

Learning to Detour: Shortcut Mitigating Augmentation for Weakly Supervised Semantic Segmentation

JuneHyung Kwon¹, Eunju Lee², Yunsung Cho², YoungBin Kim^{1,2}

¹Department of Artificial Intelligence, Chung-Ang University, Korea

²Graduate School of Advanced Imaging Science, Multimedia & Film, Chung-Ang University, Korea

{dirchdmltnv, dmswn5829, cho4062002, ybkim85}@cau.ac.kr

Abstract

Weakly supervised semantic segmentation (WSSS) employing weak forms of labels has been actively studied to alleviate the annotation cost of acquiring pixel-level labels. However, classifiers trained on biased datasets tend to exploit shortcut features and make predictions based on spurious correlations between certain backgrounds and objects, leading to a poor generalization performance. In this paper, we propose shortcut mitigating augmentation (SMA) for WSSS, which generates synthetic representations of object-background combinations not seen in the training data to reduce the use of shortcut features. Our approach disentangles the object-relevant and background features. We then shuffle and combine the disentangled representations to create synthetic features of diverse object-background combinations. SMA-trained classifier depends less on contexts and focuses more on the target object when making predictions. In addition, we analyzed the behavior of the classifier on shortcut usage after applying our augmentation using an attribution method-based metric. The proposed method achieved the improved performance of semantic segmentation result on PASCAL VOC 2012 and MS COCO 2014 datasets.

1. Introduction

Acquiring pixel-level labels for semantic segmentation is both cost- and time-intensive. To alleviate this problem, weakly supervised semantic segmentation (WSSS) is actively researched by different groups. WSSS uses a weak label that contains less information about the location of an object than a pixel-level label but has a cheaper annotation cost. Examples of such weaker forms of labels are image-level class labels [20, 22, 25], bounding boxes [16, 24, 32], points [4, 18], and scribbles [28, 35]. Among these weak labels, we focus on the image-level class label, which is the most accessible and has the lowest annotation cost.

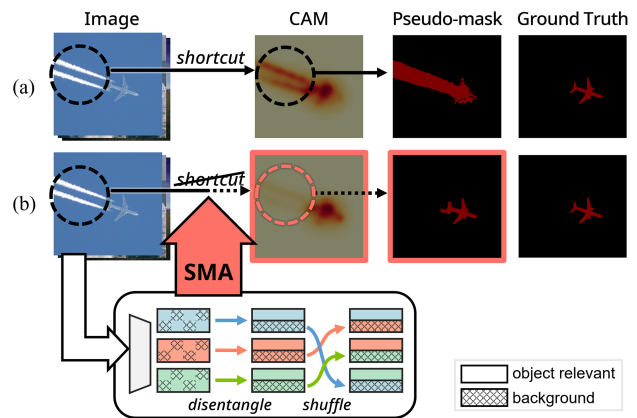


Figure 1. (a) The classifier trained with CutMix utilizes background cues as shortcuts producing suboptimal localization map. (b) Our proposed SMA relies less on contexts and focuses more on objects to generate fine-grained localization maps.

In most studies on WSSS utilizing image-level class labels, class activation map (CAM) is used as the initial seed which can be seen as estimation of the region covered by the object. The classifier is trained to predict the corresponding category of the image and identify the target object region. However, the classifier unexpectedly highlights the background region, which in turn generates a blurry CAM [28]. This is because the classifier exploits the bias of the dataset as a shortcut instead of making predictions using the information related to the target object. Here, a “shortcut” refers to a decision rule that performs well in typical contexts but fails in slightly different circumstances [13]. This background bias originates from the biased dataset composed of images in which a specific object frequently co-occurs with a certain background context [13]. For instance, an object corresponding to the “sheep” class almost always appears in “grass landscape” and the visual layout is similar in many images. Among the Pascal VOC 2012 training datasets, more than 20% of the images with “sheep”

class contain “grass landscape” as the context. The same trend is observed for cow–grass landscape, boat–water, and aeroplane–sky train–track pairs [23]. Therefore, a classifier trained with a biased dataset depends not only on the target object but also leverages the background bias as a shortcut to make predictions. Because the classifier benefits from this non-generalizable shortcut feature, it occasionally assigns high class scores on the background regions, and fails to activate the target object region, where such objects appear outside typical scenes.

Several studies have been conducted on data augmentation [9, 43, 44] that enables the training of classifiers with enhanced robustness [6]. These methods minimize the effect of biases on the classifier and aid in the identification of non-discriminative regions of the target object while making predictions. However, these approaches are susceptible to dataset bias as they do not take into account the contextual information of the image during the augmentation of the training samples. Figure 1 (a) illustrates the localization map produced by applying a conventional augmentation. The CutMix [43]-trained classifier leverages “sky” region (especially contrail part), which often appears together with object “aeroplane”, as a shortcut, resulting in an inaccurate CAM and the pseudo-mask.

In this paper, we propose **Shortcut Mitigating Augmentation (SMA)**, which prevents the use of shortcuts in the classifiers, for WSSS. We reduce the shortcut usage degree by presenting the classifier with synthetic features of the object–background combinations that do not frequently appear in the training dataset. First, we disentangle the object-relevant and background features, because these features are highly entangled with each other and mislead the classifier into identifying background as objects of interest. Each feature is exclusively aggregated, and contrastive learning is adopted for further separation. Second, based on these representations, we randomly shuffle the background representation in each mini-batch, while the foreground representation is fixed, and vice versa. Through the proposed SMA, the classifier receives representations of various object-background combinations, allowing it to focus more on the target object without being biased towards a specific background. As a result, the dependence of the classifier on the shortcuts is reduced, and a high-quality CAM is obtained.

We performed an additional experiment to evaluate the extent to which each augmentation method relied on shortcuts. To analyze the classifier’s shortcut behavior, we introduced a metric based on attribution methods. The experimental results showed that compared to the existing augmentation method, the classifier trained with SMA employed more object-relevant information and less background information when identifying the target objects. By leveraging the shortcut mitigation effect of SMA, the mean

Intersection over Union (mIoU) of the localization map and pseudo-mask were enhanced on the Pascal VOC 2012 and MS COCO 2014 datasets.

The main contributions of this study are as follows: We (1) proposed SMA, a data augmentation method for WSSS that synthesizes the novel representation of various object-background combinations (2) evaluated the effectiveness of our method on reducing shortcut usage by adopting attribution method-based metric (3) enhanced the performance of the localization map and pseudo-mask by applying the SMA method on various WSSS methods using the Pascal VOC 2012 and MS COCO 2014 benchmarks for the WSSS.

2. Related work

2.1. Weakly supervised semantic segmentation

WSSS methods that use image-level class labels generate a localization map based on the initial seed CAM, and then produce a pseudo-mask through an additional refinement process. Because the initial seed identifies only the discriminative regions in the image, numerous studies have been conducted to expand such regions. AdvCAM [22] identifies more regions of objects by manipulating the attribute map through adversarial climbing of the class scores. DRS [17] suppresses the most discriminative region, thereby enabling the classifier to capture even the non-discriminative regions. SEAM [38] regularizes the classifier so that the differently transformed localization maps are equivalent. AMN [25] leverages a less discriminative part through per-pixel classification.

Further, several studies have been conducted to develop feasible methods to prevent the classifier from learning misleading correlations between the target object and the background. SIPE [7] captures the object more accurately through prototype modeling of the background. ICD [11] includes an intra-class discriminator that discriminates the foreground and background within the same class. W-OoD [23] utilizes out-of-distribution data as extra supervision to train the classifier to suppress spurious cues. C^2AM [41] generate class-agnostic activation maps by adopting foreground-background contrastive learning. While C^2AM leverages separated representation to refine CAM, SMA synthesizes the disentangled representation to diversify the views learned by the classifier. In addition, various studies have employed a saliency map as an additional supervision or for post-processing [12, 21, 26, 39, 40, 42]. In contrast, our proposed method disentangles the background information in the representation space, and thus, no additional supervision is required.

2.2. Data augmentation

Data augmentation improves the generalization ability of a classifier for unseen data by improving the diversity of the

training data. The image erasing method removes one or more sub-regions in an image and replaces them with zero or random values. Cutout [9] randomly masks a specific part of the image, and Hide-and-Seek [31] allows the classifier to identify the class relevant features after randomly hiding the patch in the image. In contrast, the image mix-based method mixes two or more images. Mixup [44] interpolates two images and labels, and CutMix [43] replaces a certain region of an image with a patch of another image. However, because the method that uses this regional patch randomly occludes the sub-regions, including both the object and background areas, the classifier trained with this method cannot distinguish the foreground from the context. In addition, there are methods to augment the representation in the feature space [27, 37]. Furthermore, there have been similar approaches to enhance the diversity of the training dataset by combining features of different instances [19, 45]. SMA differs in its approach by explicitly disentangling two features and bidirectionally shuffling them.

In context decoupling augmentation (CDA) [33], the copy-and-paste method is introduced to a WSSS task. CDA decouples the object and context by pasting the pseudo-mask obtained in the first stage to another image. However, this method uses a single class image to obtain an accurate object instance and restricts the scale of the mask. Consequently, the diversity of the augmented representation is limited. Our proposed method synthesizes features irrespective of the number or size of the object mask, and thus, it can provide a more diverse representation to the classifier.

3. Method

In section 3.1, we introduce the SMA method, which generates synthetic features that encapsulate combinations of objects and backgrounds that do not frequently occur together. The degree of shortcut usage of the SMA and other augmentation methods is analyzed in section 3.2. Finally, section 3.3 describes additional training schemes for performing an effective augmentation.

3.1. Shortcut mitigating augmentation

Here, we describe the overall procedure of the SMA. The proposed method disentangles the input features into object-relevant and background features, and then shuffles different training samples to generate synthetic features. A classifier trained with SMA minimizes the use of background cues as shortcuts and leverages more target objects to create higher-level localization maps.

Disentangling object-relevant and background features.

First, we disentangle the features obtained through the backbone network into object-relevant and background features. To separate distinctive features effectively, mutually exclusive attributes must be aggregated from the input feature. Global average pooling (GAP), which is generally

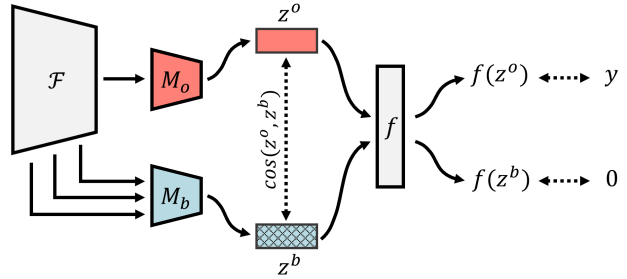


Figure 2. An illustration demonstrating how the classifier separates object-relevant features from background features. The two aggregators M_o and M_b disentangle the features obtained from the backbone network \mathcal{F} into object-relevant feature z^o and background feature z^b . Then two classification score $f(z^o), f(z^b)$ obtained by classification head $f(\cdot)$ is supervised by label y and 0 respectively. In order to further differentiate z^o, z^b , contrastive loss is additionally employed.

used for pooling, is not appropriate as it can include peripheral attributes. Therefore, inspired by previous studies [3, 20, 46], we aggregated the object-relevant and background features using different aggregators M_o, M_b respectively. In addition, to integrate the background related information more effectively, we extract features from shallow layers of the backbone network, since semantics of background are not restricted [7]. We fed extracted features to M_b . The distinct features are aggregated by M_o, M_b through the following process.

Given N images $\{x_i\}_{i=1}^N$, the feature map $\{z_i\}_{i=1}^N$ is obtained through the backbone network $\mathcal{F}(\cdot)$, in which $z_i \in \mathbb{R}^{C \times h \times w}$. C, h , and w indicate the channel size, height, and width of the feature map, respectively. Based on the extracted feature map z_i , different 1×1 convolutions $\phi(\cdot), \theta(\cdot)$ are employed, and Softmax operation is performed on the spatial dimension to produce the attention maps $o_i \in \mathbb{R}^{d \times h \times w}, b_i \in \mathbb{R}^{d \times h \times w}$. The attention maps o_i, b_i can be seen as spatial importance for and object-relevant and background regions, respectively. o_i, b_i , and z_i are flattened, i.e., $o_i \in \mathbb{R}^{d \times hw}, b_i \in \mathbb{R}^{d \times hw}$, and $z_i \in \mathbb{R}^{C \times hw}$. Then, the disentangled representation z_i^o, z_i^b can be calculated as follows:

$$z_i^o = m(o_i \otimes z_i^T), z_i^b = m(b_i \otimes z_i^T) \quad (1)$$

where \otimes and $m(\cdot)$ denote the matrix multiplication and average operations over the channel dimension, respectively; $z_i^o \in \mathbb{R}^{1 \times C}, z_i^b \in \mathbb{R}^{1 \times C}$ are the object-relevant and background representations, respectively. We effectively separate the disentangled representations by supervising them with different labels for each logit value obtained using the classification head $f(\cdot)$. We supervise logit value $f(z^o)$ with class label y , so that z^o contains object relevant attributes. Following [23, 46], we assign zero vector

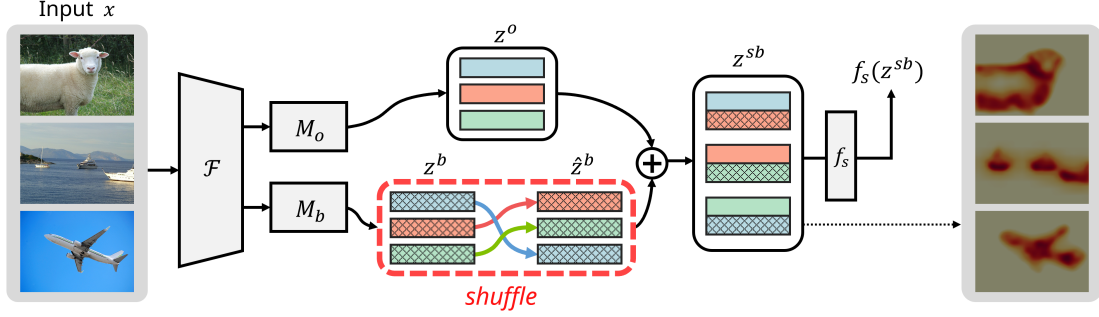


Figure 3. This is an overview of SMA when only background feature shuffling is implemented. The feature map z is disentangled into object-relevant representation z^o and background representation z^b by aggregators M_o and M_b . We perform background shuffling to acquire \hat{z}^b in which the background representation is permuted on the mini-batch. Then novel representation z^{sb} is synthesized by concatenating z^o and \hat{z}^b . $f_s(z^{sb})$ denotes prediction by f_s and \oplus denotes channel-wise concatenation operation.

label $y = (0, 0, \dots, 0)$ to score $f(z^b)$ to ensure z^b only includes background attributes. Further, binary cross entropy (BCE) is used for training the classifier, and the classification loss is determined as follows:

$$L_{cls} = BCE(f(z_i^o), y) + BCE(f(z_i^b), 0) \quad (2)$$

We adopt contrastive loss to separate distinct representations further. Our training objective forces the object-relevant representation z^o and background representation z^b apart. The contrastive loss used in this case is calculated as follows:

$$L_{contr} = -\frac{1}{N} \sum_{i=1}^N \log(1 - sim(z_i^o, z_i^b)) \quad (3)$$

where $sim(z_i^o, z_i^b)$ denotes the cosine similarity between the two representations; z_i^o and z_i^b are disentangled such that the distance between them on the feature space increases through the contrastive loss. The overall process of feature disentanglement can be seen in Figure 2.

Shuffling disentangled representations. We randomly shuffle the disentangled representations and combine them to produce synthetic representation. Synthesizing these distinct representation of different samples is crucial, because the optimal classifier should consistently identify target object irrespective of background. The features are separated to be mutually exclusive with z^o and z^b . z^o is highly related to the class label prediction, whereas z^b is correlated with the object, although it is not necessary to predict the class label. In other words, for an image x , even if the background representation z^b is replaced with another representation z^{b*} , the predictions are assumed to remain unaffected. Therefore, the optimal classifier f^* should provide consistent predictions without being affected by biases. Our hypothesis can be expressed as Eq 4.

$$f^*([z^o, z^b]) = f^*([z^o, z^{b*}]) \quad (4)$$

$[\cdot, \cdot]$ denotes the channel-wise concatenation; that is, to achieve such an invariant prediction, shuffling the separate representation is employed. First, we randomly permute the disentangled background representation and obtain \hat{z}^b . Then, we concatenate \hat{z}^b with z^o to create a novel representation z^{sb} represents a fixed object-relevant representation combined with a swapped background representation from other image in the mini-batch. Additionally, to provide a more diverse representation to the classifier, the augmentation is also performed in the opposite direction. That is, \hat{z}^o is obtained by randomly shuffling z^o and then obtaining $z^{so} = [\hat{z}^o, z^b]$ by concatenating with z^b . Then, we feed z^{sb} to the classifier f_s and supervise the classification score with the target label y . In the case of z^{so} , the target object is shuffled in the mini-batch, and thus, the target label y is also rearranged to \hat{y} , according to the permuted index.

$$L_{shuffle} = BCE(f_s(z^{sb}), y) + BCE(f_s(z^{so}), \hat{y}) \quad (5)$$

The objective function for training the classifier augmented with the shuffled representation can be expressed as Eq 5.

$$L = L_{cls} + \lambda L_{contr} + L_{shuffle} \quad (6)$$

Thus, the total loss function can be described as Eq 6, where λ represents the balancing scalar. The overview of SMA are shown in Figure 3. More results on λ can be found in supplementary material.

Discussions. Two-way shuffling combines the object-relevant attributes in the representation space with the background attributes that do not appear frequently with the corresponding class. Consequently, the classifier learns the representation that rarely appears in the biased dataset and is thus less dependent on the shortcut features. For example, when images of an “aeroplane” with “sky” and “cows and sheeps” with “grass landscape” are used as the input

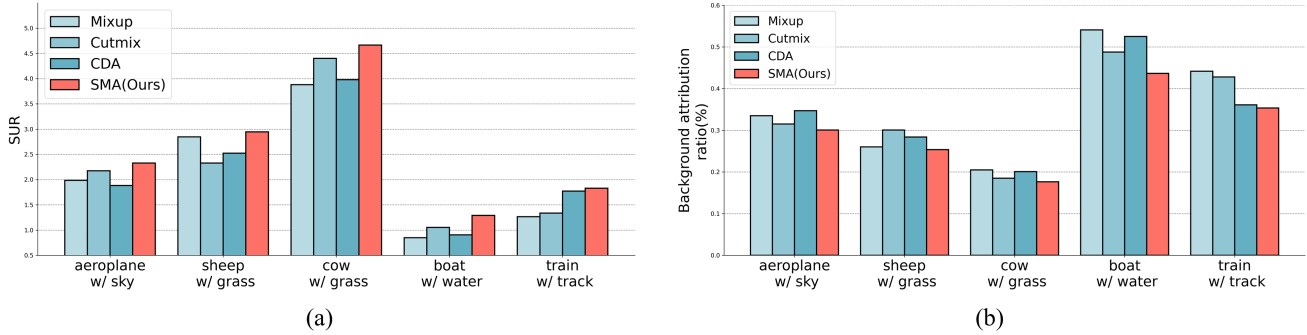


Figure 4. For each object–background pair, we evaluated (a) shortcut usage ratio (SUR) (higher the better) and (b) background attribution ratio (lower the better) for each augmentation including Mixup [44], CutMix [43], CDA [33], and SMA(Ours).

images, the classifier can learn the representations related to “aeroplane with grass landscape” and “cows and sheeps appearing in sky” in the representation space, which the images of such scenes are absent in the training dataset. Furthermore, the diversity of the representation is guaranteed, because the representations in the mini-batch are randomly combined in every iteration.

3.2. Analysis of the shortcut behaviors

In this section, we designed a toy experiment to analyze the shortcut learning behavior of the classifier. First, we collected aeroplane–sky, sheep–grass, cow–grass, boat–water, and train–track images, which are the most frequent object–background pairs in the Pascal VOC 2012 dataset [10]. The Pascal CONTEXT [29] dataset was used to determine whether the target background was included. Then, ResNet-50 [15] pretrained with ImageNet [8] was trained to predict the class labels with the Pascal VOC 2012 dataset after applying different augmentation methods, i.e., Mixup [44] and CutMix [43], which are popular augmentation methods used in classification tasks, as well as CDA [33], which is applied to WSSS tasks.

The degree of shortcut usage by the classifier can be measured by the contribution of object region (R_o) pixels compared to background region (R_b) pixels in predicting class labels. To quantify this, we leverage Integrated Gradients (IG) [34], which is a gradient-based method that integrates m gradients along a straight-line path from a baseline x_{base} to an input x . Here, we use a black image with the same resolution as the input image x as the baseline input x_{base} [34]. The IG of the i -th pixel of input image x can be expressed as follows:

$$I(x_i) = (x_i - x_{base}) \cdot \sum_{k=1}^m \frac{\partial f(x_{base} + \frac{k}{m}(x_i - x_{base}))}{\partial x_i} \cdot \frac{1}{m} \quad (7)$$

We define the **shortcut usage ratio (SUR)** of an image as the ratio of IG in the background region (R_b) to that of the

object region (R_o). SUR indicates the extent to which the information from the object region is utilized compared to the background region. SUR can be calculated as follows:

$$SUR(x) = \frac{\sum_{i \in R_o} I(x_i)}{\sum_{i \in R_b} I(x_i)} \quad (8)$$

In addition, to directly evaluate the extent of mitigated shortcut usage, we calculated the **background attribution ratio (BAR)** between the attributions in the background region and the sum of all attributions when predicting the target class. Shortcut is an unintended decision rule when making predictions, and in our case, the background represents this. Therefore, it is possible to measure the extent to which a classifier uses shortcuts by 1) object-relevant attributions and 2) background attributions when predicting class labels, which can be measured by SUR and BAR respectively.

We computed the average SUR and BAR for all samples of each object–background pair, which can be seen in Figure 4 (a) and (b), respectively. Experimental results show that SMA accomplished highest SUR in all cases of object–background pairs compared to the existing augmentation method. Also, our proposed method achieved a lowest background attribution ratio among the existing methods in all cases. It can be confirmed that the classifier trained through SMA is less dependent on shortcut features and leverage object relevant features more when making predictions.

3.3. Training schemes

We used several training schemes to improve the performance of the classifier. First, our proposed augmentation method was applied after a certain epochs. We assumed that the two representations are sufficiently separated, before implementing shuffling. However, if the synthesizing is performed when the features are still entangled, then the classifier will be trained with the wrong signals. The classifier preferentially learns the bias-aligned sample (e.g., boat

Method	Seed	Mask
PSA [2] _{CVPR'18}	48.0	61.0
+ SMA (Ours)	51.4	64.1
IRN [1] _{CVPR'19}	48.3	66.3
+ SMA (Ours)	52.4	68.6
AdvCAM [22] _{CVPR'21}	55.6	69.9
+ SMA (Ours)	57.8	70.4
AMN [25] _{CVPR'22}	62.1	72.2
+ SMA (Ours)	64.4	72.7

Table 1. We evaluate the mIoU (%) of localization map (Seed) and pseudo-mask (Mask) on various WSSS methods when SMA was applied on Pascal VOC 2012 *train* set.

with water) at the beginning of the learning process and learns the bias conflicting sample (e.g., boat with railroad) later [30]. In other words, it is difficult for the classifier to distinguish the target-object-related features from the bias in the early stages of learning. Therefore, the augmentation is implemented after a specific epochs t_{aug} , in which the two features are sufficiently disentangled.

In addition, we do not update the concatenated background features based on the shuffle loss $\mathcal{L}_{shuffle}$. The shuffled feature is supervised by the target labels y and \hat{y} , and if the background representation is affected by such a supervision, then the classifier can be wrongly updated. Therefore, to prevent this unintended representation learning, the corresponding feature is detached, so that the shuffle loss does not backpropagate to the background features z^b and z^{sb} .

4. Experiment

4.1. Experimental setup

Dataset and evaluation metric. We used the Pascal VOC 2012 [10] dataset with a total of 21 classes (20 object categories and 1 background) and 10,582 image-level class label images augmented by [14] for the training. For the evaluation and testing, 1,449 and 1,456 pixel-level labels were used. We also performed training and evaluation on the MS COCO 2014 [5] dataset. The MS COCO 2014 dataset consists of 81 classes, including background, with 82,783 and 40,504 images for training and validation, respectively. We analyzed the performances of the generated localization map, pseudo-mask, and semantic segmentation benchmark using the mIoU metric, which is generally used for evaluating segmentation results.

Implementation details. We adopted ResNet-50 [15] pre-trained on ImageNet [8] as the classifier, and the stochastic gradient descent was used as the optimizer. The learning rate was set to 0.1 at the beginning of the training, and then decreased for every iteration by using a polynomial decay

Method	<i>val</i>	<i>test</i>
PSA [2] _{CVPR'18}	61.7	63.7
+ SMA (Ours)	65.9	66.8
IRN [1] _{CVPR'19}	63.5	64.8
+ SMA (Ours)	68.6	68.7
AMN [25] _{CVPR'22}	69.5	69.6
+ SMA (Ours)	70.9	70.8

Table 2. We show the results of training the semantic segmentation network with the pseudo-mask obtained by applying SMA to the existing WSSS methods for the Pascal VOC 2012 *val* and *test* set.

function. The classifier was trained for 10 epochs, and the point at which the augmentation was applied, t_{aug} was set to 6 through extensive experiments. The additional results of t_{aug} are found in supplementary material.

4.2. Experimental results

Quantitative results of localization maps and pseudo-masks. Table 1 compares the initial seeds (Seed) and pseudo-masks (Mask) obtained by using the WSSS baseline and our proposed methods on Pascal VOC 2012 dataset. When our method was applied to different baselines [1, 2, 25], their performances showed considerable improvements. In particular, when we applied our method on AMN [25], an mIoU value of 72.7% was achieved for the pseudo-mask, and this value exceeded those obtained using the previous methods. Compared to other methods, SMA focuses more on the target object while exploiting less background cues. Figure 5 (a) shows examples of the pseudo-masks generated through the CDA [33], IRN [1], and proposed method. These examples also show that compared to the CDA, our method captures the target object more accurately and does not mis-assign the background region to the foreground region.

Quantitative results for the semantic segmentation network. Table 2 shows the results of training the semantic segmentation network with the pseudo-mask obtained by applying the proposed augmentation method to PSA [2], IRN [1], and AMN [25] and the evaluated mIoU value for the Pascal VOC 2012 *val* and *test* sets. As a result of applying SMA, PSA improved by 4.2%p and 3.1%p, and IRN improved by 5.1%p and 3.9%p for the *val* and *test* sets. In addition, when our proposed method was employed on AMN, mIoU values of 70.9% and 70.8% were achieved for the *val* and *test* sets, respectively. These observations confirm that our proposed method can be effectively applied to the previous WSSS methods and can notably boost their semantic segmentation performance. We adopted the MS COCO 2014 dataset, which is a large-scale benchmark with a large number of classes to demonstrate the generalization ability and robustness of our proposed method. Table 3 shows the

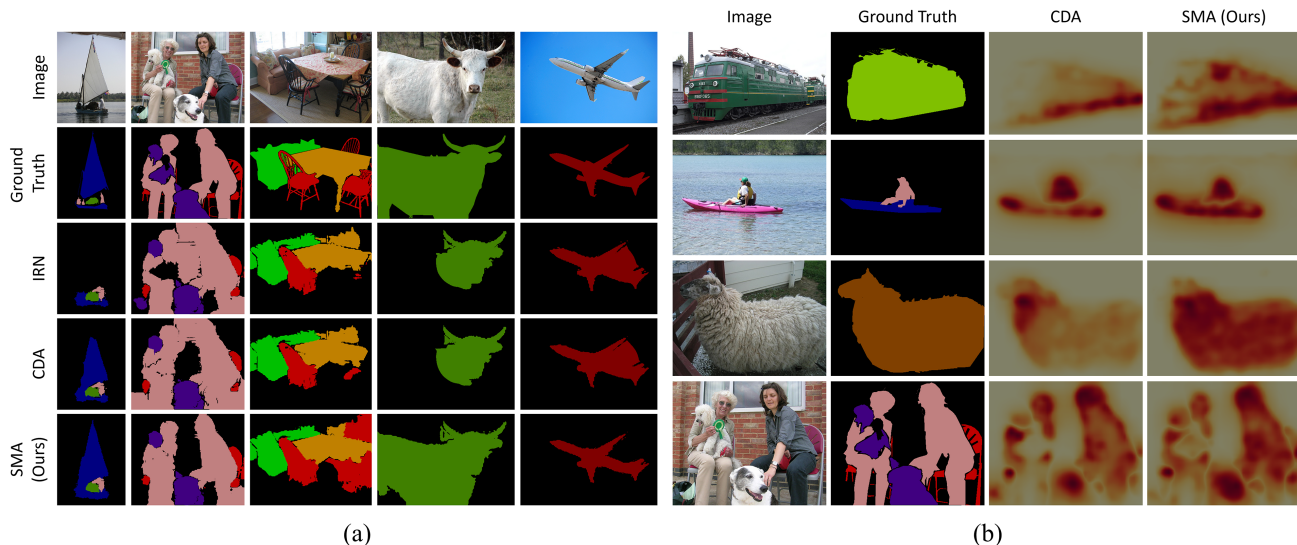


Figure 5. Qualitative examples of (a) pseudo-masks from IRN [1], CDA [33] and SMA(Ours) on PASCAL VOC 2012 *val* images (b) localization maps which are obtained by applying CDA [33] and SMA (Ours).

results for the MS COCO validation set. When SMA was employed on AMN, it achieves an improvement of 1.2%p compared with AMN, and outperforms the other competitive methods. Qualitative results for MS COCO 2014 are in supplementary material.

Comparison with other augmentation methods. Table 4 compares the performance of the localization maps obtained by applying different augmentations including Manifold mixup [37], Cutout [9], Mixup [44], CutMix [43], and CDA [33]. The experimental results show that compared with the other augmentations, including the ones used in the image classification and WSSS tasks, our proposed method showed the highest mIoU value. Evidently, when classical augmentation was used, the classifier becomes confused between the object-relevant and background cue. In addition, the result was 1.6%p higher than that obtained using the CDA [33]. Figure 5 (b) shows localization maps generated by CDA and our proposed method.

4.3. Analysis and discussions

Ablations for loss terms on localization maps. We performed an ablation study for each loss term as shown in Table 5, where baseline denotes the baseline loss without the proposed augmentation method. $\mathcal{L}_{shuffle}^{\dagger}$ and $\mathcal{L}_{shuffle}^{\ddagger}$ are the first and second terms of $\mathcal{L}_{shuffle}$ and, indicate the losses corresponding to the background and object relevant synthesizing, respectively. When the features were disentangled by adding \mathcal{L}_{cls} and \mathcal{L}_{contr} , the performance improved by 1.5%p compared to that obtained using the baseline, implying that the classifier focuses more on the target object region. As a result of randomly shuffling the background representation by adding $\mathcal{L}_{shuffle}^{\dagger}$, the performance

Method	<i>val</i>
ADL [37] _{TPAMI'20}	30.8
CONTA [37] _{NeurIPS'20}	33.4
CDA [43] _{ICCV'21}	33.7
IRN [9] _{CVPR'19}	41.4
RIB [44] _{CVPR'21}	43.8
AMN [25] _{CVPR'22}	44.7
AMN [25] + SMA (Ours)	45.9

Table 3. We evaluated semantic segmentation results (mIoU) on MS COCO 2014 *val* set.

Method	Seed
w/o augmentation	48.3
Manifold mixup [37] _{ICML'19}	48.7
Cutout [9] _{preprint'17}	48.9
Mixup [44] _{ICLR'18}	49.0
CutMix [43] _{ICCV'19}	49.2
CDA [33] _{ICCV'21}	50.8
SMA (Ours)	52.4

Table 4. We compare localization map performance of various augmentation methods on Pascal VOC 2012 *train* set.

increased by 1.1%p, compared to that of \mathcal{L}_{contr} . Furthermore, when object relevant representation was shuffled by adding the $\mathcal{L}_{shuffle}^{\dagger}$ term, an mIoU of 52.4% was achieved. These experimental results show that SMA effectively diversifies the representations and suppress the use of the shortcut features.

Loss	(a)	(b)	(c)	(d)	(e)	Seed
baseline	✓					48.3
\mathcal{L}_{cls}	✓	✓				49.2
\mathcal{L}_{contr}	✓	✓	✓			49.8
$\mathcal{L}_{shuffle}^\dagger$	✓	✓	✓	✓		50.9
$\mathcal{L}_{shuffle}^\ddagger$	✓	✓	✓	✓	✓	52.4

Table 5. Effectiveness of each loss term on localization map on Pascal VOC 2012 *train* set.

Class (w/ background)	Co-occurrence ratio	Method	
		IRN [1]	SMA
airplane (w/ sky)	0.23	83.72	87.34
sheep (w/ grass)	0.22	85.95	86.10
cow (w/ grass)	0.20	86.24	87.79
boat (w/ water)	0.18	75.05	76.36
train (w/ track)	0.11	68.83	71.69

Table 6. We evaluated mIoU (%) of localization map for images of object-background pairs that frequently appear in the Pascal VOC 2012 training dataset to analyze spurious correlations.

Analysis for spurious correlations. The mIoU values presented in Table 6 were obtained for bias-aligned images, which corresponds to frequently appearing object classes and background pairs. We used images with the same object-background combination used in section 3.2. The co-occurrence ratio denotes the ratio in which the corresponding background appears among the images corresponding to a specific class label. For example, in the entire airplane image, the sky coincides by more than 23%. The removal of harmful correlations was verified by evaluating the performance of the images of the above-mentioned combinations. All the combinations showed results that were higher than those obtained using the IRN [1]. Thus, it can be confirmed that our proposed method can effectively alleviate the high usage of shortcuts caused by a biased dataset. A more experiment on bias-conflicting samples is provided in supplementary material.

Manifold visualization. We performed manifold visualization using T-SNE [36], which is a dimensionality reduction method, to evaluate the utility of the disentangled object-relevant and background representations. We used the intermediate features z^o and z^b obtained using the aggregators M_o and M_b . Figure 6 reveals that the features related to the foreground and background are semantically different.

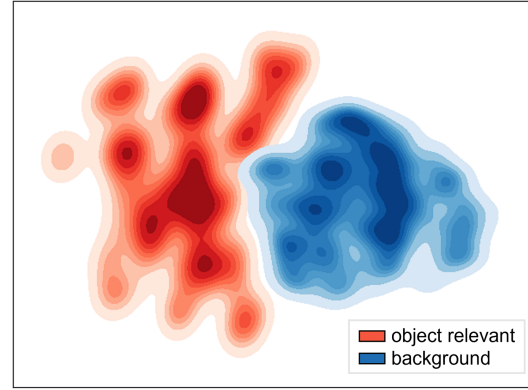


Figure 6. Visualization of the features corresponding to the object relevant and background by using the T-SNE [36] dimension reduction method.

5. Conclusion

In this study, we propose SMA that disentangles the object-relevant and background representations and shuffles them in mini-batches to create synthetic feature. Evidently, the classifier trained with SMA effectively mitigates the use of shortcuts by focusing on the object itself to perform the predictions. Thus, the performance of the localization map and pseudo-mask obtained through the classifier could be improved by applying our proposed augmentation method. Our method improved localization map and pseudo-mask obtained from a classifier and exhibit increased semantic segmentation performance on Pascal VOC 2012 and MS COCO 2014 datasets. In the future, we intend to combine representations based on distance to allow the classifier to leverage more information related to objects for improving the performance of localization maps.

Limitations. Our proposed SMA involves a two-step process, which requires additional training iterations. Furthermore, while SMA to some extent separates the two representations, it faces challenges in effectively decoupling hard attributes. Despite these limitations, we effectively trained the classifier to detour shortcuts and focus on the target.

6. Acknowledgements

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2022R1C1C1008534), and Institute for Information & communications Technology Planning & Evaluation (IITP) through the Korea government (MSIT) under Grant No. 2021-0-01341 (Artificial Intelligence Graduate School Program, Chung-Ang University).

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. 6, 7, 8
- [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. 6
- [3] Nikita Araslanov and Stefan Roth. Single-stage semantic segmentation from image labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4253–4262, 2020. 3
- [4] 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. 1
- [5] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018. 6
- [6] Chengtai Cao, Fan Zhou, Yurou Dai, and Jianping Wang. A survey of mix-based data augmentation: Taxonomy, methods, applications, and explainability. *arXiv preprint arXiv:2212.10888*, 2022. 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*, pages 4288–4298, 2022. 2, 3
- [8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5, 6
- [9] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 2, 3, 7
- [10] 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. 5, 6
- [11] Junsong Fan, Zhaoxiang Zhang, Chunfeng Song, and Tieniu Tan. Learning integral objects with intra-class discriminator for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4283–4292, 2020. 2
- [12] Junsong Fan, Zhaoxiang Zhang, Tieniu Tan, Chunfeng Song, and Jun Xiao. Cian: Cross-image affinity net for weakly supervised semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10762–10769, 2020. 2
- [13] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. 1
- [14] 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
- [15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Computer Science*, 2015. 5, 6
- [16] 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. 1
- [17] Beomyoung Kim, Sangeun Han, and Junmo Kim. Discriminative region suppression for weakly-supervised semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1754–1761, 2021. 2
- [18] Beomyoung Kim, Youngjoon Yoo, Chae Eun Rhee, and Junmo Kim. Beyond semantic to instance segmentation: Weakly-supervised instance segmentation via semantic knowledge transfer and self-refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4278–4287, 2022. 1
- [19] Namyup Kim, Taeyoung Son, Jaehyun Pakh, Cuiling Lan, Wenjun Zeng, and Suha Kwak. Wedge: web-image assisted domain generalization for semantic segmentation. In *ICRA*, 2023. 3
- [20] Jungbeom Lee, Jooyoung Choi, Jisoo Mok, and Sungroh Yoon. Reducing information bottleneck for weakly supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 34:27408–27421, 2021. 1, 3
- [21] 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
- [22] 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. 1, 2, 6
- [23] Jungbeom Lee, Seong Joon Oh, Sangdoon Yun, Junsuk Choe, Eunji Kim, and Sungroh Yoon. Weakly supervised semantic segmentation using out-of-distribution data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16897–16906, 2022. 2, 3
- [24] 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. 1
- [25] Minhyun Lee, Dongseob Kim, and Hyunjung Shim. Threshold matters in wsss: Manipulating the activation for the robust and accurate segmentation model against thresholds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4330–4339, 2022. 1, 2, 6, 7

- [26] Seungho Lee, Minhyun Lee, Jongwuk Lee, and Hyunjung Shim. Railroad is not a train: Saliency as pseudo-pixel supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5495–5505, 2021. [2](#)
- [27] Soon Hoe Lim, N Benjamin Erichson, Francisco Utrera, Winnie Xu, and Michael W Mahoney. Noisy feature mixup. *arXiv preprint arXiv:2110.02180*, 2021. [3](#)
- [28] 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](#)
- [29] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 891–898, 2014. [5](#)
- [30] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33:20673–20684, 2020. [6](#)
- [31] Krishna Kumar Singh, Hao Yu, Aron Sarmasi, Gautam Pradeep, and Yong Jae Lee. Hide-and-seek: A data augmentation technique for weakly-supervised localization and beyond. *arXiv preprint arXiv:1811.02545*, 2018. [3](#)
- [32] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3136–3145, 2019. [1](#)
- [33] 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*, pages 7004–7014, 2021. [3](#), [5](#), [6](#), [7](#)
- [34] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017. [5](#)
- [35] Meng Tang, Abdelaziz Djelouah, Federico Perazzi, Yuri Boykov, and Christopher Schroers. Normalized cut loss for weakly-supervised cnn segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1818–1827, 2018. [1](#)
- [36] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. [8](#)
- [37] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *International Conference on Machine Learning*, pages 6438–6447. PMLR, 2019. [3](#), [7](#)
- [38] 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. [2](#)
- [39] 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](#)
- [40] Yunchao Wei, Huaxin Xiao, Honghui Shi, Zequn Jie, Jiashi Feng, and Thomas S Huang. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7268–7277, 2018. [2](#)
- [41] Jinheng Xie, Jianfeng Xiang, Junliang Chen, Xianxu Hou, Xiaodong Zhao, and Linlin Shen. C2am: contrastive learning of class-agnostic activation map for weakly supervised object localization and semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 989–998, 2022. [2](#)
- [42] Qi Yao and Xiaojin Gong. Saliency guided self-attention network for weakly and semi-supervised semantic segmentation. *IEEE Access*, 8:14413–14423, 2020. [2](#)
- [43] Sangdoon Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. [2](#), [3](#), [5](#), [7](#)
- [44] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. [2](#), [3](#), [5](#), [7](#)
- [45] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. In *ICLR*, 2021. [3](#)
- [46] Lei Zhu, Qi She, Qian Chen, Xiangxi Meng, Mufeng Geng, Lujia Jin, Zhe Jiang, Bin Qiu, Yunfei You, Yibao Zhang, et al. Background-aware classification activation map for weakly supervised object localization. *arXiv preprint arXiv:2112.14379*, 2021. [3](#)