Cross-domain Semantic Decoupling for Weakly-Supervised Semantic Segmentation

Zaiquan Yang
zaiquanyangcat@gmail.com
Zhanghan Ke
kezhanghan@outlook.com
Gerhard Hancke
gp.hancke@cityu.edu.hk
Rynson Lau
rynson.lau@cityu.edu.hk

Department of Computer Science City University of Hong Kong, Hong Kong SAR, China

Abstract

Weakly-supervised semantic segmentation (WSSS) aims to obtain pixel-wise pseudo labels from image-level labels for segmentation supervision. However, due to the co-occurrence of multiple categories in an image, it is difficult to obtain accurate pseudo labels for supervision, leading to the unsatisfactory performances of current methods. In this paper, we observe that accurate pseudo labels are easier to obtain from images with only a single semantic object (i.e., single-label images) compared to those with multiple semantic objects (i.e., multi-label images). This inspires us to treat the localization maps from single-label images (referred to as the source domain) as good prior knowledge and transfer to multi-label images (referred to as the target domain). Specifically, we present a cross-domain semantic decoupling (CSD) method that first splits image data into source and target domains, and then utilizes the co-occurrence oriented copy-and-paste scheme to enforce pixel-wise consistency and regularize the network responses to the same objects in the two domains. Such a design reduces semantic ambiguity and generates more accurate class boundaries for the pseudo labels. Our method can be seamlessly incorporated into existing WSSS models. Extensive experiments on PASCAL-VOC 2012 demonstrate that the proposed CSD can significantly improve the quality of pseudo labels and final segmentation results.

1 Introduction

Semantic segmentation is a fundamental task in the field of computer vision. However, this task highly relies on pixel-wise ground-truth labels. Considering the huge cost of collecting pixel-level semantic labels, the weakly-supervised semantic segmentation (WSSS) task is proposed. One popular branch of WSSS is to predict pixel-wise pseudo labels from image-level class labels for supervision [1, 33], which is challenging but requires less efforts than using other forms of labels such as bounding boxes [17], scribbles [31], and points [10].

The existing methods can improve the WSSS performance notably. IRN [2] integrates the interpixel relations on the attention maps in the training. MCTformer [29] proposes

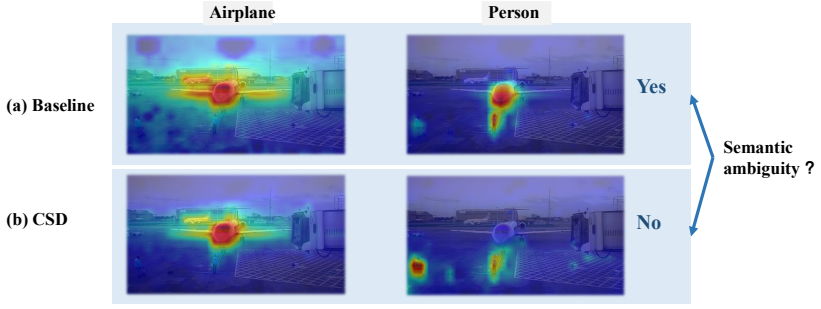

Figure 1: Illustration of the Class Activation Map (CAM) generated by a baseline model MCTformer [29] and our method CSD. (a) The MCTformer suffers from pixel-wise semantic ambiguity , *e.g.*, person (left column) or airplane (right column) when multiple categories co-occur, while (b) our method effectively reduces pixel-wise semantic ambiguity and produces more accurate class boundaries.

a transformer-based framework and considers the self-attention of tokens as the class activation map. However, obtaining precise pixel-wise labels remains challenging. Though OoD [18] proposes to obtain accurate object boundaries by collecting specific out-of-distribution data and CDA [24] utilize copy-and-paste augmentation to remove the dependence between target objects and contextual background, these methods are less efficient and also rarely consider the co-occurrence of multiple target objects (*e.g.*, airplane and person). As shown in Fig. 1(a), the co-occurrence of multiple targets can cause semantic ambiguity and even errors in object boundaries. In this work, we observe that quite a lot of multi-label images are included in the dataset of WSSS, each of which has multiple target objects and hinder the model from getting more complete semantic discrimination. As a result, semantic ambiguity often occurs in the pseudo mask of multi-label images. On the contrary, the complete pseudo mask can be obtained easily from single-label images with only a single target object. More details can be found in the supplementary materials.

Based on the above observation, we propose a novel cross-domain semantic decoupling (CSD) method. Our insight is that the high-quality pseudo-labels from the single-label images can be leveraged as good prior knowledge to decouple the target objects in multi-label images. The CSD first splits the image dataset into single-label and multi-label image subsets. The single-label image subset, which is also referred to as the *source domain*, is composed of images with each of them containing only a single object class, while the multi-label image subset, referred to as the *target domain*, is composed of images with each of them containing more than one object class. Inspired by augmentation-based consistency learning in semi-supervised learning [23], we regard the foreground regions of the source domain as reliable prior knowledge and transfer them to the target domain with the proposed co-occurrence oriented copy-and-paste augmentation scheme. By enhancing the pixel-wise semantic consistency, we regularize the network responses to the same object regions in the two domains. Furthermore, to balance the activation map between foreground and background, we also utilize the background regions for consistency learning. As shown in Fig. 1(b), our method obviously alleviates the semantic ambiguity in multiple target object classes.

To sum up, the main contributions of this work are three-fold: (1) We demonstrate that the co-occurrence of multiple target classes resulted in semantic ambiguity and propose to

decouple the multiple target classes from a cross-domain perspective. (2) We introduce a co-occurrence oriented copy-paste scheme for transferring the prior knowledge. To balance the activation of foreground and background, we propose dual semantic consistency learning framework which respectively considers the localization map of foreground and background as prior knowledge. (3) We conduct experiments on PASCAL-VOC 2012 benchmark to demonstrate that our method can be a plug-in to boost various popular WSSS methods.

2 Related Work

2.1 Weakly-supervised Semantic Segmentation

Prevailing WSSS methods with image-level labels commonly adopt a multi-stage framework [4, 19, 29] which firstly generate class activation map (CAM) as pseudo-masks from a classification network and then use the pseudo-masks as supervision to train a segmentation network. Partial activation is one of the critical problems, [14, 32] proposes to use erasing strategies to discover more areas. In addition, [29] propose a multi-class token transformer, which uses multiple class tokens to learn interactions between the class and the patch tokens as the alternative of CAM. Besides, some recent work focus on the co-occurrence of target objects and background objects. [19, 28] utilize both the off-the-shelf saliency map to overcome the challenge. [24] present an augmentation-based method to change the inherent context and decouple between the object instances and contextual information. Different from these prior works, we pay more attention to the coupling problem between target categories. The collected image data often contains multiple target categories, which cause ambiguous boundaries between target objects and even mistakenly recognize the object together with other objects.

2.2 Consistency Regularization

Consistency regularization [3, 15] is an important idea in semi-supervised learning (SSL). It utilizes unlabeled data by relying on the assumption that the model should output similar predictions when fed perturbed versions of the same image. Typically, the FixMatch [23] leverages weak and strong two kinds of augmentations on the same inputs and then train the model to make its prediction on the strongly-augmented version match the weak version. In this paper, devoted to mitigating the semantic confusion caused by co-occurrence, we regard the input from the single-label domain as the weak-augmented version. Correspondingly, we obtain strongly-augmented versions by copying and pasting the input into multi-label images. By encouraging the model to obtain consistent predictions for the same image regions but in different domains, we enhance the pseudo label's ambiguity of co-occurrence semantic categories.

2.3 Copy-and-Paste Augmentation

Copy-paste augmentation is first presented in [7]. By cutting object patches from the source image and pasting to the target one, a combinatorial amount of synthetic training data can be acquired and improve the detection/segmentation performance. The method [9] has further polished the augmentation by context adaptation. It is then verified by subsequent works from many fields [6, 11, 13]. In the prior WSSS work, [24] also employ copy-and-paste

for changing the context and decoupling the target objects and non-target context objects. Differently, we aim to decouple the target objects existing together in one image. In the implementation, we impose the copy-and-paste on the input from the single-label domain for strongly-augmented version. Besides, we focus on pixel-wise consistency regularization instead of the global classification and we think it helps narrow the gap between segmentation and classification.

3 Proposed Method

3.1 Framework

We focus on alleviating the semantic ambiguity caused by co-occurrence between multiple target objects. As a result, we propose a new framework for Weakly-supervised semantic segmentation, called Cross-domain Semantic Decoupling (CSD). Considering the image data with only a single label is little ambiguous when predicting pixel-wise pseudo labels, we split the whole dataset as single-label images (source domain) and multi-label images (target domain) as shown in Fig. 2. By leveraging the copy-and-paste based augmentation, we transfer the prior knowledge from simple source domain to complex target domain. In the following, we will describe implementations in detail.

3.2 Cross-domain Semantic Decoupling

Prior Localization Map Obtaining. As previously described, our aim is to transfer accurate prior knowledge of simple domain to complex target domain. Therefore, our approach consists of two stages like [24]. The first stage is to obtain the localization map of single-label images by using off-the-shelf WSSS methods. Then in the second stage, we utilize these relatively complete localization map as pseudo mask to implement blending with multi-label domain data and train the classification network in a online manner.

Co-occurrence Oriented Resampling. A plain paste copy can effectively change the context of the original data, but it is not sufficient for our purposes. In fact, the frequency of co-occurrence is different for different combinatorial categories. Taking the “horse" as an example, it often appears together with “person", but the probability of appearing together with “airplane" is low. As a result, semantic decoupling is less effective if samples from different categories are sampled equally for copy-and-paste based augmentation. In this paper, we propose co-occurrence oriented resampling strategy.

We first count the prior correlation matrix $\mathcal{P}$ according to the co-occurrence of different categories as follow:

$$\mathcal{P}_{i,j} = \frac{\mathcal{N}_{i,j}}{\sum_{k=1}^{C} \mathcal{N}_{i,k}}, \tag{1}$$

where C is the number of classes in whole dataset. $\mathcal{N}_{i,j}$ denotes the number of samples with i_{th} class and j_{th} class. The $\mathcal{P}_{i,j}$ mean the probability of j_{th} class occurring when class i_{th} is present. Given the input sampled from multi-label images, we then decide which classes to sample for the copy-and-paste augmentation. Specifically, for the input with classes $\{c_1, c_2, \cdots, c_n\}$ and we would consider $\mathcal{P}_j = \sum_{k=1}^{n} \mathcal{P}_{k,j}$ as the probability to sample j_{th} class. Empirically when the category co-occurs with other categories more frequently or has more single-label domain samples, it will be more likely to be sampled.

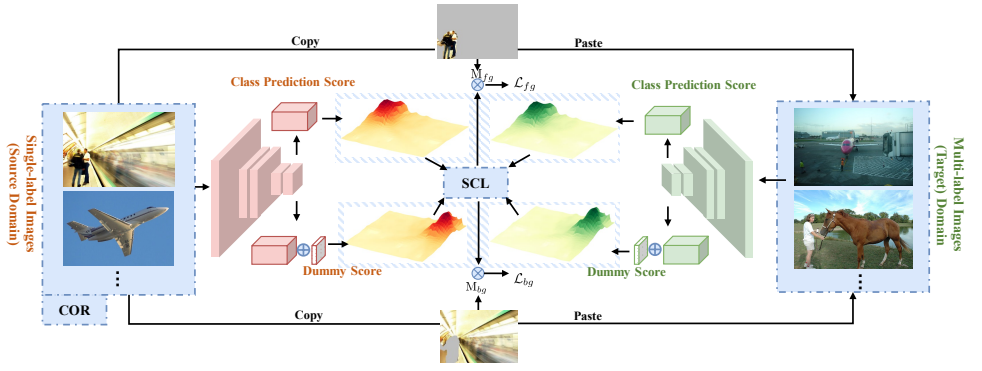

Figure 2: Overview of the proposed framework. Given a multi-label image, the single-label image is sampled according to the proposed co-occurrence oriented resampling (COR) scheme for copy-and-paste. We predict the class activation score by the classification network, which is shared between the source domain and target domain, and then implement semantic consistency learning (SCL) respectively on foreground and background regions.

Augmentation-based Semantic Consistency Learning. According to the above sampling strategy, given the input multi-label image x_m we will selectively sample a single-label image x_s. Then we employ the random blending method to paste the foreground objects of single-label image into the multi-label image. We denote the new multi-label image augmented by foreground objects as x_{sm}^{fg}. As a results, for the same foreground objects, we acquire two views in different contexts. Inspired by augmentation-based consistency learning in semi-supervised learning [23], we train the model to make its prediction on the complex view match the reliable pseudo-label from simple view.

Specifically, we denotes the model function as $\mathcal{F}(\cdot)$. Then the last output feature maps for the simple view and corresponding complex view are considered as class prediction score as follow:

$$\begin{aligned} \mathrm{M}_s &= \mathcal{F}(x_s), \quad \mathrm{M}_s \in \mathbb{R}^{C\times H\times W}, \\ \mathrm{M}_{sm}^{fg} &= \mathcal{F}(x_{sm}^{fg}), \quad \mathrm{M}_{sm}^{fg} \in \mathbb{R}^{C\times H\times W}. \end{aligned} \tag{2}$$

Considering that there are some false predictions in foreground regions, it is not optimal to optimize the cross-entropy loss directly. Here we consider the M_s as the pixel-wise pseudo label distribution and impose the kullback-leibler divergence loss on the foreground regions:

$$\mathcal{L}_{fg} = \sum_{h=1}^{H}\sum_{w=1}^{W} \mathbb{I}_{(h,w)\in \mathrm{M}_{fg}} \mathrm{KL}(\mathrm{M}_s || \mathrm{M}_{sm}^{fg}), \tag{3}$$

where the M_{fg} denotes the foreground object mask of single-label image which is obtained during object instances collecting. It is noted that we do not perform additional post-processing on the predictions like min-max normalization which is commonly used for the class activation map.

Dual Semantic Consistency Learning. However, with only foreground semantic consistency learning, the model would lead to over-activation especially on the background area. To prevent the problem, we also implement similar consistency learning for the reliable

Method	Backbone	Seed	Mask	Val	Test
PSA [1] CVPR'2018	ResNet38	48.0	61.0	61.7	63.7
SEAM [25] CVPR'2020	ResNet38	55.4	63.6	64.5	65.7
CONTA [30] NeurIPS'2020	ResNet38	56.2	66.1	66.1	66.7
EDAM [26] CVPR'2021	ResNet101	52.8	68.1	70.9	70.6
AdvCAM [16] CVPR'2021	ResNet38	55.6	68.0	68.1	68.0
CDA [24] ICCV'2021	ResNet38	55.4	63.4	66.1	66.8
SIPE [5] CVPR'2022	ResNet101	58.6	69.3	68.8	69.7
CLIMS [27] CVPR'2022	ResNet101	56.6	70.5	70.4	70.0
ViT-PCM [22] ECCV'2022	ResNet101	63.6	67.1	70.3	70.9
Improvement over baseline:					
IRN [2] CVPR'2019	ResNet50	48.3	66.5	63.5	64.8
IRN **w/CSD** BMVC'2023	ResNet50	**50.6** ↑ 2.3	**68.2** ↑ 1.7	**64.9** ↑ 1.4	**66.1** ↑ 1.3 [1]
MCTformer [29] CVPR'2022	ResNet38	61.7	69.1	70.0	71.6
MCTformer **w/CSD** BMVC'2023	ResNet38	**63.8** ↑ 2.1	**70.9** ↑ 1.8	**71.4** ↑ 1.4	**72.8** ↑ 1.2 [2]

Table 1: Evaluation (mIoU (%)) of the initial seed (Seed), the refined pseudo mask (Mask), and the segmentation mask [1] on PASCAL VOC 2012 val and test set.

background part. We paste the background objects of single-label image into the multi-label image and denote the new multi-label image augmented by the background part as x_{sm}^{bg}. Since the prediction of background categories is absent in the prediction results of the model, we introduce constant values M' as the dummy score to enable the optimization as follows:

$$\begin{aligned} \mathrm{M}_s' &= [\mathrm{M}',\ \sigma(\mathcal{F}(x_s))], \quad \mathrm{M}_s' \in \mathbb{R}^{2\times H\times W}, \\ \mathrm{M}_{sm}^{bg} &= [\mathrm{M}',\ \sigma(\mathcal{F}(x_{sm}^{bg}))], \quad \mathrm{M}_{sm}^{bg} \in \mathbb{R}^{2\times H\times W}, \end{aligned} \tag{4}$$

where σ denotes the sum operation along the classes channels and results a prediction score about saliency. Similarly, we can get the background-aware kullback-leibler divergence loss:

$$\mathcal{L}_{bg} = \sum_{h=1}^{H}\sum_{w=1}^{W} \mathbb{I}_{(h,w)\in \mathrm{M}_{bg}} \mathrm{KL}(\mathrm{M}_s || \mathrm{M}_{sm}^{fg}), \tag{5}$$

where the M_{bg} denotes the background object mask of single-label image which is obtained during object instances collecting.

Loss Function. Our loss includes two parts: classification loss and semantic decoupling loss. The latter further include the foreground consistency learning loss and the background one. The total loss is formulate as:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda_1 * \mathcal{L}_{fg} + \lambda_2 * \mathcal{L}_{bg}. \tag{6}$$

where λ_1 and λ_2 are the factor for balancing the activation of background and foreground.

In general, our CSD can effectively alleviate the pixel-wise coupling problem between all target categories without any extra data. In prior work, CDA [24] also leverage the copy-and-paste for decoupling the high correlation between objects and their contextual background. AttBN [20] transfers the foreground prior from a simple single-label dataset to another complex multi-label dataset by adversarial learning [21]. However, they still cannot further narrow the gap between classification and segmentation tasks from the pixel level.

[1] http://host.robots.ox.ac.uk:8080/anonymous/74JEAT.html
[2] http://host.robots.ox.ac.uk:8080/anonymous/UYDBF7.html

4 Experiments

4.1 Experimental Settings

Datasets. We evaluate our proposed method on PASCAL VOC 2012 segmentation dataset [8], one of the standard benchmarks for WSSS. The dataset consists of 21 classes including a background, with 1,464, 1,449, and 1,456 images for train, validation, and test set, respectively. Following the common practice in semantic segmentation community, we also use the augmented train set that consists of 10,582 images [12] for training. We report the mean Intersection-over-Union (mIoU) for evaluation, and the mIoU on the VOC test set is obtained from the official evaluation server.

Baselines. We choose two popular models, IRN [2] and MCTformer[29] as our baselines. IRN integrates the interpixel relations on the attention maps in the training, which achieves outstanding performance. MCTformer proposes a ViT-based multiple class tokens learning framework and considers the self-attention of tokens as the activation map. Built on top of these models, we evaluate the effectiveness of our proposed CSD.

Implementation details. The general training pipline includes multi-label image classification, a pseudo-mask generation, and the final segmentation training three stages. We strictly follow the same settings (e.g., image augmentation) as reported in the official codes. When imposing our proposed contrast to MCTformer and IRN, we set $\lambda_1 = 0.01$ and $\lambda_2 = 0.1$ in order to keep balance with classification loss. As for the training epoch, learning rate, learning rate decay policy, weight decay rate, and optimizer, we follow the same setting as MCTformer and IRN. More details can be found in the supplementary materials.

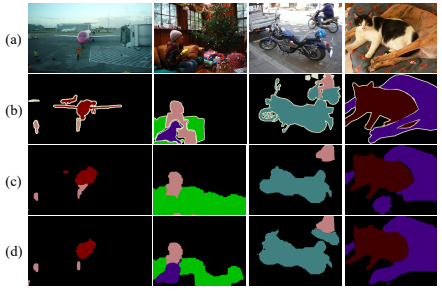

Figure 3: The visualization of pseudo-masks on PASCAL-VOC. (a) Input images. (b) GT masks. (c) Pseudo masks produced by baseline model MCTformer [29]. (d) Pseudo masks generated by our method. Our method can obtain pseudo masks with better boundaries and completeness.

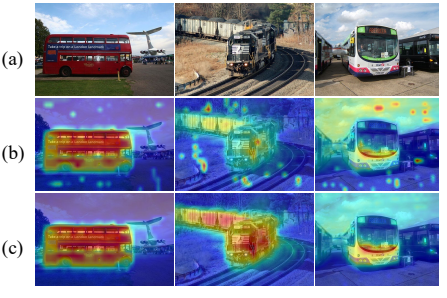

Figure 4: The class activation map (CAM) comparison. (a) Input images. (b) The CAM when seting $\lambda_1 = 0.01, \lambda_2 = 0.0$. (c) The CAM when setting $\lambda_1 = 0.01, \lambda_2 = 0.1$.

4.2 Comparison with State-of-the-arts

Accuracy of seed and pseudo-mask. To verify the effectiveness of our CSD, we evaluate CAM seed regions and pseudo-masks respectively. The seed are obtained by applying a range of thresholds to separate the foregrounds and backgrounds in the raw CAMs. As

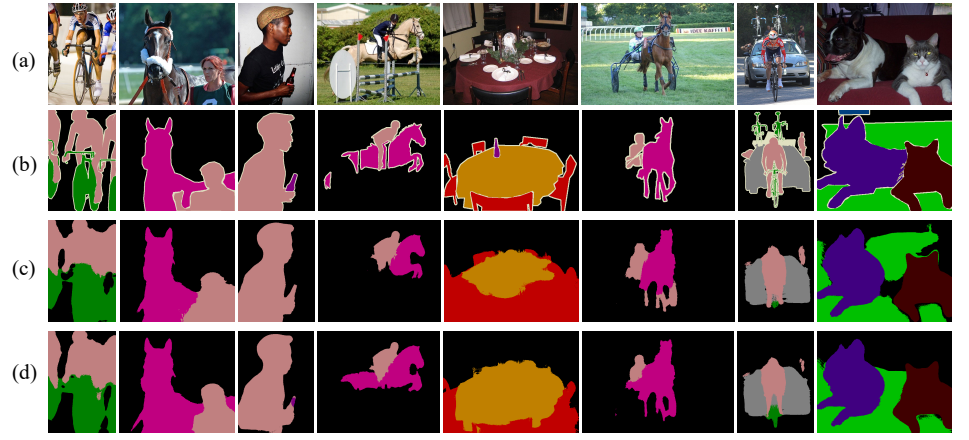

Figure 5: Qualitative results on PASCAL-VOC. (a) Input images. (b) GT masks. (c) Segmentation results produced by baseline model MCTformer [29]. (d) Segmentation results generated by our method.

for the pseudo-masks, we are in line with the baseline models MCTformer and IRN. For MCTformer, we the PSA [1] for refinement. For the IRN, we use the proposed interpixel relations [2] for refinement. The results are shown in the Tab. 1. As can be seen, built on the strong baseline models, our CSD still improve the MCTformer by 2.1% and 1.8% mIoU on the initial seed and pseudo masks which show the superior performance. When built on the IRN baseline model, our CSD also achives improvement with 2.2%, 1.7% on the initial seed and pseudo masks. In Fig. 3, we visualize the pseudo mask. Our CSD can accurately distinguish the categories of co-occurrence and localize the precise object parts. It indicates that the pseudo mask by our method not only can improve the completeness of the objects but also the boundary between different objects. Furthermore, compared to the recent counterparts CDA [24], SIPE[5], AdvCAM [16], our method also outperforms them by large margins.

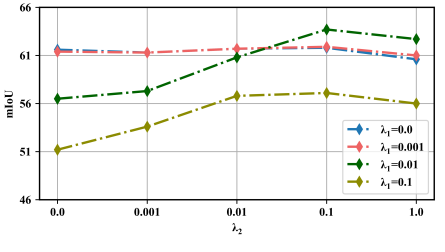

Figure 6: The performance (mIoU (%)) influence of the loss coefficient λ_1 and λ_2 for penalizing loss.

Copy-and-paste	Resampling	mIoU (%)
×	×	61.7
✓	×	63.1
✓	✓	63.8

Table 2: The ablation study of the different components.

Accuracy of segmentation map. To make a fair comparison, we report Deeplab-based segmentation performance which is in line with the baseline methods. For the MCTformer, we reproduce the segmentation performance and achieve the 70.0%, 71.6% mIoU on the

PASCAL-VOC val and test set. When equipped with our method, the MCTformer are further imporved and achives 71.4% and 72.8% mIoU on the PASCAL-VOC val and test set, which obtain the state-of-the-art performance In addition, we also avluate the performance of the IRN integrated with our method. We also achieve better segmentation performance. We also present some segmentation results in the Fig. 5. From the results, we can find that our method can clearly improve the semantic boundary not only the simple but also the complex scenes with co-occurrence of multiple objects. We stress that though we propose utilizing the simple single-label images as the source of prior knowledge, our method still show non-trivial improvement on the complex multi-label images. It indicates that our method do effectively transfer the advantageous prior knowledge from the single-label images (source domain) to the multi-label images (target domain), which confirm the motivation described in the previous section.

4.3 Ablation Studies

To analyze how each component in our proposed method helps to improve WSSS, we present extensive ablation studies in this section. Here, all experiments are done with MCTformer on PASCAL VOC 2012 dataset.

Effectiveness of each component. In this section, we demonstrate the effectiveness of each component. The results are shown in Fig. 2. First, we only impose the proposed dual semantic consistency learning which employs a plain copy-and-paste scheme. We improve the mIoU of seed from 61.7% to 63.1%. Then we carefully choose the source domain sample by the co-occurrence oriented resampling strategy. We further improve the mIou of the initial seed. The results validate the effectiveness of our proposed components.

Coefficient λ_1 and λ_2 of loss. In our work, only utilizing foreground-based copy-and-paste scheme for semantic consistency learning even causes a performance drop. We explain that when imposing the semantic consistency learning with background class absent there will be more background areas activated by foreground categories as shown in Fig. 4. To contend with the phenomenon of over-activation, it is essential to impose similar semantic consistency learning based on the background prior knowledge. To balance the two loss, we empirically choose λ_1 in set $\{0.0, 0.001, 0.01, 0.1\}$ and λ_2 in set $\{0.0, 0.001, 0.01, 0.1, 1.0\}$. The ablation results are shown in Fig. 6. From the results, we can see there is no significant improvement and even lead to serious degeneration when imposing the foreground or background semantic consistency learning alone. However, when combining the two for balance, we obtain a non-trivial improvement. Specially, when setting λ_1=0.01 and λ_2=0.1, we get the optimal performance.

5 Conclusion

This work proposes CSD, a novel weakly-supervised semantic segmentation method that propose to decouple the multiple target objects in an image from a cross-domain perspective. The CSD introduces co-occurrence oriented copy-and-paste augmentation to transfer the prior knowledge from single-label images (i.e., source domain) to multi-label images (i.e., target domain). Furthermore, it also proposes the dual semantic consistency learning for balancing the activation between foreground and background. It effectively alleviates the semantic ambiguity existing in the multiple target objects and produces a more accurate semantic boundary for supervision. We conduct extensive experiments to validate that CSD

can significantly improve the baseline performance on the PASCAL-VOC dataset. In the future, we consider carrying out comprehensive verification on more datasets.

References

[1] 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] 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.

[3] Philip Bachman, Ouais Alsharif, and Doina Precup. Learning with pseudo-ensembles. *Advances in neural information processing systems*, 27, 2014.

[4] Yu-Ting Chang, Qiaosong Wang, Wei-Chih Hung, Robinson Piramuthu, Yi-Hsuan Tsai, and Ming-Hsuan Yang. Weakly-supervised semantic segmentation via sub-category exploration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8991–9000, 2020.

[5] 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.

[6] Jiangfan Deng, Dewen Fan, Xiaosong Qiu, and Feng Zhou. Improving crowded object detection via copy-paste. *arXiv preprint arXiv:2211.12110*, 2022.

[7] Debidatta Dwibedi, Ishan Misra, and Martial Hebert. Cut, paste and learn: Surprisingly easy synthesis for instance detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1301–1310, 2017.

[8] 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*, 88:303–308, 2009.

[9] Hao-Shu Fang, Jianhua Sun, Runzhong Wang, Minghao Gou, Yong-Lu Li, and Cewu Lu. Instaboost: Boosting instance segmentation via probability map guided copy-pasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 682–691, 2019.

[10] Shuyong Gao, Wei Zhang, Yan Wang, Qianyu Guo, Chenglong Zhang, Yangji He, and Wenqiang Zhang. Weakly-supervised salient object detection using point supervison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 670–678, 2022.

[11] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2918–2928, 2021.

[12] 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.

[13] Hanzhe Hu, Fangyun Wei, Han Hu, Qiwei Ye, Jinshi Cui, and Liwei Wang. Semi-supervised semantic segmentation via adaptive equalization learning. *Advances in Neural Information Processing Systems*, 34:22106–22118, 2021.

[14] Hyeokjun Kweon, Sung-Hoon Yoon, Hyeonseong Kim, Daehee Park, and Kuk-Jin Yoon. Unlocking the potential of ordinary classifier: Class-specific adversarial erasing framework for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6994–7003, 2021.

[15] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.

[16] 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.

[17] 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.

[18] Jungbeom Lee, Seong Joon Oh, Sangdoo 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.

[19] 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.

[20] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Attention bridging network for knowledge transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5198–5207, 2019.

[21] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[22] Simone Rossetti, Dam Zappia, Marta Sanzari, Marco Schaerf, and Fiora Pirri. Max pooling with vision transformers reconciles class and shape in weakly supervised semantic segmentation supplementary material.

[23] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

[24] 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.

[25] 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.

[26] Tong Wu, Junshi Huang, Guangyu Gao, Xiaoming Wei, Xiaolin Wei, Xuan Luo, and Chi Harold Liu. Embedded discriminative attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16765–16774, 2021.

[27] Jinheng Xie, Xianxu Hou, Kai Ye, and Linlin Shen. Clims: cross language image matching for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4483–4492, 2022.

[28] Lian Xu, Wanli Ouyang, Mohammed Bennamoun, Farid Boussaid, Ferdous Sohel, and Dan Xu. Leveraging auxiliary tasks with affinity learning for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6984–6993, 2021.

[29] 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*, pages 4310–4319, 2022.

[30] Dong Zhang, Hanwang Zhang, Jinhui Tang, Xian-Sheng Hua, and Qianru Sun. Causal intervention for weakly-supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 33:655–666, 2020.

[31] Jing Zhang, Xin Yu, Aixuan Li, Peipei Song, Bowen Liu, and Yuchao Dai. Weakly-supervised salient object detection via scribble annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12546–12555, 2020.

[32] Tianyi Zhang, Guosheng Lin, Weide Liu, Jianfei Cai, and Alex Kot. Splitting vs. merging: Mining object regions with discrepancy and intersection loss for weakly supervised semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 663–679. Springer, 2020.

[33] 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.