Separate and Conquer: Decoupling Co-occurrence via Decomposition and Representation for Weakly Supervised Semantic Segmentation

Zhiwei Yang[1,2,3] Kexue Fu[4]
Minghong Duan[2,3] Linhao Qu[2,3] Shuo Wang[2,3]* Zhijian Song[1,2,3]*
[1]Academy for Engineering and Technology, Fudan University
[2]Digital Medical Research Center, School of Basic Medical Sciences, Fudan University
[3]Shanghai Key Laboratory of Medical Image Computing and Computer Assisted Intervention
[4]Shandong Computer Science Center (National Supercomputer Center in Jinan)

Abstract

Weakly supervised semantic segmentation (WSSS) with image-level labels aims to achieve segmentation tasks without dense annotations. However, attributed to the frequent coupling of co-occurring objects and the limited supervision from image-level labels, the challenging co-occurrence problem is widely present and leads to false activation of objects in WSSS. In this work, we devise a 'Separate and Conquer' scheme SeCo to tackle this issue from dimensions of image space and feature space. In the image space, we propose to 'separate' the co-occurring objects with image decomposition by subdividing images into patches. Importantly, we assign each patch a category tag from Class Activation Maps (CAMs), which spatially helps remove the co-context bias and guide the subsequent representation. In the feature space, we propose to 'conquer' the false activation by enhancing semantic representation with multi-granularity knowledge contrast. To this end, a dual-teacher-single-student architecture is designed and tag-guided contrast is conducted, which guarantee the correctness of knowledge and further facilitate the discrepancy among co-contexts. We streamline the multi-staged WSSS pipeline end-to-end and tackle this issue without external supervision. Extensive experiments are conducted, validating the efficiency of our method and the superiority over previous single-staged and even multi-staged competitors on PASCAL VOC and MS COCO. Code is available here.

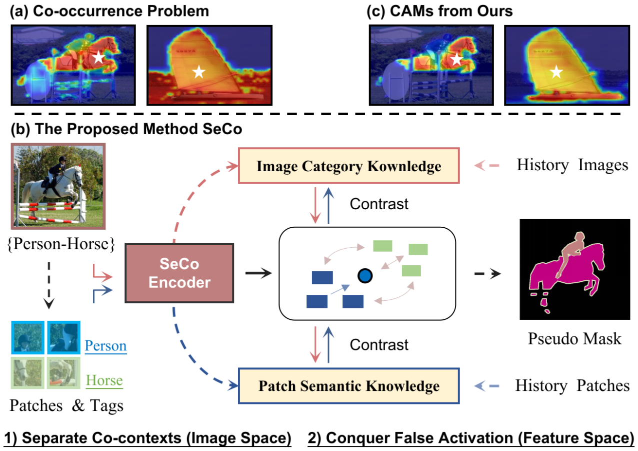

Figure 1. (a) Co-occurrence issue. Targets marked by stars (horse and boat) are falsely activated. (b) To solve this issue, we propose a single-staged framework SeCo, which acts in a 'separate and conquer' manner that efficiently tackles co-occurrence issue without external supervision. It initially separates spatial con-texts in the image space and then conquers false activation in feature space. (c) The proposed SeCo accurately localizes the co-categories.

1. Introduction

Weakly supervised semantic segmentation (WSSS) as an annotation-efficient alternative to fully supervised semantic segmentation, has enjoyed enormous popularity in recent years [50]. It aims to densely classify every pixel of an input image by only leveraging more accessible labels than pixel-wise labels, such as points [3], scribbles [25, 42], bounding boxes [11, 21], and image-level labels [1, 32]. Among these forms of annotations, image-level labels are the most economical yet challenging annotation form to accomplish the segmentation task, as they only indicate the presence of objects and contain the least semantic information.

Formally, the pipeline for WSSS with image-level labels consists of three steps, i.e., initially training a classification model to generate CAM seeds [44], then refining CAMs to generate pseudo labels, and finally training a segmentation network with the pseudo labels and taking it as the final model to inference [2, 47]. Such a WSSS paradigm can be further divided into multi-staged and single-staged groups.

*Corresponding author.

For multi-staged WSSS methods [24, 27, 46], the classification and segmentation models need to be trained progressively. It intends to have better segmentation performance while more complicated at training streamline. The single-staged methods [2, 31, 36, 37] share the encoder for classification and segmentation networks, thus can be trained end-to-end. It is more efficient to optimize while holding inferior performance to the multi-staged. In this study, we focus on the most challenging WSSS paradigm with image-level annotations and streamline the paradigm end-to-end.

For both single- and multi-staged WSSS, generating reliable CAMs from image-level labels is the first and fundamental step for the performance [19, 48]. However, since objects intend to co-occur together, such as {train, railroad}, {boat, water}, {horse, person}, etc., it is inevitable to tackle the co-occurrence of objectives in WSSS. The challenging co-occurrence problem is widespread and often leads to false activation [5, 20, 22, 46], as shown in Fig. 1 (a). Although most existing works have succeeded in completing CAMs, they barely pay attention to such issue, consequently limited to tackling the false activation and bottlenecked in WSSS performance. Recently, introducing external supervision or human prior is proposed to tackle this problem. [27, 46] leverage the vision-language matched CLIP model [33] to help distinguish among coupled contexts. [20] elaborately applies hard out-of-distribution samples to suppress spurious background cues. [39] constructs additional co-categories to address this issue. Although impressive, they heavily rely on external data or elaborate designs to tackle co-occurrence, which impedes real-world applications with complex relations among categories.

Essentially, the co-occurrence problem arises because the co-appeared contexts coupled in images confuse the networks and incur wrong semantic bias during the feature representation, resulting in false positive pixels activated with high probability. Previous methods ignore the importance of separating co-contexts before representation, thus requiring external data or designs to tackle this issue. Based on the analysis above, we argue that **initially separating the coupled objects to remove the bias and then enhancing category-specific representation provides a potential insight** to address this issue without external supervision.

As illustrated in Fig. 1 (b), we propose a single-staged WSSS framework SeCo that does not require any extra supervision to tackle the problem. **Our key insight lies in the 'separate and conquer' training scheme** that decouples co-occurrence in image space and feature space by image decomposition and representation enhancement, respectively. **(1) To separate the spatial dependence of co-contexts in image space**, we propose to decompose the integral image into multiple patches containing single category information. Previous patch-based method [17] simply gives image-level labels to patches, but the image labels cannot help differentiate the co-contexts at patch level (see Sec. 2.1). Instead, we further explore the strategy to spatially separate coupled objects and focus on tagging each patch. Specifically, category tags from CAMs are designed for each patch, which help identify the co-contexts at patch level. Since CAMs inevitably bring noise in tags, a similarity-based rectification method is designed to revise the noisy tags. In addition, a tag memory pool is constructed to store all history tags, guiding the subsequent patch representation. **(2) To conquer the co-context confusion and enhance the semantic representation in feature space**, we design a dual-teacher single-student architecture to promote the discrepancy among co-categories. We first build a global teacher to extract category knowledge from integral images. The knowledge provides class centroids for the student in patch representation and helps to push apart co-contexts. Considering the trade-offs between the separation of co-categories and the destruction on global contexts, we share the encoders of both branches to provide complementary information for patch and image semantics. In addition, patch semantic knowledge is further extracted from a patch-level semantic reservoir maintained by a local teacher. Guided by the tags from memory pool, the knowledge helps remove the bias during the representation and pushes apart co-contexts while pulling together those within the same category at a fine-grained level. **(3) Along with the category tags and the extracted knowledge, multi-granularity contrast is further proposed** across the whole dataset to decouple the co-contexts deeply.

Extensive experiments are conducted on PASCAL VOC and MS COCO, validating its effectiveness in tackling co-occurrence (as shown in Fig. 1 (c)) and the superiority over previous single-staged and multi-staged competitors.

2. Related Works

2.1. Learning from Local Semantics

The image-level labels provide limited supervision to generate high-quality CAMs, which motivates many researches to dig complementary information from local semantics. Decomposing a whole image into local patches offers a practical implementation. L2G [17] crops patches to mine different views of semantics and generates more complete CAMs. PPL [23] utilizes feature patches to explore scattered local details and generates CAMs to cover the whole object. ToCo [37] extracts local semantics from patches and activates more non-discriminative areas. However, simply giving image-level labels to each patch like L2G cannot identify co-categories at patch level, or grouping them into foregrounds and backgrounds as ToCo fails to tell co-categories in the foregrounds. Both of them aggravate the unexpected co-occurring bias during patch representation and suffer from false activation. To this end, we

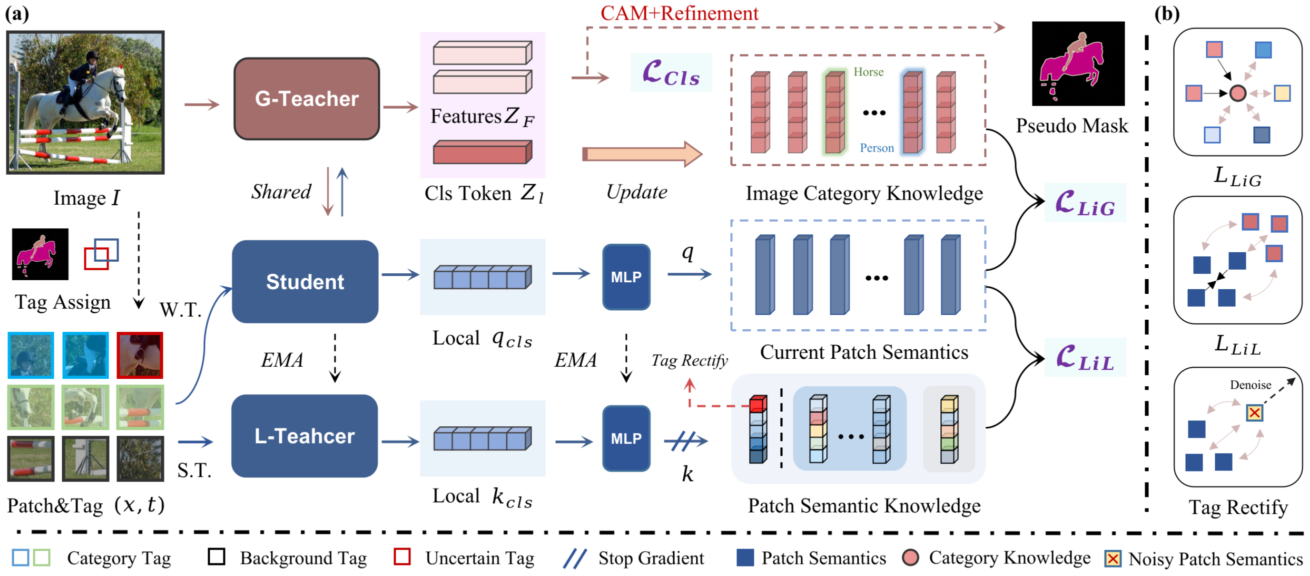

Figure 2. (a) Architecture of the proposed SeCo to tackle co-occurrence issue. Specifically, integral images are firstly sent to the global teacher (G-Teacher) to extract the category knowledge and CAMs. Then three types of category tags, i.e., single-category, background and uncertain tags, are generated from CAMs and allocated to patches accordingly. With tags, two views of patches by different augmentations, i.e., weak data augmentation (W.T.) and strong augmentation (S.T.), are sent to the student and local teacher (L-Teacher) branch, respectively. The local teacher stores all the history patches and category tags and generalizes the patch semantic knowledge. Finally, two contrastive losses, L_{LiG} and L_{LiL}, are conducted to guarantee the decoupling. In addition, CAMs from global teacher are refined as pseudo labels to train the segmentation network. Since the encoders of segmentation model and classification model are shared, our WSSS pipeline can be trained end-to-end. (b) Illustrated essence of the key components in SeCo. More details are introduced in Sec. 3.2

observe that rich semantics from CAMs can be potentially used to tag each patch, which helps differentiate co-contexts and guide category representation at a patch level. To the best of our knowledge, we are the first to report that assigning class-specific tags to patches effectively separates co-categories and reduces the bias during the representation.

2.2. Contrastive Learning & Knowledge Distillation

Contrastive learning (CL) [9, 15, 43] and knowledge distillation (KD) [4, 16] are two prevalent techniques to promote feature representation. Inspired by it, RCA [52] and ToCo [37] conduct region-level contrast to focus on non-discriminative pixels. PPC [13] leverages prototype-based [51] contrast to expand CAMs. Apart from incorporating CL, SCD [49] and L2G [17] introduce knowledge distillation into WSSS and succeed in completing CAMs. However, since the co-occurring pixels intend to be falsely activated with high confidence, simply conducting CL or KD without separating co-contexts could bring much noise to the feature representation, consequently limited to suppressing false activation and tackling co-occurrence. Different from previous works, we focus on reducing noise during feature representation from a new perspective of a 'separate and conquer' paradigm with those techniques, and highlight removing the dependence among co-contexts from two dimensions of image space and feature space, respectively.

3. Methodology

3.1. Problem Definition

The co-occurrence problem stems from the fact that the coupling of co-contexts confuses networks and introduces noise during feature representation, leading to false positive pixels still activated with high confidence. Therefore, the key of SeCo to addressing the problem is to initially separate the co-occurring objects before feature extraction and then enhance category-specific representations to promote the discrepancy among co-contexts.

3.2. Framework Overview

Fig. 2 (a) specifically depicts the pipeline of SeCo. Given a space of input images X and a space of classification labels $\mathcal{Y} = \{1, 2, \ldots, K\}$, where the number of categories is denoted as K, the training dataset is defined as $D = \{(I_i, Y_i)\}_{i=1}^{V}$. Each tuple (I, Y) in D is the input, where $I \in \mathbb{R}^{3\times H\times W}$ is the image and $Y \in \mathcal{Y}$ is the class label. In our framework, we decompose an image I into patches x and remove the spatial dependence among co-categories. Then we build a dual-teacher-single-student architecture to extract multi-granularity knowledge and conduct semantic contrast to facilitate discrepancy among co-contexts.

Fig. 2 (b) illustrates the essence of the proposed losses L_{LiG}, L_{LiL} and tag rectification strategy. L_{LiG} means the loss between local patches and global images. The category knowledge (pink circle) from images acts in centroids for the patch representation (colored squares). L_{LiL} means the loss among local patches. It pushes apart patch semantics in different categories and pulls together those within the same. The rectification strategy excludes noisy patches in abnormal similarity with the help from category tags.

3.3. Image Decomposition

Assignment of Category Tags. In order to remove the spatial dependence on co-contexts, we propose to separate co-categories with image decomposition. As shown in Fig. 2 (a), given (I, Y) as input, we decompose the integral image I into multiple patches $x = \left\{x_i \in \mathbb{R}^{3\times h\times w}\right\}_{i=1}^{n}$ and the cropping process can be denoted as $x = crp(I)$, where $crp(\cdot)$ is the cropping operation, $h \times w$ is the size of each patch, and n represents the number of local patches.

The essence of contrastive learning is the construction of positive pairs [9]. Simply cropping images into patches cannot help differentiate co-categories [17,37]. To conduct contrast among co-categories, we assign a category tag t_i from the raw CAM seeds to each patch x_i and leverage tags to guide the subsequent class-specific contrast. Specifically, samples with the same tags are viewed as positive pairs, while samples with the different are negative pairs. We firstly generate CAM seeds by introducing an auxiliary classification head in the teacher network. It is found that the auxiliary head from intermediate features helps generate more diverse CAMs than that from final features [37]. Although CAMs cannot provide precise supervision, this raw signal effectively guides the assignment of category tags. We obtain the auxiliary pseudo mask M_{aux} from CAM_{aux}:

$$CAM_{aux} = Relu\left(W_{\lambda}^{T} Z_{F}^{\lambda}\right), \tag{1}$$

where Z_F^{λ} represents the features from the intermediate λ-th layer of the teacher encoder, W_{λ} is the mapping matrix in the corresponding classification head, and $Relu(\cdot)$ is the activation function. With CAM_{aux}, we obtain the auxiliary pseudo mask M_{aux} to guide the allocation of category tags $t = \{t_i\}_{i=1}^{n}$ for x. We have $m = crp(M_{aux})$ and $m = \left\{m_i \in \mathbb{R}^{h\times w}\right\}_{i=1}^{n}$, where m_i represents the pseudo mask patch for the image patch x_i.

As shown in Fig. 2 (a), we divide patches into background type $t_i = 0$, single category type $t_i = y_i$, and uncertain type $t_i = -1$, and allocate category tags accordingly. Specifically, a proportion threshold φ determines the tag types based on the proportion of target pixels to m_i. The uncertain tags represent noisy cases of separating co-occurrence and are excluded from the subsequent contrast.

However, since CAM seeds inevitably bring noise, the assigned category tags are possibly incorrect when guiding the contrast among co-contexts. Therefore, we design a rectification strategy to revise the noisy tags and guarantee the contrastive representation. It is detailed in Sec. 3.5.

Representation of Local Patches. Following the setups of popular contrastive approaches [15], we generate two augmented views from local patches x by implementing weak data augmentation (W.T.) $Aug_q(\cdot)$ and strong augmentation (S.T.) $Aug_k(\cdot)$, and send them to the student encoder $g_q(\cdot)$ and local teacher encoder $g_k(\cdot)$ to extract class embedding q_{cls} and k_{cls}, respectively, as shown in Fig. 2 (a). Class token in ViT is used to represent the embedding as it generalizes high-level semantics [4]. We further adopt a MLP operation $O_q(\cdot)$ and $O_k(\cdot)$ on the obtained embeddings to strengthen the feature and obtain the final representation $q = \left\{q_i \in \mathbb{R}^{1\times C}\right\}_{i=1}^{n}$ and $k = \left\{k_i \in \mathbb{R}^{1\times C}\right\}_{i=1}^{n}$, which denotes the local semantics from patches. The patch representation is formulated as:

$$q = O_q\left(g_q\left(\mathrm{Aug}_q(x)\right)\right), k = O_k\left(g_k\left(\mathrm{Aug}_k(x)\right)\right). \tag{2}$$

3.4. Representation with Category Knowledge

Class-specific Knowledge Extraction. After spatially separating the co-occurring context, we build a global teacher to dynamically extract category knowledge from integral images. Notably, considering the image decomposition may destruct the semantic context of patches while global CAMs lack local details, we train both teacher and student and share the encoders to facilitate the knowledge communication between local and global semantics.

Previous works [8,28] extract semantics based on CAM. SeCo extracts category knowledge P by utilizing the virtue of class token in ViT [4]. It represents the high-level semantics of each category and avoids the noise from false localization of CAMs. In particular, the knowledge set from images consists of K prototypes and each prototype generalizes the corresponding category semantics, i.e.,$P = \left\{P_l \in \mathbb{R}^{1\times C}\right\}_{l=1}^{K}$. Given the input image I with categories l, we extract the class token Z_l from the global teacher encoder $f(\cdot)$ to denote the category representation. The process can be formulated as $Z_l = f(I)$.

To reduce the noise from co-occurring objects and comprehensively generalize the knowledge, we propose an adaptive updating strategy to gather all semantics across the dataset. Given the token Z_l with multi-class, we calculate the cosine similarity with the corresponding prototypes and leverage the similarity scores after $softmax(\cdot)$ as the weights $W = \left\{W_l \in \mathbb{R}^{1\times C}\right\}_{l=1}^{K}$ to estimate the relevance to the corresponding category knowledge. The updating process can be formulated as:

$$P_l \leftarrow Norm\left(\eta P_l + W_l \cdot (1-\eta) Z_l\right). \tag{3}$$

Particularly, based on the prior that class tokens from single-category images are most relevant to the corresponding category semantics, the single class tokens are only used to update the prototypes and $W_l = 1.0$ at this time.

Knowledge Guided Contrast for Co-categories. With the global category knowledge and patch semantics q, we design L_{LiG} loss inspired by InfoNCE [30] to guide the student training. To promote the discrepancy of co-categories, only the co-categories are viewed as negative pairs while the semantics within the same category is positive pairs. Hence, the diversity comparison is held between the filtered local semantics $q_s = \{q_i\}_{i=1}^{u}$ and the knowledge P_s with the appeared categories, where u is the number of patches. The contrast between patch semantics and category prototypes is achieved by:

$$L_{LiG} = -\frac{1}{N_g^+}\sum_{i=1}^{u}\log\frac{\exp\left(q_i^T P_l^+/\tau_g\right)}{\sum_{P_l\in P_s}\exp\left(q_i^T P_l/\tau_g\right)}, \quad (4)$$

where N_g^+ counts the number of positive pairs between patches and prototypes, P_l^+ is the positive prototypes within the same category l with q_i and τ_g is the temperature factor.

3.5. Representation with Patch Semantics

Local Semantics Extraction. Following the memory setup in contrastive learning, we store patch semantics across the dataset. However, simply taking two views of a patch as a positive pair is not helpful to learn the difference among co-contexts. We propose a category tag pool to match the memory bank and guide the contrast among co-categories. Both patch semantics and category tags are stored as supportive knowledge to decouple the co-context at patch level. Specifically, we build a local teacher to extract features from history patches and update the reservoir and tag pool chronologically by storing the most recent key semantics k and its corresponding tag t_i while dequeuing the oldest. Mathematically, given the input (x, t), the current query and key embeddings with tags are denoted as B_q and B_k, the oldest are B_{-q} and B_{-k}, respectively. Then the reservoir paired with tags is defined as:

$$R(x,t) = B_q \cup B_k \cup queue\backslash\{B_{-q}\cup B_{-k}\}, \quad (5)$$

where $queue \in \mathbb{R}^{N\times C}$ is the history local semantics paired with tags in the reservoir and N is the reservoir capacity.

Importantly, we update local teacher from student with EMA to keep the memories consistent for contrast and avoid the dramatic variance between the older memories and the newest in the reservoir [15].

Rectification of Noisy Category Tags. The tags from CAM seeds are inevitably noisy and bring noise in contrast. To remedy it, we propose a similarity-based rectification strategy to denoise the tags. Since the similarity between two patches with the same category should be significantly higher than those different [52], we leverage the memories in the reservoir to rectify noisy tags in an unsupervised manner. When embedding k_i with a tag t_i updates the reservoir, we compute the inner product between its query view q_i and history embeddings $R(x, t_i)_+$ to measure the similarity. Then the average similarity $\mu(q_i, t_i)$ is calculated with:

$$\mu(q_i, t_i) = \frac{1}{|R(x,t_i)_+|}\sum_{k_+\in R(x,t_i)_+} q_i^T k_+, \quad (6)$$

where k_+ is the embedding from $R(x, t_i)_+$. Embeddings in the reservoir with noisy category tags hold a smaller fraction, thus the average similarity between the falsely tagged samples and the true positive samples is lower than that between true positive samples. Once the number of abnormal-similarity pairs exceeds a certain proportion σ, we consider q_i as a noisy embedding eventually. At this point, we change the category tags t_i to uncertain and exclude them from the contrast. The rectification process is denoted as:

$$t_i \leftarrow -1, if\,\frac{N_v}{|R(x,t_i)_+|} > \sigma, \quad (7)$$

where $N_v = \sum_{k_+\in R(x,t_i)_+}\mathbb{1}\left(q_i^T k_+ < \mu(q_i, t_i)\right)$ is the number of noisy pairs.

Tag Guided Contrast for Co-categories. With the category tags and history patch semantics, we design a contrastive loss L_{LiL} to differentiate co-categories at patch level. The semantics with the same category tag is viewed as positive pairs and the noisy patch semantics is excluded. With the query semantics q and local embeddings $R(x, t_i)_+$, we apply the loss to supervise the above process. The contrast among patch semantics is denoted as:

$$L_{LiL} = -\frac{1}{N_l^+}\sum_{i=1}^{n}\sum_{k_+} M_f \log\frac{\exp\left(q_i^T k_+/\tau_l\right)}{\sum_{k'\in R(x,t)}\exp\left(q_i^T k'/\tau_l\right)}, \quad (8)$$

where $M_f = \mathbb{1}(t_i \neq -1)$ is the rectification mask to exclude noisy patches, N_1^+ is the number of positive pairs, n is the number of patches and τ_l is a temperature factor.

3.6. Training Objectives

As shown in Fig. 2 (a), loss functions for SeCo consist of two contrast losses, i.e., L_{LiG} and L_{LiL}, and a classification loss L_{cls}. Apart from it, we also implement an auxiliary classification loss L_{cls}^{aux} to supervise the generation of auxiliary pseudo masks and allocate tags to local patches. Both L_{cls} and L_{cls}^{aux} adopt multi-label soft margin loss. The loss objectives of our SeCo are:

$$L_{SeCo} = L_{cls} + L_{cls}^{aux} + \alpha L_{LiG} + \beta L_{LiL}. \quad (9)$$

It is noted that the proposed framework SeCo generates the pseudo masks online and is trained end-to-end to achieve the dense segmentation task. The loss for segmentation adopts cross-entropy loss L_{seg}. Thus, the overall loss is: $L = L_{SeCo} + \gamma L_{seg}$. Following previous approaches [36, 37,40], we leverage regularization losses to enforce the spatial consistency of CAMs and the predicted masks.

Table 1. Comparisons with SOTAs in mIoU(%). $\mathcal{M}$:multi-staged, $\mathcal{S}$:single-staged. $\mathcal{I}$:image labels. $\mathcal{SA}$:saliency maps. $\mathcal{E}$: external data.

(a) Performance on PASCAL VOC [14].

Methods	Type	Backbone	CAM	Mask	Val	Test
AdvCAM [19] CVPR'2021	$\mathcal{M}$	ResNet101	55.6	68.0	68.1	68.0
GSM [24] AAAI'2021		ResNet101	-	-	68.2	68.5
CDA [39] CVPR'2021		ResNet38	58.4	66.4	66.1	66.8
W-OoD [20] CVPR'2022		ResNet38	59.1	72.1	70.7	70.1
CLIMS [46] CVPR'2022		ResNet101	56.6	70.5	70.4	70.0
L2G [17] CVPR'2022		ResNet101	-	71.9	72.1	71.7
FPR [6] ICCV'2023		ResNet101	63.8	66.4	70.3	70.1
OCR [10] CVPR'2023		ResNet38	61.7	69.1	72.7	72.0
1Stage [2] CVPR'2020	$\mathcal{S}$	ResNet38	-	66.9	62.7	64.3
AFA [36] CVPR'2022		MiT-B1	65.0	68.7	66.0	66.3
ViT-PCM [35] ECCV'2022		ViT-B/16	67.7	71.4	70.3	70.9
ToCo [37] CVPR'2023		ViT-B/16	71.6	72.2	71.1	72.2
SeCo(Ours)		**ViT-B/16**	**74.8**	**76.5**	**74.0**	**73.8**

(b) Performance on MS COCO [26].

Methods	Type	Backbone	Sup.	Val
EPS [22] CVPR'2021	$\mathcal{M}$	ResNet101	$\mathcal{I} + \mathcal{SA}$	35.7
RCA [52] CVPR'2022		ResNet101		36.8
L2G [17] CVPR'2022		ResNet101		44.2
CDA [39] CVPR'2021		ResNet38	$\mathcal{I} + \mathcal{E}$	33.2
CLIP-ES [27] CVPR'2023		ResNet101		45.4
MCTformer [48] CVPR'2022		ResNet38	$\mathcal{I}$	42.0
FPR [6] ICCV'2023		ResNet101		43.9
OCR [10] CVPR'2023		ResNet38		42.5
1Stage [2] CVPR'2020	$\mathcal{S}$	ResNet38	$\mathcal{I}$	-
SLRNet [31] IJCV'2022		ResNet38		35.0
AFA [36] CVPR'2022		MiT-B		38.9
ToCo [37] CVPR'2023		ViT-B/16		42.3
SeCo(Ours)		**ViT-B/16**		**46.7**

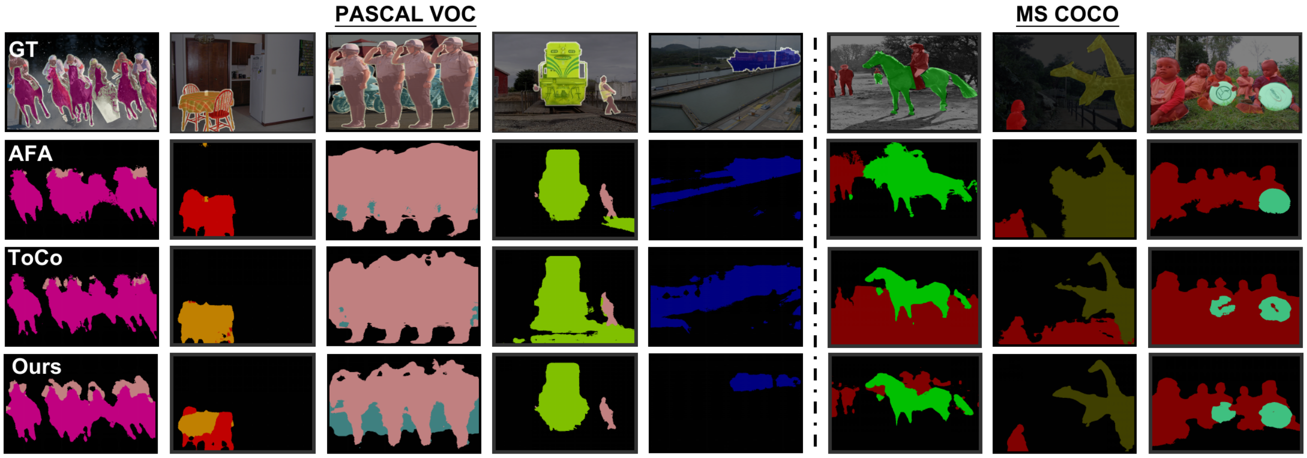

Figure 3. Qualitative segmentation results of AFA [36], ToCo [37] and ours on VOC and COCO. SeCo differentiates co-contexts precisely.

4. Experiments and Results

4.1. Experimental Settings

Datasets and Evaluation Metrics. The proposed method is evaluated on PASCAL VOC 2012 [14] and MS COCO 2014 [26]. PASCAL VOC contains 21 semantic categories. Following the practice [29, 36, 37], we use the augmented dataset with $10,582$, $1,449$, and $1,456$ images for training, validating, and testing, respectively. MS COCO includes 81 classes. $82,081$ images are used for training, and $40,137$ images are used for validating. Mean Intersection-Over-Union (mIoU) is used as evaluation criteria. Confusion ratio, i.e., the number of false positive pixels / that of true positives, is designed to evaluate the efficacy of suppressing false positives from co-occurrence.

Implementing Details. Encoders in the dual-teacher single-student framework all adopt ViT-B [12] as the backbone and are initialized with pre-trained weights on ImageNet [34]. Our decoder adopts a simple segmentation head with four 3×3 convolution layers. Following the training strategy in [36,37], we use AdamW optimizer to train SeCo with a polynomial scheduler. The crop size and the number of the local patches are set to 64×64 and 12. The capacity of the semantic reservoir is 4608. The loss weight factors (α, β, γ) in sec.3.5 are set as $(0.5, 0.5, 0.12)$. All experiments are conducted on RTX 3090 GPU. **Please refer to Supplementary Materials for more details**.

4.2. Main Results

Evaluation of CAMs and Pseudo Masks. Tab. 1 (a) quantitatively reports the quality of the initial CAMs and pseudo masks generated by SeCo and other recent competitors on VOC train set. It shows that SeCo generates better CAM seeds with 74.8% mIou, even surpassing other methods refined with post-processing [18]. With the simple multi-scale refinement [37], the quality of pseudo masks further improves to 76.5%, significantly higher than both single-staged and multi-staged methods by at least 4.3%.

Performance of Semantic Segmentation. Tab. 1 (a) reports the performance of the semantic segmentation of SeCo on VOC. The proposed SeCo achieves 74.0% and 73.8% mIoU on the val set and test set, respectively. Tab. 1 (b) compares our segmentation performance with other recent methods on COCO val set. Without external data, SeCo achieves 46.7% mIoU and outperforms multi-staged [27] competitors tackling co-occurrence and single-

Table 2. Abalation study of SeCo on VOC val set.

Conditions	LiG	LiL	Tag Rec.	Recall	Precision	mIoU
Baseline (ViT-B)				-	-	54.2
w/o LiG		✓	✓	81.1	79.2	69.1
w/o LiL	✓			82.7	80.9	70.3
w/o Tag Rec.	✓	✓		83.8	82.6	72.4
SeCo	✓	✓	✓	**85.0**	**84.0**	**74.0**

Table 3. Comparison of IoU and confusion ratio (in the bracket) with recent methods tackling co-occurrence on VOC val set.

	AFA [36]	ToCo [37]	**SeCo(Ours)**	
Train w/(Railroad)	59.6 (0.63)	58.0 (0.75)	62.2	**(0.54)**
Boat w/(Water)	64.6 (0.42)	43.6 (1.11)	68.4	**(0.32)**
Aeroplane w/(Sky)	79.3 (0.12)	77.3 (0.19)	86.3	**(0.07)**
Chair w/(Sofa)	29.6 (1.09)	35.6 (0.65)	38.3	**(0.48)**
Sofa w/(Chair)	44.6 (0.57)	43.8 (0.77)	57.4	**(0.35)**
Horse w/(Person)	76.0 (0.14)	83.4 (0.09)	85.9	**(0.05)**
All Categories	66.0 (0.36)	71.1 (0.32)	74.0	**(0.23)**

Table 4. Efficiency performance of SeCo compared to others. The experiment is conducted on PASCAL VOC with RTX 3090.

$\mathcal{M}$	CAM	Refine	Decoder	Val	Test
CLIMS [46]	101 mins	332 mins	635 mins	70.4	70.0
$\mathcal{S}$					
AFA [36]		554 mins		66.0	66.3
ToCo [37]		506 mins		71.1	72.2
SeCo(Ours)		**417 mins**		**74.0**	**73.8**

staged [37] SOTAs by 1.3% and 4.4% mIoU, respectively.

Prediction results on VOC and COCO are visualized in Fig. 3. It illustrates that SeCo can precisely localize the co-occurring objects on both datasets. For example, our method is capable of filtering the distracting backgrounds (water, railroad) from objects, or differentiating the co-occurring foregrounds (horse, person and bicycle), which demonstrates the competence at addressing co-occurrence.

5. Ablation Study and Further Analysis

5.1. Efficacy of Key Components

Ablative experiments on the key components of SeCo are conducted. Tab. 2 shows the segmentation results on VOC val set. Here, w/o LiG means no category prototypes is extracted, w/o LiL means that the local semantic reservoir and category tag pool are not maintained, and w/o Tag Rec. means the tag rectification is not incorporated. The category prototypes from G-teacher act as class centroids for training while the patch knowledge from L-teacher makes the category semantics more compact, which facilitates the difference among co-contexts. As can be seen, without LiG, the precision and recall drop heavily by 4.8% and 3.9%, respectively. It verifies that the proposed method can effectively help suppress the false positives and generate more complete masks. Without LiL, the precision, recall and mIoU drop heavily as well. The tag rectification improves precision from 82.6% to 84.0%. It works by guaranteeing the right constructions of positive pairs and negative pairs, reducing the noise in the contrastive representation.

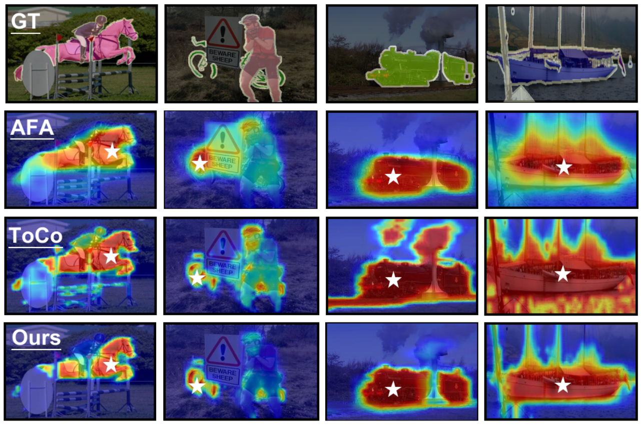

Figure 4. CAMs for co-contexts on VOC between SeCo and competitors [36, 37]. SeCo accurately activates the targets (star).

5.2. Effectiveness of Tackling Co-occurrence

In Tab. 3, representative co-occurring pairs (e.g., {Boat, water}, {train, railroad}, etc.) in VOC val set are selected to validate the effectiveness of our method to tackle co-occurrence. The metrics of IoU and confusion ratio (in bracket) are adopted. Confusion ratio is calculated by FP/TP, the lower the better. It reports that SeCo demonstrates significantly lower confusion ratio than ToCo, such as boat (−79%), train (−21%), and higher IoU in all representative co-occurring pairs. For all categories, SeCo surpasses other competitors with 0.23 confusion ratio, lower by at least 9%, which shows the superiority of our method to suppress the false positives from co-occurrence.

In addition, we visualize CAMs from SeCo and recent impressive methods in Fig. 4. For the co-occurring objects marked with white stars (horse, bicycle, train, boat), previous methods cannot solve the co-occurrence properly. They are frequently confused by co-contexts and intend to falsely activate the co-occurring backgrounds (water, railroad) or the related foregrounds (person), which can be attributed to the ignoring of noise from co-occurrence during the feature representation. Instead, contributing to the proposed 'separate and conquer' scheme, SeCo shows strength at accurately activating the co-contexts, which could plainly demonstrate the effectiveness in tackling the challenge.

5.3. Further Analysis

Hyper-parameter Sensitive Analysis. The analysis of key parameters, such as *patch size, loss weights, reservoir capacity, EMA momentum, temperature factors, etc.*, is specifically discussed in **Supplementary Materials**.

Table 5. The comparison to the fully-supervised counterparts on VOC val set. $\mathcal{I}$:image-level labels. $\mathcal{E}$: external data.

Methods	Backbone	Sup.	Val	Ratio
DeepLabV2 [7] TPAMI'2017	ResNet101	$\mathcal{F}$	77.7	-
WideResNet [45] PR'2019	ResNet38		80.8	-
Segfromer [38] ICCV'2021	MiT-B1		78.7	-
DeepLabV2 [7] TPAMI'2017	ViT-B/16		82.3	-
Multi-staged methods				
CDA [39] CVPR'2021	ResNet38	$\mathcal{I}+\mathcal{E}$	66.1	81.8%
W-OoD [20] CVPR'2022	ResNet38		70.7	87.5%
CLIMS [46] CVPR'2022	ResNet101		69.3	89.2%
AdvCAM [19] CVPR'2021	ResNet101	$\mathcal{I}$	68.1	87.6%
PPL [23] TMM'2023	ResNet38		67.8	87.3%
MCTformer [48] CVPR'2022	ResNet38		71.9	89.0%
Single-staged methods				
1Stage [2] CVPR'2020	ResNet38	$\mathcal{I}$	62.7	77.6%
SLRNet [31] IJCV'2022	ResNet38		69.3	85.8%
AFA [36] CVPR'2022	MiT-B1		66.0	83.9%
ViT-PCM [35] ECCV'2022	ViT-B/16		70.3	85.4%
ToCo [37] CVPR'2023	ViT-B/16		71.1	86.4%
SeCo(Ours)	**ViT-B/16**		**74.0**	**89.9%**

Training Efficiency Analysis. SeCo is designed in a single-staged paradigm to efficiently tackle the co-occurrence issue. The training efficiency comparisons are reported in Tab. 4. CLIMS [46] leverages CLIP model to tackle co-occurrence and consists of 3 progressive steps, which takes 1068 minutes to finish the WSSS workflow. Compared to it, SeCo takes 417 minutes to finish the workflow and outperforms it by a significant margin. Notably, SeCo achieves more favorable performance compared to other single-staged competitors [36, 37] as well.

Fully-supervised Counterparts. Since competitors in Tab. 1 use different backbones, we report the upper bound performance on VOC val set for fair comparison in Tab. 5. Although ViT intends to have advantageous performance in vision tasks, our method achieves 74.0 mIoU and 89.9% to its fully-supervised performance, which significantly outperforms other single-staged methods with ViT backbone and holds superiority over other multi-staged competitors [20, 39, 46] tackling co-occurrence with external data.

Feature Representation Analysis. We visualize the co-context feature representation at image level to validate the effectiveness. As shown in the right of Fig. 5, we compute the similarity among the category knowledge from images. It is observed that each prototype in the knowledge is only highly related to itself, which suggests that the co-occurring semantics is separated. As shown in the left of Fig. 5, the visualization with t-SNE [41] also validates the efficacy.

In addition, we further visualize the co-context feature representation at patch level. We decompose a demo image with co-occurring objects {dining table, chair} into 16 patches, as shown in the left of Fig. 6. Each star (index from 1 to 16) denotes a patch, orange stars for table semantics and blue ones for chair semantics. We calculate the similarity between patches and category prototypes. As shown in the right of Fig. 6, the patch semantics has the correct relationship to the corresponding prototypes. Moreover, the patches with index $(3, 7, 10)$ containing both table and chair semantics show a high relationship to both category information. This validates that SeCo can successfully recognize co-contexts instead of being biased to the one or another.

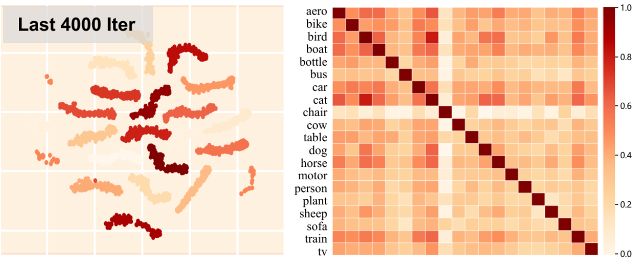

Figure 5. Category representation of SeCo on PASCAL VOC. Left: category prototypes from last $4,000$ iterations are visualized with t-SNE [41]. Right: similarity among the category prototypes.

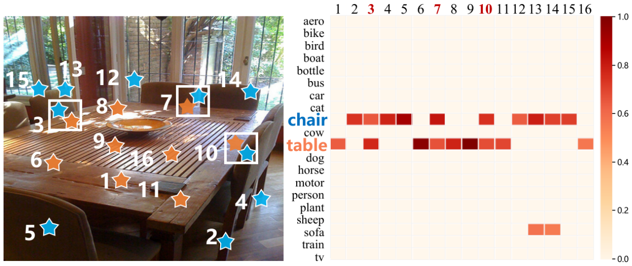

Figure 6. Patch feature representation. Left: sample image with co-contexts {table, chair}. The orange star represents the table patch and the blue is the chair patch. Right: similarity among patch semantics and category prototypes. The x-axis means the patch index and the y-axis is the 20 category prototypes on VOC.

6. Conclusion

In this paper, we propose to tackle the widespread co-occurrence problem in WSSS from a new perspective of 'separate and conquer' manner by designing image decomposition and contrastive representation. Extensive experiments are conducted on PASCAL VOC and MS COCO, validating the effectiveness of tackling co-occurrence issue.

One limitation is that, although we make the attempt to allocate tags and reduce the bias, it inevitably remains co-category patches and allocates wrong tags. In the future, leveraging patches with adaptive size or adopting other denoising techniques is the potential research for WSSS.

7. Acknowledgement

This work was supported by National Natural Science Foundation of China (82072021), Shanghai Municipal Science and Technology Committee (23410710400, 22YF1409300).

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. 1

[2] 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. 1, 2, 6, 8

[3] 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

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9650–9660, 2021. 3, 4

[5] Liyi Chen, Chenyang Lei, Ruihuang Li, Shuai Li, Zhaoxiang Zhang, and Lei Zhang. Fpr: False positive rectification for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1108–1118, 2023. 2

[6] Liyi Chen, Chenyang Lei, Ruihuang Li, Shuai Li, Zhaoxiang Zhang, and Lei Zhang. Fpr: False positive rectification for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1108–1118, 2023. 6

[7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 8

[8] Qi Chen, Lingxiao Yang, Jian-Huang Lai, and Xiaohua Xie. Self-supervised image-specific prototype exploration for weakly supervised semantic segmentation. In *CVPR*, pages 4288–4298, June 2022. 4

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 3, 4

[10] Zesen Cheng, Pengchong Qiao, Kehan Li, Siheng Li, Pengxu Wei, Xiangyang Ji, Li Yuan, Chang Liu, and Jie Chen. Out-of-candidate rectification for weakly supervised semantic segmentation. In *CVPR*, pages 23673–23684, 2023. 6

[11] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *ICCV*, pages 1635–1643, 2015. 1

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 6

[13] Ye Du, Zehua Fu, Qingjie Liu, and Yunhong Wang. Weakly supervised semantic segmentation by pixel-to-prototype contrast. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4320–4329, June 2022. 3

[14] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 111:98–136, 2015. 6

[15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 3, 4, 5

[16] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 3

[17] Peng-Tao Jiang, Yuqi Yang, Qibin Hou, and Yunchao Wei. L2g: A simple local-to-global knowledge transfer framework for weakly supervised semantic segmentation. In *CVPR*, pages 16886–16896, 2022. 2, 3, 4, 6

[18] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *NeurIPS*, 24, 2011. 6

[19] 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, 6, 8

[20] 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. 2, 6, 8

[21] 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

[22] 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, 6

[23] Jinlong Li, Zequn Jie, Xu Wang, Yu Zhou, Xiaolin Wei, and Lin Ma. Weakly supervised semantic segmentation via progressive patch learning. *IEEE Transactions on Multimedia*, 2022. 2, 8

[24] Xueyi Li, Tianfei Zhou, Jianwu Li, Yi Zhou, and Zhaoxiang Zhang. Group-wise semantic mining for weakly supervised semantic segmentation. In *AAAI*, volume 35, pages 1984–1992, 2021. 2, 6

[25] 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

[26] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 6

[27] Yuqi Lin, Minghao Chen, Wenxiao Wang, Boxi Wu, Ke Li, Binbin Lin, Haifeng Liu, and Xiaofei He. Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. *arXiv preprint arXiv:2212.09506*, 2022. 2, 6

[28] Weide Liu, Xiangfei Kong, Tzu-Yi Hung, and Guosheng Lin. Cross-image region mining with region prototypical network for weakly supervised segmentation. *IEEE Transactions on Multimedia*, 2021. 4

[29] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 6

[30] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 5

[31] Junwen Pan, Pengfei Zhu, Kaihua Zhang, Bing Cao, Yu Wang, Dingwen Zhang, Junwei Han, and Qinghua Hu. Learning self-supervised low-rank network for single-stage weakly and semi-supervised semantic segmentation. *IJCV*, 130(5):1181–1195, 2022. 2, 6, 8

[32] Pedro O Pinheiro and Ronan Collobert. From image-level to pixel-level labeling with convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1713–1721, 2015. 1

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[34] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021. 6

[35] Simone Rossetti, Damiano Zappia, Marta Sanzari, Marco Schaerf, and Fiora Pirri. Max pooling with vision transformers reconciles class and shape in weakly supervised semantic segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXX*, pages 446–463. Springer, 2022. 6, 8

[36] Lixiang Ru, Yibing Zhan, Baosheng Yu, and Bo Du. Learning affinity from attention: end-to-end weakly-supervised semantic segmentation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16846–16855, 2022. 2, 5, 6, 7, 8

[37] Lixiang Ru, Heliang Zheng, Yibing Zhan, and Bo Du. Token contrast for weakly-supervised semantic segmentation. *arXiv preprint arXiv:2303.01267*, 2023. 2, 3, 4, 5, 6, 7, 8

[38] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7262–7272, 2021. 8

[39] 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. 2, 6, 8

[40] Meng Tang, Federico Perazzi, Abdelaziz Djelouah, Ismail Ben Ayed, Christopher Schroers, and Yuri Boykov. On regularized losses for weakly-supervised cnn segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 507–522, 2018. 5

[41] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 8

[42] Paul Vernaza and Manmohan Chandraker. Learning randomwalk label propagation for weakly-supervised semantic segmentation. In *CVPR*, pages 7158–7166, 2017. 1

[43] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. Exploring cross-image pixel contrast for semantic segmentation. In *ICCV*, pages 7303–7313, 2021. 3

[44] Xiang Wang, Shaodi You, Xi Li, and Huimin Ma. Weakly-supervised semantic segmentation by iteratively mining common object features. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1354–1362, 2018. 1

[45] Zifeng Wu, Chunhua Shen, and Anton Van Den Hengel. Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognition*, 90:119–133, 2019. 8

[46] 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. 2, 6, 7, 8

[47] 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. 1

[48] 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. 2, 6, 8

[49] Rongtao Xu, Changwei Wang, Jiaxi Sun, Shibiao Xu, Weiliang Meng, and Xiaopeng Zhang. Self correspondence distillation for end-to-end weakly-supervised semantic segmentation. *arXiv preprint arXiv:2302.13765*, 2023. 3

[50] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, pages 2921–2929, 2016. 1

[51] Tianfei Zhou, Wenguan Wang, Ender Konukoglu, and Luc Van Gool. Rethinking semantic segmentation: A prototype view. In *CVPR*, pages 2582–2593, 2022. 3

[52] Tianfei Zhou, Meijie Zhang, Fang Zhao, and Jianwu Li. Regional semantic contrast and aggregation for weakly supervised semantic segmentation. In *CVPR*, pages 4299–4309, 2022. 3, 5, 6