificity. The detailed proposed pipeline is shown in Figure 5. a. We do not give too specific introduction for this method in this paper. Just in brief, with only image-level labels, we designed a sample selection strategy by using multi-features: CAM, seed and shallow features, to select data from high-resolution shallow features with sufficient semantics, and label them by the value in the seed, which enables to create a pixel-wise data set for training an image-specific Support Vector Machine (SVM) classifier to infer the pixel-wise prediction for the entire image. Using this expansion method, we can get better predictions compared with the original seed. With further refinement process, our enhanced pseudo masks are also promising. Some predictions examples from the SVM are shown in the last column in Figure 4.

The critical part of the proposed expansion method is the sample selection process. It directly influences the accuracy of the labels in the pixel-wise train set and thus the quality of the prediction. In order to evaluate the full potential of our expansion

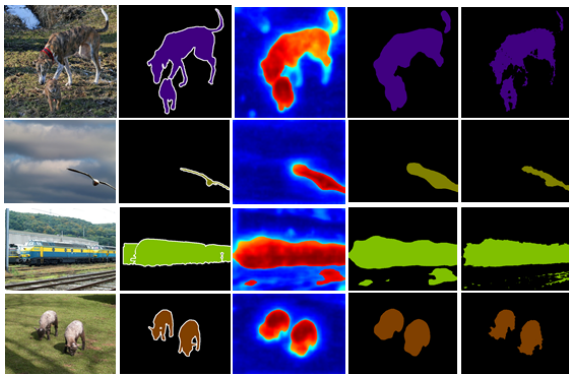


Figure 4: Qualitative results from the proposed expansion method in weakly supervised. From the left to right: image, ground truth, CAM, seed and the result from the SVM in our expansion method.

method, we conduct, in this study, experiments assuming that the ground truth is available, which enables us to define a pixel-wise training set free of labelling error. This training set can be regarded as the best result for the sample selection process. Thus, sample selection step of the original method described in Figure 5.a is replaced by a uniform random sampling from features of the classification network that are labelled using ground truth, as shown in Figure 5.b.

For the feature, we explore various options by choosing activation map values from different layers in the classification network. Next, for each image on the training set, a SVM is trained to label its pixels. We show that, when ground truth is available, the prediction results from SVM is particularly promising under the condition that the features used strike a balance between resolution and semantics.

The rest of the paper is organized as follows: Section 2 presents related work about the expansion process in the WSSS framework. Section 3 describes the proposed expansion method when assuming ground truth is available. Section 4 provides the experimental setup and substantial results. We conclude and outline future research directions in Section 5.

## 2 RELATED WORK

Generally, weakly supervised semantic methods with only image-level labels requires a 2-step pipeline: pseudo mask generation and segmentation model training. Since the quality of the pseudo masks directly impacts the performance of the segmentation model, most methods put main efforts on improving the accuracy of those pseudo masks. These methods can be divided into 2 groups. The first one tries to ad-

just classification models to obtain improved CAM, by strategies likes improving training mechanisms (Wang et al., 2020) and contrastive learning (Yuan et al., 2023). The main challenge in these methods is finding an effective connection between the implemented modifications and the resulting enhancement of CAMs.

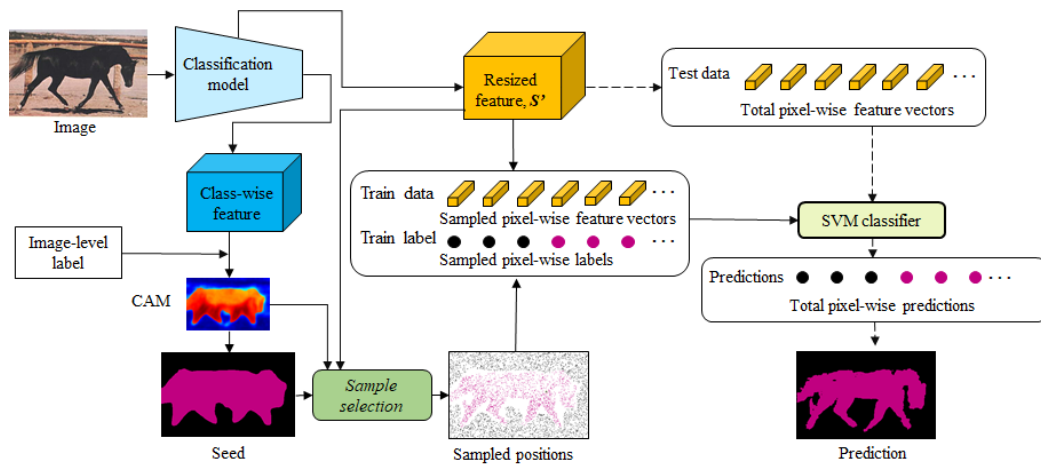
The second one aims at designing better expansion methods and using refinement processes to obtain high quality pseudo masks from CAM (Krähenbühl and Koltun, 2011; Ahn and Kwak, 2018; Ahn et al., 2019; Li et al., 2021; Jo et al., 2023). For example, PMM (Li et al., 2021) strives to overcome the partial response problem in CAM by using a smoothing method to expand the localization area, and generate class-specific background for each image to independently obtain the pixel-wise label when obtaining the pseudo mask. In the same way, by generating class-wise centroids prototype from unsupervised features among the whole dataset, the MARS method (Jo et al., 2023) is proposed to exclude false activation made by co-occurrences between the background elements and associated objects in CAM.

Different with fuzzy localization map provided by CAM, details are comparatively recovered after the well-designed expansion method with refinement process. We observe that improved expansion methods can be achieved by using high resolution shallow features, such as the color information of the original image, and relying on the specific semantic information for the given image. However, we argue that, color information may not contain sufficient semantics to well represent the class in the image, which may lead to incorrect predictions (Krähenbühl and Koltun, 2011; Li et al., 2021). Besides, valuable image-specific feature may not be effectively utilized in expansion methods which are implemented globally (Ahn and Kwak, 2018; Ahn et al., 2019; Jo et al., 2023).

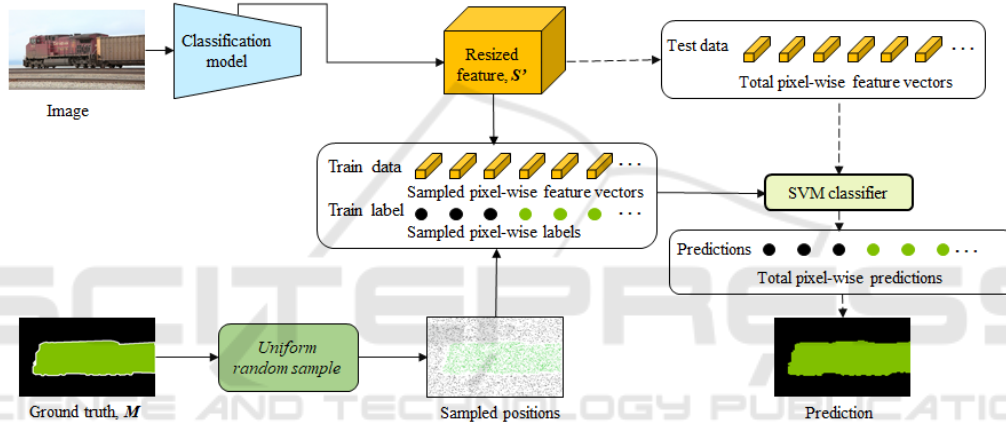
The proposed expansion method in our previous work introduces an image-specific pixel-wise classifier with multi-features to solve the limitations in the recent expansion method. In this paper, we want to study the potential of this expansion method by assuming ground truth is available, and find the optimal options for the feature used by revealing the relevance between performance and resolution-semantics balance.

## 3 METHOD

In this section, we describe the details of proposed expansion method for fully supervised semantic seg-



a. Pipeline of the proposed expansion method for weakly supervised semantic segmentation.



b. Pipeline of the proposed expansion method for fully supervised semantic segmentation.

Figure 5: Detailed pipeline of the proposed expansion method for WSSS (a) and its corresponding fully supervised approach (b). The solid lines are related to training phase of the SVM classifier, and the dot lines are related to the inference phase.

mentation, as shown in Figure 5.b. We implement this method assuming the availability of ground truth, with the goal of assessing its potential upper performance limit for the WSSS task and finding the clues about the feature to use.

The proposed method is divided into 2 steps: Firstly, for each image we construct a training set composed of a data vector, which can be a set of selected pixels of the image itself or from resized features of the classification model (outputted activation maps of a layer in the trained classification network), and the corresponding training labels given by the ground truth. Secondly, we start by training a pixel-wise classifier, here we consider a SVM model, using the training set. Then, we output pixel-wise predictions by inferring all the pixels of the original image or the considered resized features of the classification model (called the test data in what follows). Finally, the pixel-wise predictions are reshaped into a two-

dimensional prediction mask.

### 3.1 Construction of the Training Set and Test Data

The first part in our method is to construct the pixel-wise training set, which includes training data and training labels, and the test data.

The training set is constructed by sampling data from the feature generated in the classification model or directly from the original image, and labeling the data by the ground truth. The details processes are illustrated as followed: firstly, a classification model is trained by the given images with image-level labels. Assuming that  $l$  classes are labelled in the data set, we define  $O = \{1, 2, \dots, l\}$  as the set of classes for image-level labels. The data set is a collection of image  $I \in \mathbb{R}^{W \times H \times 3}$  with image-level labels  $\mathcal{Y}$ , with



$\mathcal{Y} \subset \mathcal{O}$ . Then, we uniform random sample  $n$  training samples for each class in the image, as shown in Figure 6. Each selected sample  $\mathcal{S}'_i$  is characterized by the  $i^{\text{th}}$  pixel-wise feature deduced from the original image or outputted activation maps  $\mathcal{S}$  of a layer in the trained classification network after bilinear interpolation to image size. Finally, the pixel-wise training set is built with the  $(|\mathcal{Y}| + 1)n$  pairs,  $(\mathcal{S}'_i, \mathbf{M}_i)$ , where  $\mathcal{S}'_i$  is the pixel-wise feature vector and  $\mathbf{M}_i$  is the ground truth label in pixel  $i$ .  $|\mathcal{Y}|$  is the number of the object class in the image.

Notice that, in order to avoid training errors caused by imbalanced data, the number of samples is kept the same for each object class and background. It is essential to stress that the accuracy of SVM results is related to the total number of samples. When using a large amount of samples, the results will get better, at the cost of training time. It is also interesting to set the total number of samples linked to the resolution of the feature.

Depending on the layer, the resolution of the chosen feature varied from 1/2 to 1/16 of the original image. The details resolution for the chosen feature can be seen in Section 4.1. Normally, the shallow feature generated from the early layers in the classification network tends to have shallow-semantic information with high-resolution. In contrast, the deep features have high-semantic information with low-resolution. Different features used in generating localization maps can bring varied performance. Low-resolution features are prone to result in smoothed boundaries and missing details (Zhou et al., 2016), while insufficient semantics can lead to noisy and inaccurate predictions (Krähenbühl and Koltun, 2011). Thereby, by comparing the predictions given by the classifiers trained with different features, the appropriate balance between the semantics and resolution for the ideal prediction can be observed.

To construct the so-called test data, we select all the pixels of the original image or the feature  $\mathcal{S}'$ . The test data serves as the input for the trained pixel-wise classifier during the inference stage.

### 3.2 Training and Infer of the Pixel-Wise Classifier

As already mentioned and shown in Figure 5, the pixel-wise classifier, implemented as a Support Vector Machine (SVM) in our study, conducts a training phase using the constructed image-specific pixel-wise training set.

During training, the SVM learns to distinguish between different classes based on the sampled feature data. The one-vs-all strategy is employed in multi-



Figure 6: Labeled samples from the ground truth. From left to the right: image, ground truth, and the selected labeled samples.

class classification task for SVM, where the classifier is trained for each class against the others.

Once the SVM is trained, the pixel-wise features from the test data are inferred to assign class labels. These predictions are then reconstructed into a segmentation mask.

## 4 EXPERIMENTS

In this section, we first give the details for experimental settings like dataset, evaluation metrics and implementation details. Then, we exhibits our experiment results quantitatively and qualitatively.

### 4.1 Experimental Settings

**Dataset and Evaluation Metrics.** Our preliminary study is evaluated on PASCAL VOC 2012 dataset (Everingham et al., 2015), which has 20 foreground

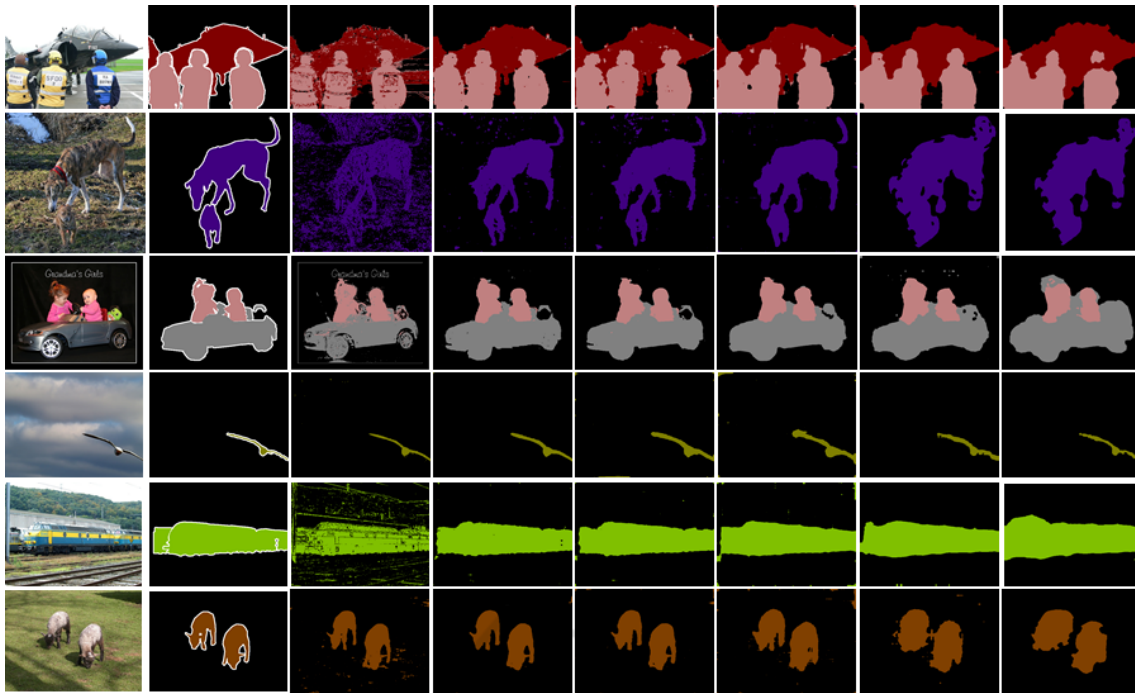


Figure 7: Qualitative segmentation results on PASCAL VOC 2012 train set. From left to right: image, ground truth, prediction when SVM is trained by using the original image as feature, and features generated from the different layers in the backbone classification network, namely  $F_1$  to  $F_5$ .

Table 1: Comparison of segmentation mIoU (%) scores using different features on the PASCAL VOC 2012 train set. The best is highlight by bold.

Feature	bkg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mean
img	86.98	68.79	35.91	64.40	57.83	58.97	68.77	62.19	77.31	58.65	73.76	69.01	71.25	73.22	57.99	68.02	61.07	70.89	74.65	64.80	60.96	65.97
$F_1$	<b>98.98</b>	<b>96.70</b>	<b>89.97</b>	<b>96.95</b>	<b>96.39</b>	95.61	<b>96.80</b>	<b>96.74</b>	<b>98.32</b>	<b>95.76</b>	<b>97.77</b>	97.74	<b>97.82</b>	<b>97.75</b>	<b>95.22</b>	<b>96.87</b>	<b>96.17</b>	<b>97.63</b>	97.82	<b>96.70</b>	96.13	<b>96.67</b>
$F_2$	98.65	92.87	85.61	95.38	93.90	95.34	95.66	95.63	98.16	94.89	96.77	96.76	97.23	96.86	93.93	95.85	95.02	97.22	97.47	95.80	95.19	95.44
$F_3$	98.58	87.62	84.83	93.02	91.31	<b>96.36</b>	96.74	95.63	97.91	95.57	95.73	<b>97.82</b>	96.44	96.31	93.88	96.27	95.31	96.53	<b>98.29</b>	96.52	<b>97.50</b>	95.15
$F_4$	96.98	78.48	75.53	86.10	84.00	93.96	94.35	91.75	95.39	91.04	87.35	96.74	90.63	88.29	89.19	93.85	91.03	88.22	96.43	93.71	96.69	90.46
$F_5$	90.19	55.38	66.41	61.51	60.54	77.25	80.48	75.81	82.78	80.74	72.03	91.36	72.52	71.02	73.94	85.84	81.31	72.71	82.62	73.93	86.00	75.92

classes and 1 background class. The official dataset is split into train set, validation set and test set, which contains 1464, 1449 and 1456 images, respectively. There is an additional annotations provided by Semantic Boundary Dataset (Hariharan et al., 2011) which augment the train set to 10582 images. The augmented train set is used to train the classification network, which is able to generate image-specific features for the given image. We evaluate the segmentation results on all 1464 images from the train set. Mean-intersection-over-union is used to evaluate segmentation results.

**Implementation Details.** In our experiments, we use the classification network designed in the weakly supervised semantic segmentation framework called Self-supervised Image-specific Prototype Exploration (SIPE) (Chen et al., 2022) to obtain features for the given image. Following the set in SIPE, we use ResNet-50 (He et al., 2016) as the backbone classification network. The architecture of ResNet-50 is di-

vided into 5 stage, each of which consists in the convolutional and pooling layers. We named the output features from each stage in the backbone classification network as  $F_1$  to  $F_5$ , whose spatial size is  $1/2$ ,  $1/4$ ,  $1/8$ ,  $1/16$  and  $1/16$  of the original image’s size, respectively. These features are z-score normalized to prevent certain parts of features from dominating the training process due to their larger magnitude. For each category, including background, in the image, we set  $n = 2500$  to sample, which corresponds to a small fraction of image pixels.

## 4.2 Experiment Results

Figure 7 shows the SVM predictions when using varied features from different stages. We observe that, when training SVM using the original image as feature, high-resolution results with relatively detailed object contours are generated. However, due to the limited semantics there was a higher occurrence of

misclassifications in background regions, leading to a relatively imprecise prediction. Using  $F_5$  feature, which is in low-resolution with richer semantics, the predicted segmentation boundaries is blurred and less accurate. Both of these two features have advantages in terms of resolution or semantics for SVM results, but the imbalance between the two factors leads to imperfect results. In comparison, as shown from column 4 to column 6 in Figure 7, when using shallow features with sufficient semantics, like  $F_1$ ,  $F_2$  and  $F_3$ , most of the wrong predictions in the background are clearly avoided and the objects boundaries are quite clear.

Segmentation results using different features for each class category and the mean performance in the PASCAL VOC 2012 train set are shown in Table 1. Results given by using features which strike the balance between resolution and semantics, i.e.,  $F_1$ ,  $F_2$  and  $F_3$ , shows best segmentation performance in all categories. Using  $F_1$  feature makes the best performance in most categories, which suggests that when ground truth is available, the pixel-wise classifier trained using high-resolution with sufficient semantics features is able to generate clear segmentation boundaries and reasonably accurate segmentation results. Besides, we also did experiments by simply concatenating selected features, i.e., concatenating more than one features along the channel axis. However, the output results from the classifier trained by the concatenated features are close to the results from the classifier trained by the feature which has the most number of channels among the selected features. It reveals that, when using more than one feature, it is important to find an appropriate method to combine them, especially taking into account the differences in feature depths.

The experiments results shows the great potential for our expansion method for a segmentation task. Since we implement the expansion method with ground truth in this paper, the results can be regarded as the upper performance limit of the proposed approach. It reveals that segmentation results in our WSSS task, which corresponds to the one initially targeted, should be improved when the labels of the selected samples are sufficiently precise and the feature used represents a good compromise between resolution and semantics.

## 5 CONCLUSIONS

When we investigated methods to improve WSSS performances, we observed that the main critical part is to get better pseudo masks. Compared with the

ground truth, the pseudo masks are not perfect yet. We start from a proposed expansion method made to improve them by training a pixel-wise SVM classifier for each image in the training set. In this study, we use ground truth to label samples, which is regarded as the best sampling selection process, to evaluate the potential of this expansion method. We found that promising prediction is generated when the used feature keeps balance between semantics and resolution. It shows that our expansion method has some potential and that high-resolution shallow features with sufficient semantics brings effective gain in generating high quality pseudo masks.

Beside these considerations, it also appears that the performance of WSSS segmentation networks are close to the performance obtained with pseudo masks, thus improving pseudo masks is valuable if the segmentation network is able to take advantage of their quality as shown in figure 8. Indeed, FSSS DeeplabV2, used as backbone in all considered WSSS methods, reaches almost 79%. The gain we can expect with this backbone from perfect pseudo masks is then significant but not that large. Our experiments show that by improving the sample selection process, we could reach more than 95% mIoU for pseudo masks which is worth the effort if using the best FSSS network that reaches 90.6% mIoU. So finally, we can conclude that the efforts put to improve pseudo masks should be adapted according to the expected performance of the backbone network in FSSS.

We must also mention that another line of research is to consider segmentation models that self correct the pseudo masks during training so that post-processing could be avoided leading to less demanding methods from computational point of view. Works in that direction are under investigation.

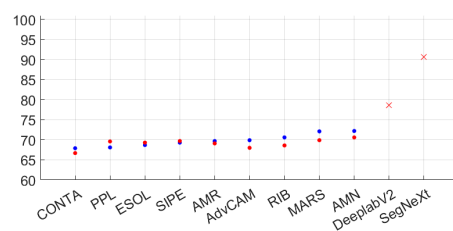


Figure 8: mIoU (%) of the pseudo mask on train set (red points), mIoU of segmentation (model trained from the pseudo masks, blue points) on test set of PASCAL VOC 2012 for several WSSS methods, and mIoU of the segmentation for fully supervised models: DeepLabV2 and SegNeXt (Guo et al., 2022) on train set (red crosses).



## REFERENCES

- Ahn, J., Cho, S., and Kwak, S. (2019). Weakly supervised learning of instance segmentation with inter-pixel relations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.
- Ahn, J. and Kwak, S. (2018). 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 (CVPR)*.
- Amy, B., Olga, R., Vittorio, F., and Li, F. (2016). What's the point: Semantic segmentation with point supervision. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer.
- Chen, L., Zhu, Y., Papandreou, G., F. Schroff, F., and Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*. Springer.
- Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Chen, Q., Yang, L., Lai, J., and Xie, X. (2022). Self-supervised image-specific prototype exploration for weakly supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.
- Dai, J., He, K., and Sun, J. (2015). Boxesup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the International Conference on Computer Vision (ICCV)*. IEEE.
- Guo, M., Lu, C., Hou, Q., Liu, Z., Cheng, M., and Hu, S. (2022). Segnext: Rethinking convolutional attention design for semantic segmentation. *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 35.
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.
- Jo, S., Yu, I., and Kim, K. (2023). Mars: Model-agnostic biased object removal without additional supervision for weakly-supervised semantic segmentation. *arXiv preprint arXiv:2304.09913*.
- Krähenbühl, P. and Koltun, V. (2011). Efficient inference in fully connected crfs with gaussian edge potentials. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc.
- Lee, J., Choi, J., Mok, J., and Yoon, S. (2021a). Reducing information bottleneck for weakly supervised semantic segmentation. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc.
- Lee, J., Kim, E., Mok, J., and Yoon, S. (2022a). Anti-adversarially manipulated attributions for weakly supervised semantic segmentation and object localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Lee, M., Kim, D., and Shim, H. (2022b). Threshold matters in wsss: Manipulating the activation for the robust and accurate segmentation model against thresholds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.
- Lee, S., Lee, M., Lee, J., and Shim, H. (2021b). Railroad is not a train: Saliency as pseudo-pixel supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.
- Li, J., Jie, Z., Wang, X., Wei, X., and Ma, L. (2022a). Expansion and shrinkage of localization for weakly-supervised semantic segmentation. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc.
- Li, J., Jie, Z., Wang, X., Zhou, Y., Wei, X., and Ma, L. (2022b). Weakly supervised semantic segmentation via progressive patch learning. *IEEE Transactions on Multimedia*.
- Li, Y., Kuang, Z., Liu, L., Chen, Y., and Zhang, W. (2021). Pseudo-mask matters in weakly-supervised semantic segmentation. In *Proceedings of the International Conference on Computer Vision (ICCV)*.
- Qin, J., Wu, J., Xiao, X., Li, L., and Wang, X. (2022). Activation modulation and recalibration scheme for weakly supervised semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. MIT Press.
- Strudel, R., R.Garcia, , Laptev, I., and Schmid, C. (2021). Segmenter: Transformer for semantic segmentation. In *Proceedings of the International Conference on Computer Vision (ICCV)*. Springer.
- Vernaza, P. and Chandraker, M. (2017). Learning random-walk label propagation for weakly-supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.
- Wang, Y., Zhang, J., Kan, M., Shan, S., and Chen, X. (2020). Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Yuan, K., SChaefer, G., Lai, Y., Wang, Y., Liu, X., Guan, L., and Fang, H. (2023). A multi-strategy contrastive learning framework for weakly supervised semantic segmentation. *Pattern Recognition*.
- Zhang, D., Zhang, H., Tang, J., Hua, X., and Sun, Q. (2020). Causal intervention for weakly-supervised semantic segmentation. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc.
- Zhou, B., Khosla, A., Lapedriza, A., et al. (2016). Learning deep features for discriminative localization. In *Proceedings of the International Conference on Computer Vision (ICCV)*. Springer.