

High-fidelity Pseudo-labels for Boosting Weakly-Supervised Segmentation

Arvi Jonnarth* Yushan Zhang Michael Felsberg†

Department of Electrical Engineering, Linköping University, Sweden

{arvi.jonnarth, yushan.zhang, michael.felsberg}@liu.se

Abstract

Image-level weakly-supervised semantic segmentation (WSSS) reduces the usually vast data annotation cost by surrogate segmentation masks during training. The typical approach involves training an image classification network using global average pooling (GAP) on convolutional feature maps. This enables the estimation of object locations based on class activation maps (CAMs), which identify the importance of image regions. The CAMs are then used to generate pseudo-labels, in the form of segmentation masks, to supervise a segmentation model in the absence of pixel-level ground truth. Our work is based on two techniques for improving CAMs; importance sampling, which is a substitute for GAP, and the feature similarity loss, which utilizes a heuristic that object contours almost always align with color edges in images. However, both are based on the multinomial posterior with softmax, and implicitly assume that classes are mutually exclusive, which turns out suboptimal in our experiments. Thus, we reformulate both techniques based on binomial posteriors of multiple independent binary problems. This has two benefits; their performance is improved and they become more general, resulting in an add-on method that can boost virtually any WSSS method. This is demonstrated on a wide variety of baselines on the PASCAL VOC dataset, improving the region similarity and contour quality of all implemented state-of-the-art methods. Experiments on the MS COCO dataset further show that our proposed add-on is well-suited for large-scale settings. Our code implementation is available at <https://github.com/arvijj/hfpl>.

1. Introduction

Recent years have witnessed significant advancements in deep learning methods, impacting many computer vision tasks, including semantic segmentation. Automatic image segmentation has proven useful in many applications, including autonomous driving [9], video surveillance [12],

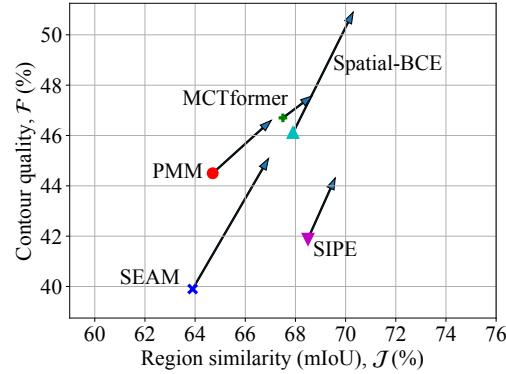


Figure 1. Region similarity vs. contour quality. The arrows represent improvements from our proposed approach, which boosts the performance of five state-of-the-art methods. The results are five-run averages, which required reimplementations of all methods.

and medical image analysis [28]. Although remarkable results have been achieved by fully-supervised segmentation frameworks, the utilization of large datasets with pixel-wise annotated images necessitates a significant manual labeling effort, which increases with the dataset size. To address this challenge, weakly-supervised semantic segmentation (WSSS) aims to alleviate the labelling effort by other alternatives, such as image-level labels [33], bounding boxes [10], scribbles [24], and points [4]. Out of all the available options, image-level classification labels are the most cost-effective and can be obtained from various sources, making them the most popular choice, and addressed in this work.

WSSS with image-level supervision aims to produce pseudo-masks for images from classification labels only. For this purpose, class activation maps (CAMs) [38] are exploited in many methods. While this approach has been shown to have a remarkable ability for localization, it does not take the object shape into account. This leads to activation mostly around discriminative regions and overly smooth CAMs, without well-defined prediction contours. This is due to the nature of global average pooling, where all the pixels in the image are treated as the same category, which is problematic in background areas. In the particular case of SEAM [33], two complementary techniques have been proposed to overcome this drawback [16]: (1) The

*Affiliation: Husqvarna Group, Huskvarna, Sweden.

†Co-affiliation: University of KwaZulu-Natal, Durban, South Africa.

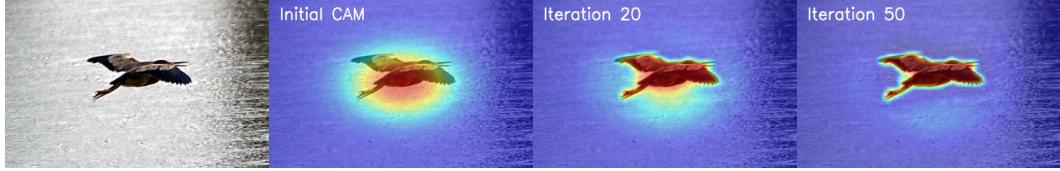


Figure 2. Illustration of how our improved feature similarity loss improves an initial CAM.

importance sampling loss (ISL), which applies the classification loss to pixels sampled from a distribution based on the CAM; (2) the feature similarity loss (FSL), which aligns prediction contours with edges in the image. However, they are based on the multinomial posterior with softmax, and implicitly assume that classes are mutually exclusive, which we find is suboptimal. We aim to address this drawback.

The mutual exclusivity assumption for ISL and FSL is suboptimal for two reasons: (1) It is violated in practice, as the losses are applied to downsampled CAMs, where a single pixel represents a larger image region. Thus, two classes can share the space occupied by a single pixel, especially near class borders. (2) The techniques do not easily generalize beyond the SEAM baseline, as most methods treat the classes independently based on the sigmoid activation. Moreover, some methods do not explicitly model the background, which is necessary for ISL with softmax, as the foreground probability otherwise sums to 1 in every pixel.

To solve both issues at once, we model multiple binary problems, with independent binomial posteriors for each class. We find that, as no assumption is made on class co-occurrences, the performance is improved. Figure 2 highlights our improved FSL, where foreground and background are fully decoupled for an initial coarse CAM. Additionally, as this approach aligns with the majority of WSSS methods, our proposed method generalizes seamlessly, and boosts the performance of all tested state-of-the-art baselines, as shown in Fig. 1. Compared to the approach in the short paper [16], our contributions are three-fold:

- We analyse the weaknesses of the multinomial-based approach in [16] that limit performance and generality, restricting the method to the SEAM baseline.
- We model the independent binomial posteriors as opposed to the multinomial posterior, to formulate an add-on method that generalizes to virtually any baseline.
- In extensive ablations and experiments, we verify the choices in our approach and show that it improves several state-of-the-art methods on the common mIoU and the complementary contour quality, as shown in Fig. 1.

2. Related work

Types of weak supervision. Full supervision of semantic segmentation requires pixel-wise annotations, which are difficult to acquire and can suffer from inaccuracies. Therefore, it is of interest to investigate weak supervision that

requires less manual labeling. Bounding boxes are a natural substitute for pixel-wise labels, and are exploited in different works [10, 17, 22, 29]. Inspired by interactive image segmentation, scribbles [24, 32] and points [5] have been used to ease the annotation. Image classification labels [38], which are the easiest to acquire, are also the most common.

Pseudo-label generation from image classification labels constitutes a core part of WSSS. Zhou *et al.* [38] introduced class activation maps (CAMs) by revisiting the global average pooling (GAP) layer in classification networks, and showed that they had a remarkable localization ability, despite being trained with only classification labels. However, CAMs usually only activate over sparse discriminative regions. Subsequent works aim to overcome this drawback. Kolesnikov *et al.* [18] introduced a global weighted rank pooling to replace GAP, and a constrain-to-boundary loss to learn precise boundaries during training. Wei *et al.* [34] proposed an adversarial erasing approach to mining and expanding object regions continuously. Huang *et al.* [14] utilized the classical seeded region growing method to expand the discriminative regions to cover the entire objects. Lee *et al.* [20] proposed FickleNet using dropout to stochastically select hidden units, allowing a single network to generate localization maps for different object parts. Wang *et al.* [33] proposed a self-supervised equivariant attention mechanism (SEAM) by imposing consistency regularization in a siamese network. Closely related to our work, Jonnarth and Felsberg [16] performed importance sampling as a substitute for GAP, and introduced an auxiliary feature similarity loss as additional supervision for SEAM [33]. The main differences are: (1) We analyse the weaknesses that limit their performance and generality; (2) we model the binomial as opposed to the multinomial posterior; (3) we generalize ISL and FSL beyond SEAM, to virtually any baseline.

Pseudo-label refinement methods enhance the pseudo-masks based on the CAMs. Ahn *et al.* [2] proposed AffinityNet that predicts class-agnostic pixel-level semantic affinity. They further explored inter-pixel relations (IRN) [1], where confident seed areas are identified, and propagated via random walk. Chen *et al.* [6] explored object boundaries explicitly and used the explored boundaries to constrain the localization map propagation. Lee *et al.* [21] introduced an anti-adversarial technique to manipulate images for increasing its classification score. Krähenbühl and Koltun [19] proposed a refinement method based on conditional random fields (CRF), which is commonly used in WSSS.

3. Preliminaries

3.1. The weakly-supervised segmentation task

We consider the image-level weakly-supervised semantic segmentation (WSSS) task. As opposed to the fully supervised case, no pixel-wise annotations are used for training. Given a dataset of images and corresponding image-level labels, we aim to train a model that predicts the semantic class for every pixel in an image. Commonly, this is achieved by first training a multi-label classification network, from which pseudo-labels are generated, commonly based on CAMs [33]. Optionally, a pseudo-label refinement method is applied [2]. Finally, a segmentation model is trained in the fully-supervised setting, where the pseudo-labels are used as ground truth. Our contributions focus on the first stage of generating high-fidelity pseudo-labels.

3.2. Class activation maps

Class activation maps (CAM) of different forms have become a core component in the majority of weakly supervised segmentation methods [7, 8, 33]. Zhou *et al.* [38] introduced the CAM concept, and showed that classification networks that use a global average pooling (GAP) layer possess the ability to localize objects without any supervision on their locations, such as bounding boxes or segmentation masks. In their network architecture, they perform GAP on the feature map $f_k(i, j)$ over the spatial coordinates i and j for each channel k , where $f_k(i, j)$ is the output of the final convolutional layer. They then apply a single fully connected layer with weights w , and thus, the *image-level* score for class c is

$$U_c = \sum_k w_k^c \sum_{i,j} f_k(i, j) = \sum_{i,j} \sum_k w_k^c f_k(i, j). \quad (1)$$

The weight w_k^c can be seen as the importance of feature k on class c . Thus, they formulate the CAM as

$$M_c(i, j) = \sum_k w_k^c f_k(i, j), \quad (2)$$

which is the weighted sum over the features, and directly indicates the importance at spatial grid position (i, j) [38].

Since CAMs provide the ability to localize objects using only classification labels, they are a natural choice for weakly supervised segmentation. However, most WSSS methods modify the original CAM, by removing the fully connected layer, and explicitly represent the output of the final convolutional layer as the spatial class score $s_c(i, j)$, with the same number of channels, C , as there are classes [7, 33, 36]. In this case, the CAM is equivalent to s_c . GAP is then applied to s_c to get the image-level class score

$$S_c = \sum_{i,j} s_c(i, j). \quad (3)$$

Furthermore, in WSSS, the CAMs are usually normalized to the range $[0, 1]$ for pseudo-label generation. A common method involves ReLU and max normalization [7, 33, 37]

$$r_c(i, j) = \frac{\text{ReLU}(s_c(i, j))}{\max_{m,n} \text{ReLU}(s_c(m, n))}, \quad (4)$$

which we refer to as max-normalized CAMs.

A third option for creating CAMs is to model the pixel-wise multinomial posterior, prior to global pooling [16]. The normalized CAMs are computed using softmax

$$a_c(i, j) = \frac{e^{s_c(i, j)}}{\sum_{k=1}^C e^{s_k(i, j)}} \cong \Pr(z_c(i, j) = 1|x), \quad (5)$$

given image x , where C is the number of classes, $z_c(i, j)$ is the true class label, which is 1 if class c occupies position (i, j) , or 0 otherwise. This is useful in the fully-supervised case, as the posterior can be directly supervised using ground-truth masks in full image resolution.

3.3. Importance sampling loss

Recently, importance sampling has been proposed as a substitute for global pooling during weakly supervised training [16]. It aims to solve the problem of global max pooling (GMP), which mainly targets the most discriminative region of an object. The authors in [16] model the multinomial posterior as in (5), and define a probability mass function over pixel coordinates (i, j) ;

$$p_c(I, J|x) = \Pr(I = i, J = j|x, c) = a_c(i, j)/Z(a_c), \quad (6)$$

where $Z(a_c) = \sum_{m=1}^W \sum_{n=1}^H a_c(m, n)$ is a normalizing constant for an image of width W and height H . A pixel is then drawn from this distribution for each class, and the image-level prediction \tilde{y} is computed as

$$\tilde{y}_c = a_c(\hat{i}, \hat{j}), \quad (\hat{i}, \hat{j}) \sim p_c(I, J|x). \quad (7)$$

Finally, the importance sampling loss (ISL) is the binary cross-entropy loss of the sampled prediction

$$\mathcal{L}_{\text{is}}(y, \tilde{y}) = -\frac{1}{C} \sum_{c=1}^C y_c \log \tilde{y}_c + (1 - y_c) \log(1 - \tilde{y}_c), \quad (8)$$

where y is a vector of the image-level classification labels, and y_c is the label for class c , which is 1 if class c is present anywhere in the image, and 0 otherwise.

3.4. Feature similarity loss

Empirically, objects are in most cases separated from one another, or the background, by a color edge. Jonnarth and Felsberg [16] encapsulate this prior knowledge into the self-supervised feature similarity loss function

$$\mathcal{L}_{\text{fs}}(a, x) = -\frac{1}{HW} \sum_{i,j=1}^{H \cdot W} w_{ij} g(a_i, a_j) f(\delta(x_i, x_j)), \quad (9)$$

which aligns the prediction contours with edges in the image. It is a function of the posterior $a \in [0, 1]^{H \times W \times C}$ in (5), and the RGB image $x \in [0, 1]^{H \times W \times 3}$. Note that the spatial indexing differs from above. Similar to [16], a single index is used for both of the spatial coordinates. The loss term sums over all *pixel pairs*, and consists of three components. First, a Gaussian spatial weight is computed

$$w_{ij} = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{\|p_i - p_j\|_2^2}{2\sigma^2}\right), \quad (10)$$

where p_i is the two-dimensional position vector for pixel i . This assigns a higher importance to nearby pixel pairs. Very distant pairs are disregarded, as little to no conclusion can be drawn about their semantic affinity, solely based on color information. The second component is a gating function

$$g(a_i, a_j) = \frac{1}{2} \|a_i - a_j\|_2^2, \quad (11)$$

which is the distance between the posteriors of pixels i and j . The final component is a pixel dissimilarity function

$$f(\delta) = \tanh\left(\mu + \log\left(\frac{\delta}{1-\delta}\right)\right), \quad (12)$$

$$\delta(x_i, x_j) = \frac{\|x_i - x_j\|_1}{3}, \quad (13)$$

which maps two RGB pixel values to a dissimilarity score in $[-1, 1]$. See Fig. 3 for an illustration of FSL.

The gradient is only propagated through g , as a is the network output. The weight w_{ij} controls the magnitude of the loss, while f controls the sign. A pair of similar pixels induces a positive loss, which is minimized by equal predictions. The opposite is true for dissimilar pixels. However, if the predictions are equal to begin with, then $\nabla \mathcal{L}_{\text{fs}} = 0$, allowing consistent predictions over heterogeneous image regions. The two parameters σ and μ are learned [16].

4. Modeling the binomial posterior

In this section, we first explain the problem of modeling the multinomial posterior for ISL and FSL, and proceed to present our binomial-based approach.

4.1. The limits of multinomial-based ISL and FSL

The multinomial posterior in (5), as used in the original ISL and FSL [16], assumes that classes are mutually exclusive, *i.e.* that only a single class can be present in each pixel. This assumption is reasonably sound in sufficiently high image resolution, where a single pixel is small enough to only house a single class, for all intents and purposes. Although, it may still be violated, *e.g.* at class borders where the pixel-grid does not align with the object contours, or for distant objects where the size of an object is only a few pixels. However, the assumption becomes entirely void in

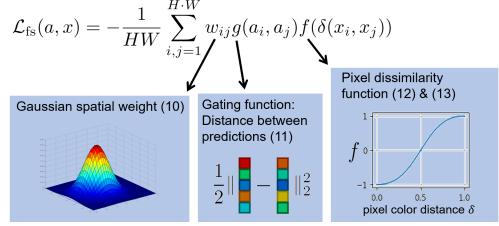


Figure 3. Illustration of the feature similarity loss (FSL).

downsampled CAMs, where the losses are applied, due to the fact that they are usually computed in a much lower resolution compared to the original input images, as a result of a fixed input size and downsampling layers in the network. This means that the assumption is not made on the pixel-level, but in fact, within a small region defined by the final downsampling scale factor. In summary, the mutual exclusivity assumption does not hold in practice. This leads to a loss in performance, as can be seen in Tab. 4.

Moreover, ISL and FSL cannot easily be applied to stronger baselines than SEAM [33], since most methods compute independent class scores [23, 31], and use GAP according to (3). We found in our initial experiments that enforcing a multinomial posterior estimation with (5) resulted in performance degradation for other baselines. A likely explanation is that, as (5) changes the characteristics of the learned CAMs [16], optimal hyperparameter values are affected. In fact, Jonnarth and Felsberg [16] found it necessary to heavily modify the background parameter α for AffinityNet [2] label generation, since the CAM characteristics had changed. Similarly, hyperparameter values inherent to other baselines may also need to be adapted.

We argue that ISL and FSL would benefit from relaxing the mutual exclusivity assumption. Instead of looking at the multinomial, we consider the C -fold binary problem, which constitutes modeling C independent binomial posteriors. We aim to adapt ISL and FSL based on this new formulation, so that they can be seamlessly implemented as add-on methods to any previous or future method, without any major modifications to the underlying baselines. Note that, in the case of unbalanced training data, the multiple binomial predictors are assumed to be calibrated, *e.g.* by means of logit-adjustment [3, 27].

4.2. Binomial-based importance sampling loss

A natural approach is to interpret the class scores, s_c , as log-odds, or logits, and model the pixel-wise binomial posterior using the sigmoid function, which is given by

$$b_c(i, j) = \frac{e^{s_c(i, j)}}{1 + e^{s_c(i, j)}} \cong \Pr(z_c(i, j) = 1|x). \quad (14)$$

The sampling distribution is attained in a similar manner as for the multinomial posterior case, namely

$$p_c(I, J|x) = b_c(i, j)/Z(b_c). \quad (15)$$

We sample the binomial $\tilde{b}_c = b_c(\hat{i}, \hat{j})$, where $(\hat{i}, \hat{j}) \sim p_c$, and the final loss is thus computed as $\mathcal{L}_{\text{is}}(y, \tilde{b})$ using (8).

Moreover, importance sampling has been shown to be superior to GMP [16]. However, Zhou *et al.* [38] found that while GMP achieves similar classification performance as GAP, GAP outperforms GMP for localization. Thus, we define the final classification loss as a convex combination of the importance sampling loss and the binary cross-entropy loss, \mathcal{L}_{ce} , based on GAP, similarly to how the original authors used GMP. Our combined classification loss is

$$\mathcal{L}_{\text{cls}} = (1 - \lambda)\mathcal{L}_{\text{ce}} + \lambda\mathcal{L}_{\text{is}}, \quad (16)$$

$$\mathcal{L}_{\text{ce}} = -\frac{1}{C} \sum_{c=1}^C y_c \log B_c + (1 - y_c) \log(1 - B_c), \quad (17)$$

where $B_c = \exp(S_c)/(1 + \exp(S_c))$ is the *image-level* GAP-based posterior estimate based on S_c in (3).

Finally, to improve stability, we propose to sample multiple pixels per class, and average their loss contributions. We write the multi-sample importance sampling loss as

$$\mathcal{L}_{\text{is}}^N = \frac{1}{N} \sum_{n=1}^N \mathcal{L}_{\text{is}}(y, \tilde{b}^{(n)}), \quad (18)$$

where N is the number of samples drawn per class, and $\tilde{b}^{(n)}$ for $n = 1, \dots, N$ are prediction vectors of length C , containing independently drawn samples. Figure 4 illustrates the difference between GMP, GAP, and multi-sample ISL.

4.3. Binomial-based feature similarity loss

The mutual exclusivity assumption made when modeling the multinomial posterior influences the feature similarity loss. For dissimilar pixels, not only are the predictions pushed apart between *pixels*, but also between *classes*. Thus, the predicted class distribution will tend towards a one-hot vector [16]. Based on the argument in Sec. 4.1, *i.e.* that multiple classes can occupy the same pixel, we argue that classes should be handled independently when reasoning about their contours. In fact, we believe that this reasoning is even more important at class borders, where two classes are likely to occupy the same pixel, especially in downsampled CAMs where FSL is applied.

Thus, we adapt FSL for binomial modeling according to (14), where the classes are considered independently, in the absence of softmax. We consider three different inputs to the gating function g in place of the multinomial posterior a : The raw scores s , the binomial b in (14), and the max-normalized CAM r in (4). A closer look at the gradients of g in (11) reveals that, when using the scores s directly,

$$\frac{\partial g(s_i, s_j)}{\partial s_{i,c}} = s_{i,c} - s_{j,c}. \quad (19)$$

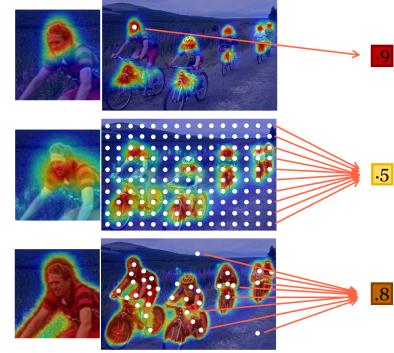


Figure 4. Illustration of global pooling methods. Top to bottom: max pooling, average pooling, multi-sample importance sampling.

The gradient is proportional to the score difference. For dissimilar pixels, g is maximized, which leads to divergence, as the gradients increase indefinitely. The gradient is bounded for the binomial posterior b , we get

$$\left| \frac{\partial g(b_i, b_j)}{\partial s_{i,c}} \right| = \left| (b_{i,c} - b_{j,c}) \frac{e^{s_{i,c}}}{(1 + e^{s_{i,c}})^2} \right| \quad (20)$$

$$\leq e^{s_{i,c}} / (1 + e^{s_{i,c}})^2 \leq 1/e^{s_{i,c}}. \quad (21)$$

The same is true for r . Assuming $s_{i,c} \geq 0$, we get

$$\left| \frac{\partial g(r_i, r_j)}{\partial s_{i,c}} \right| = \left| (r_{i,c} - r_{j,c}) \frac{1}{\max_m(s_{m,c})} \right| \quad (22)$$

$$\leq 1/\max_m(s_{m,c}). \quad (23)$$

Since the gradients are bounded for b and r , the scores do not diverge out of control whenever the prediction difference is large. Note that for r , if $s_{i,c} < s_{j,c}$ the score $s_{i,c}$ will decrease towards 0, and if $s_{i,c} < 0$ the gradient is 0. This means that if the model predicts that a class is absent, *i.e.* the score is negative, the contours are disregarded. There is no analogous lower score bound for b . Therefore, we use r for our experiments. This choice is verified in an ablation.

Furthermore, since the classes are considered independently, it no longer makes sense to compute the loss over classes that are not present. Their scores should be minimized over the entire image, which is handled by the classification loss. Enforcing prediction contours may only hurt performance, so we mask out these classes and only apply FSL on classes that are present in the image.

Finally, we find suitable FSL parameters and keep them fixed for all experiments, in contrast to learning them in [16], as this can lead to trivial solutions. For instance, the gradient with respect to μ is always negative, so μ will only increase during training. This might yield inconsistent results if the training time is changed, *e.g.* for large datasets with more gradient steps. We optimize \mathcal{L}_{fs} alone over initial Gaussian CAMs. In the spirit of WSSS, we only use one image per class to reduce the required ground-truth masks.

The resulting parameter values were $\mu = 2.5$ and $\sigma = 5$. For further details, see the supplementary material.

5. Experiments

5.1. Implementation details

Datasets. We test our approach on two datasets; *PASCAL Visual Object Classes* (VOC) [11], and the larger *MS Common Objects in Context* (COCO) [25]. VOC contains 1,464 training images, 1,449 validation images, 1,456 hold-out test images, and includes 20 foreground classes. Following the common practice, we include additional training data from SBD [13], increasing the number of training images to 10,582. For COCO, we use the 2014 train-validation split, which contains 82,783 training images, 40,504 validation images, and includes 80 classes. Note that, although pixel-level segmentation ground truth is available for the training and validation sets, they are not used for training.

Evaluation metrics. For evaluation, we use two complementary metrics as in [16]. (1) Region similarity, \mathcal{J} , which is commonly used in WSSS. It compares segmentation masks in terms of the Jaccard index [15], which is the mean intersection over union (mIoU). (2) We complement this by also evaluating the contour quality, \mathcal{F} , drawing inspiration from the well-established DAVIS benchmark for video object segmentation [30]. This is the F-score computed on the segmentation mask contours, which is efficiently approximated using morphological operations to accumulate bipartite matches of the boundaries between the predicted and ground-truth masks. Finally, we compute a combined metric, $\mathcal{J} \& \mathcal{F}$, which is simply an average of \mathcal{J} and \mathcal{F} . Mean and standard deviation is reported over five runs whenever confidence intervals are given. Note that this required reimplementation of all methods, as confidence intervals are not commonly reported in the WSSS literature.

Training details. Since we apply ISL and FSL to a wide variety of state-of-the-art baselines, we mainly use the default settings and parameters stated in the respective papers or linked code repositories. Moreover, we use the implementations referenced to in the papers, as is. Any deviations from the default settings are listed in the subsequent paragraphs. Note that we did not manage to reproduce the reported results exactly in all cases. Potential reasons include different software and hardware configurations, and the use of different evaluation code. We believe that a likely reason is that WSSS typically involves multiple steps spread out over several code repositories, which complicates the reproduction of results. For a further discussion, see the supplementary material. Still, our experiments provide good comparisons as they show how ISL and FSL can boost the performance under the same conditions as the baselines, and as the reproduced results are close to the reported results in the literature, which is evident in Tab. 1. We used four A100

Table 1. Region similarity for CAM pseudo-labels without CRF on VOC. Max and mean over five runs are shown together with the reported results from the respective papers (“Ref.” column).

Method	Set	Ref.	Reproduced			
			no ISL/FSL		with ISL/FSL	
			max	mean \pm std	max	mean \pm std
SEAM [33]	<i>train</i>	55.4	55.6	55.0 \pm 0.4	61.0	60.6 \pm 0.3
SIPE [7]	<i>train</i>	58.6	58.7	58.5 \pm 0.2	60.4	60.1 \pm 0.2
PMM [23]	<i>val</i>	56.2	55.4	55.1 \pm 0.2	59.3	58.9 \pm 0.3
MCTformer [36]	<i>train</i>	61.7	61.8	59.4 \pm 2.0	61.6	61.2 \pm 0.4
Spatial-BCE [35]	<i>train</i>	68.1	67.2	67.0 \pm 0.2	68.7	68.6 \pm 0.1

40GB GPUs in our experiments.

SEAM [33]. We set the background parameter to 1 and 24 (originally 4 and 24) for AffinityNet, when using ISL. On COCO, FSL was scaled down by a factor of 0.2.

SIPE [7]. ISL and FSL were scaled down by a factor of 0.1, since SIPE uses a higher learning rate by default. For FSL, we upsample the CAMs by a factor of 2 (from 32×32 to 64×64), to be closer to SEAM (56×56). We chose an integer factor to avoid artifacts. Note that the ResNet-101 implementation for VOC referenced by SIPE [7] uses COCO pretrained weights, which means that ground-truth segmentation masks were used, and thus, these reproduced results are not weakly supervised strictly speaking.

PMM [23]. We use ISL only during multi-scale training, and FSL during both multi-scale and multi-crop training.

MCTformer [36]. ISL and FSL were scaled down by a factor of 0.1, since MCTformer uses the AdamW optimizer [26] instead of SGD. For FSL, we upsample the CAMs by a factor of 4 (from 14×14 to 56×56) to match SEAM.

Spatial-BCE [35]. Same as the original setting, the foreground and background pixel thresholds in IRN [1] were set to 0.60 and 0.25 respectively. The semantic segmentation threshold was set to 0.30 with ISL, and 0.40 with FSL.

5.2. ISL and FSL improve pseudo-label fidelity

In Tab. 1 we evaluate pseudo-label fidelity by comparing region similarity of pseudo-labels generated from CAMs. We find that our losses have a significant effect on the pseudo-labels, resulting in an average relative performance increase of +5%. Note that we even managed to boost the performance of a transformer-based method, MCTformer [36], despite its low CAM resolution of 14×14 , and architectural differences to the CNN-based methods. Not only that, but our losses seem to stabilize its training, reducing the standard deviation over five runs from 2.0 to 0.4. While our losses increase the computational effort by 30% for the overall training pipeline, the inference time is unaffected. We further evaluate the impact of our improved pseudo-labels on final semantic segmentation performance in the weakly supervised setting.

5.3. Segmentation comparison on VOC

Fig. 1 and Tab. 2 show our reproduced final segmentation results on a wide variety of state-of-the-art baselines, including three different backbones. Our improved pseudo-labels translate into more accurate segmentation predictions, consistently outperforming the baselines. On average, \mathcal{J} , \mathcal{F} , and $\mathcal{J}\&\mathcal{F}$ are increased by +2.2%, +5.9%, and +3.7%, respectively, on the validation set. On the hold-out test set, we evaluate the best model out of five runs for each method, based on validation set performance. With ISL and FSL, the test set performance is improved by +1.6% on average. Note that, the original MCTformer [36] had a large standard deviation, and a larger *maximum* performance over five runs compared to when ISL and FSL were used. Thus, the test set performance was higher for the original method, although the *average* performance was higher with ISL and FSL. This is consistent with Tab. 1. This observation is important for practitioners who simply apply an off-the-shelf training method. A lower variance is beneficial, as the performance deviation is smaller, and the resulting model is more likely to perform as expected. Ideally, we would prefer to compute an average test set performance. However, this is prohibited by the limited number of test set submissions on VOC. Additional results are included in the supplementary material, and qualitative results are shown in Fig. 5.

5.4. Segmentation comparison on COCO

In Tab. 3, we evaluate our losses on the COCO dataset, which contains roughly 10 times more images than VOC. For SEAM [33], the segmentation results are improved in this setting as well, demonstrating that our losses are well-suited for large-scale datasets. When training the classification network for 8 epochs, we boost \mathcal{J} and \mathcal{F} by +2.5%. Our method further benefits from longer training, increasing \mathcal{J} and \mathcal{F} by an additional +2.8% and +0.7% respectively, while SEAM is less affected by the longer training.

5.5. Ablation study

For further analysis, we perform an ablation study investigating the effect of using the binomial versus multinomial, the number of sampled pixels, the loss weight λ , and the choice of gating function, with SEAM [33] as the baseline.

Binomial versus multinomial. To isolate the effect of modeling the binomial posterior versus the multinomial posterior, we evaluate the baseline method SEAM using both, and apply ISL and FSL separately. The segmentation performance is displayed in Table 4. Both models are trained with their respective optimal values for λ , which are 0.2 for the binomial, and 0.6 for the multinomial, when used with GAP. The background parameter α , for amplifying and weakening the class scores during CRF for AffinityNet, are also chosen to suit the respective models. 1 and 24 are used for the binomial, and 2 and 4 for the multinomial. With

Table 2. Final segmentation performance on VOC, in terms of region similarity (\mathcal{J}), contour quality (\mathcal{F}), and combined ($\mathcal{J}\&\mathcal{F}$).

Method	Backb.	validation			\mathcal{J}
		\mathcal{J}	\mathcal{F}	$\mathcal{J}\&\mathcal{F}$	
SEAM [33]	Res38	63.9 ± 0.5	39.9 ± 0.2	51.9 ± 0.3	65.4
+ISL/FSL (ours)	Res38	66.7 ± 0.2	44.7 ± 0.1	55.7 ± 0.1	67.6
SIPE [7]	Res38	68.0 ± 0.2	45.1 ± 0.2	56.6 ± 0.1	68.9
+ISL/FSL (ours)	Res38	68.3 ± 0.2	46.8 ± 0.2	57.6 ± 0.1	69.4
SIPE [7]	Res101	68.5 ± 0.2	41.9 ± 0.5	55.2 ± 0.3	69.4
+ISL/FSL (ours)	Res101	69.4 ± 0.2	43.9 ± 0.2	56.7 ± 0.1	70.3
PMM [23]	Res38	64.7 ± 0.5	44.5 ± 0.5	54.6 ± 0.4	65.7
+ISL/FSL (ours)	Res38	66.7 ± 0.7	46.3 ± 0.6	56.5 ± 0.6	67.0
MCTformer [36]	DeiT-S	67.5 ± 1.7	46.7 ± 0.7	57.1 ± 1.2	70.6
+ISL/FSL (ours)	DeiT-S	68.3 ± 0.7	47.3 ± 0.3	57.8 ± 0.5	70.0
Spatial-BCE [35]	Res38	68.1 ± 0.1	45.4 ± 0.1	56.7 ± 0.1	68.4
+ISL/FSL (ours)	Res38	69.3 ± 0.2	48.2 ± 0.1	58.8 ± 0.1	69.4
Spatial-BCE [35]	Res101	67.9 ± 0.2	46.1 ± 0.1	57.0 ± 0.1	68.4
+ISL/FSL (ours)	Res101	70.1 ± 0.2	50.5 ± 0.2	60.3 ± 0.2	70.6

Table 3. Final segmentation performance on the COCO validation set, in terms of region similarity (\mathcal{J}), contour quality (\mathcal{F}), and combined ($\mathcal{J}\&\mathcal{F}$). The classification networks were trained for 8 epochs (three runs) and 16 epochs (single run with fixed seed).

Method	Epochs	\mathcal{J}	\mathcal{F}	$\mathcal{J}\&\mathcal{F}$
SEAM [33]	8	35.9 ± 0.2	28.2 ± 0.3	32.1 ± 0.3
+ISL/FSL (ours)	8	36.8 ± 0.2	28.9 ± 0.1	32.9 ± 0.1
SEAM [33]	16	35.5	28.5	32.0
+ISL/FSL (ours)	16	37.8	29.1	33.4

Table 4. Comparison between modeling the multinomial posterior in (5) versus the binomial posterior in (14) for ISL and FSL, with SEAM [33]. The region similarity (\mathcal{J}), contour quality (\mathcal{F}), and averaged ($\mathcal{J}\&\mathcal{F}$) are computed on the VOC validation set.

Loss	Model	\mathcal{J}	\mathcal{F}	$\mathcal{J}\&\mathcal{F}$
ISL	Multinomial [16]	63.0 ± 0.5	41.0 ± 0.4	52.0 ± 0.4
	Binomial (proposed)	64.3 ± 0.8	42.2 ± 0.6	53.3 ± 0.7
FSL	Multinomial [16]	50.3 ± 13	41.4 ± 8.5	45.9 ± 11
	Binomial (proposed)	66.9 ± 0.4	43.5 ± 0.2	55.2 ± 0.2

ISL, modeling the binomial improves the region similarity by +1.3 points over the multinomial, and the contour quality is improved by +1.2 points. With FSL, training becomes unstable when modeling the multinomial. We observed a big variance over five runs, with the region similarity ranging from 40 to 61. However, it is worth noting that for successful runs, the contour quality reached as far as 48. But on average, modeling the multinomial underperformed the binomial. Note that, not only is the performance improved with the binomial. A further benefit is that it allows ISL and FSL to be applied seamlessly to essentially any method.

Number of sampled pixels for ISL. In Tab. 5 we show

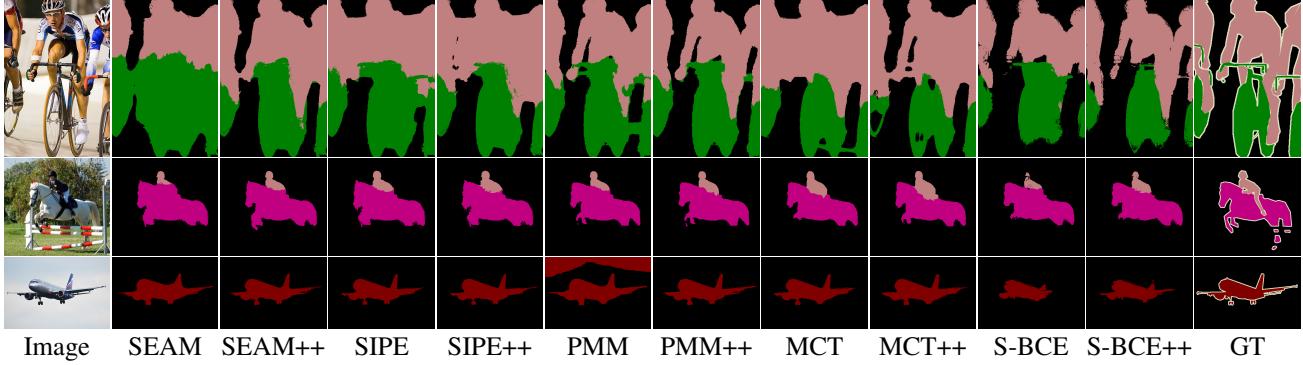


Figure 5. Qualitative results for the implemented methods on VOC. “++” indicates that ISL and FSL were used for training.

Table 5. Region similarity (\mathcal{J}), contour quality (\mathcal{F}), and averaged ($\mathcal{J} \& \mathcal{F}$) on the VOC training set when varying the number of sampled pixels, with $\lambda = 1$ and without FSL.

#pixels	\mathcal{J}	\mathcal{F}	$\mathcal{J} \& \mathcal{F}$
1	61.6 ± 0.8	40.4 ± 0.6	51.0 ± 0.7
5	62.6 ± 0.6	40.7 ± 0.6	51.6 ± 0.6
10	63.0 ± 0.4	41.2 ± 0.4	52.1 ± 0.4
50	63.0 ± 0.6	41.3 ± 0.7	52.1 ± 0.6
100	62.9 ± 0.7	41.2 ± 0.7	52.0 ± 0.7

segmentation results when varying the number of sampled pixels during importance sampling. We observe a boost in performance when increasing the number of samples up to 10, after which it plateaus. Thus, we fix the number of samples to 10, in order to keep the computational cost as low as possible while maintaining a high segmentation performance. A likely reason for why multiple samples improve results is that individual samples have less of an impact. This is vital for the early parts of training where predictions are mostly random. A single poor sample, such as a misclassified foreground pixel, is less likely to hamper learning when multiple samples are drawn. This effect diminishes beyond a certain point, as can be seen in Tab. 5.

Optimal loss weight λ for ISL. Figure 6 shows the segmentation performance when varying the loss weight λ in (16). Note that $\lambda = 0$ corresponds to the standard GAP-based classification loss in (17) without ISL, while $\lambda > 0$ uses ISL. ISL improves over the standard loss by +1.6 points with $\lambda = 0.2$. FSL further boosts the performance by +2.2 points. ISL had a larger impact on the contour quality, while FSL improved both metrics roughly the same.

Gating function in FSL. We compare $g(r_i, r_j)$ based on the max-normalized CAM r in (4), and $g(b_i, b_j)$ based on the posterior estimate b in (14). The max-normalized CAM, r , performs better with $\mathcal{J}=66.9$ (cf. 64.6), $\mathcal{F}=43.5$ (cf. 43.0), and $\mathcal{J} \& \mathcal{F}=55.2$ (cf. 53.8). This could be attributed to the fact that the gradient is suppressed too much for b , which is upper bounded by $\exp(-x)$, and diminishes faster than $1/x$ for r . See Sec. 4.3. Additionally, $g(r_i, r_j)$ induces a lower

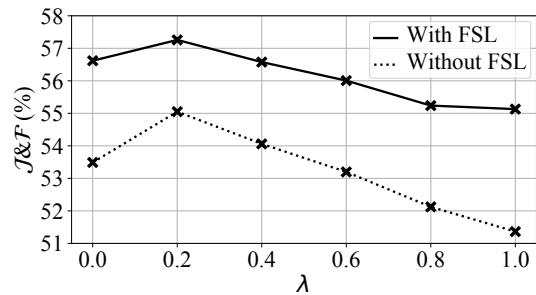


Figure 6. Combined score ($\mathcal{J} \& \mathcal{F}$) on the VOC training set, as a function of the importance sampling loss parameter λ , with and without feature similarity loss (FSL), with $\mu = 2.5$ and $\sigma = 5$.

score bound for when FSL is applied. This is not the case for $g(b_i, b_j)$, which may result in artificially large negative scores that are not reflective of the true class probability.

6. Conclusions

In this work, we generalize two previous techniques, importance sampling, and feature similarity loss, for weakly-supervised segmentation. Based on the argument that object classes are not mutually exclusive within image regions, we model multiple binary problems as opposed to the multinomial posterior. This results in an add-on method that boosts pseudo-label fidelity, which in turn improves segmentation performance, as shown on several state-of-the-art baselines.

Acknowledgments

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP), funded by the Knut and Alice Wallenberg (KAW) Foundation. The computational resources were provided by the National Academic Infrastructure for Supercomputing in Sweden (NAIIS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725, and by the Berzelius resource, provided by the KAW Foundation at the National Supercomputer Centre (NSC).

References

[1] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2209–2218, 2019. [2](#), [6](#)

[2] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4981–4990, 2018. [2](#), [3](#), [4](#)

[3] Emanuel Sanchez Aimar, Arvi Jonnarth, Michael Felsberg, and Marco Kuhlmann. Balanced product of calibrated experts for long-tailed recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19967–19977, 2023. [4](#)

[4] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What’s the point: Semantic segmentation with point supervision. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, pages 549–565. Springer, 2016. [1](#)

[5] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What’s the point: Semantic segmentation with point supervision. In *European Conference on Computer Vision*, pages 549–565. Springer, 2016. [2](#)

[6] Liyi Chen, Weiwei Wu, Chenchen Fu, Xiao Han, and Yuntao Zhang. Weakly supervised semantic segmentation with boundary exploration. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*, pages 347–362. Springer, 2020. [2](#)

[7] Qi Chen, Lingxiao Yang, Jian-Huang Lai, and Xiaohua Xie. Self-supervised image-specific prototype exploration for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4288–4298, June 2022. [3](#), [6](#), [7](#)

[8] Zhaozheng Chen, Tan Wang, Xiongwei Wu, Xian-Sheng Hua, Hanwang Zhang, and Qianru Sun. Class re-activation maps for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 969–978, June 2022. [3](#)

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. [1](#)

[10] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1635–1643, 2015. [1](#), [2](#)

[11] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010. [6](#)

[12] Monica Gruosso, Nicola Capece, and Ugo Erra. Human segmentation in surveillance video with deep learning. *Multimedia Tools and Applications*, 80:1175–1199, 2021. [1](#)

[13] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *2011 International Conference on Computer Vision*, pages 991–998. IEEE, 2011. [6](#)

[14] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7014–7023, 2018. [2](#)

[15] Paul Jaccard. Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37:241–272, 1901. [6](#)

[16] Arvi Jonnarth and Michael Felsberg. Importance sampling CAMs for weakly-supervised segmentation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 2639–2643, 2022. [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#)

[17] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 876–885, 2017. [2](#)

[18] Alexander Kolesnikov and Christoph H Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *European Conference on Computer Vision*, pages 695–711. Springer, 2016. [2](#)

[19] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *Advances in Neural Information Processing Systems*, 24:109–117, 2011. [2](#)

[20] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5267–5276, 2019. [2](#)

[21] Jungbeom Lee, Eunji Kim, and Sungroh Yoon. Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4071–4080, 2021. [2](#)

[22] Jungbeom Lee, Jihun Yi, Chaehun Shin, and Sungroh Yoon. Bbam: Bounding box attribution map for weakly supervised semantic and instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2643–2652, 2021. [2](#)

[23] Yi Li, Zhanghui Kuang, Liyang Liu, Yimin Chen, and Wayne Zhang. Pseudo-mask matters in weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6964–6973, October 2021. [4](#), [6](#), [7](#)

[24] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for

semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3159–3167, 2016. 1, 2

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014. 6

[26] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 6

[27] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations (ICLR)*, 2021. 4

[28] Shervin Minaee, Yuri Y Boykov, Fatih Porikli, Antonio J Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 2021. 1

[29] George Papandreou, Liang-Chieh Chen, Kevin P Murphy, and Alan L Yuille. Weakly- and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1742–1750, 2015. 2

[30] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 724–732, 2016. 6

[31] Yukun Su, Ruizhou Sun, Guosheng Lin, and Qingyao Wu. Context decoupling augmentation for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7004–7014, October 2021. 4

[32] Paul Vernaza and Manmohan Chandraker. Learning random-walk label propagation for weakly-supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7158–7166, 2017. 2

[33] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12275–12284, 2020. 1, 2, 3, 4, 6, 7

[34] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1568–1576, 2017. 2

[35] Tong Wu, Guangyu Gao, Junshi Huang, Xiaolin Wei, Xiaoming Wei, and Chi Harold Liu. Adaptive spatial-bce loss for weakly supervised semantic segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIX*, pages 199–216. Springer, 2022. 6, 7

[36] Lian Xu, Wanli Ouyang, Mohammed Bennamoun, Farid Boussaid, and Dan Xu. Multi-class token transformer for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4310–4319, June 2022. 3, 6, 7

[37] Sung-Hoon Yoon, Hyeokjun Kweon, Jegyeong Cho, Shinjeong Kim, and Kuk-Jin Yoon. Adversarial erasing framework via triplet with gated pyramid pooling layer for weakly supervised semantic segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIX*, pages 326–344. Springer, 2022. 3

[38] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016. 1, 2, 3, 5