DIAL: Dense Image-text ALignment for Weakly Supervised Semantic Segmentation

Soojin Jang[1], Jungmin Yun[2], Junehyoung Kwon[2], Eunju Lee[1], and Youngbin Kim[1,2]

[1] Graduate School of Advanced Imaging Science, Multimedia & Film,
Chung-Ang University, Korea
[2] Department of Artificial Intelligence, Chung-Ang University, Korea
{sujin0110,cocoro357,dirchdmltnv,dmswn5829,ybkim85}@cau.ac.kr

Abstract. Weakly supervised semantic segmentation (WSSS) approaches typically rely on class activation maps (CAMs) for initial seed generation, which often fail to capture global context due to limited supervision from image-level labels. To address this issue, we introduce DALNet, Dense Alignment Learning Network that leverages text embeddings to enhance the comprehensive understanding and precise localization of objects across different levels of granularity. Our key insight is to employ a dual-level alignment strategy: (1) Global Implicit Alignment (GIA) to capture global semantics by maximizing the similarity between the class token and the corresponding text embeddings while minimizing the similarity with background embeddings, and (2) Local Explicit Alignment (LEA) to improve object localization by utilizing spatial information from patch tokens. Moreover, we propose a cross-contrastive learning approach that aligns foreground features between image and text modalities while separating them from the background, encouraging activation in missing regions and suppressing distractions. Through extensive experiments on the PASCAL VOC and MS COCO datasets, we demonstrate that DALNet significantly outperforms state-of-the-art WSSS methods. Our approach, in particular, allows for more efficient end-to-end process as a single-stage method.

Keywords: weakly supervised semantic segmentation · image-level labels supervision · single-stage framework

1 Introduction

Semantic segmentation is the task of assigning a semantic label to each pixel in an image. While training supervised models for such tasks demands labor-intensive pixel-level annotations, weakly supervised semantic segmentation (WSSS) addresses this challenge by learning to segment objects using class labels assigned to the entire image [16, 19, 49, 54, 67].

In most WSSS methods with image-level class labels, class activation maps (CAMs) [68] are used as the initial seeds, refined to generate pseudo labels, and finally trained segmentation with the pseudo labels. However, CAMs often

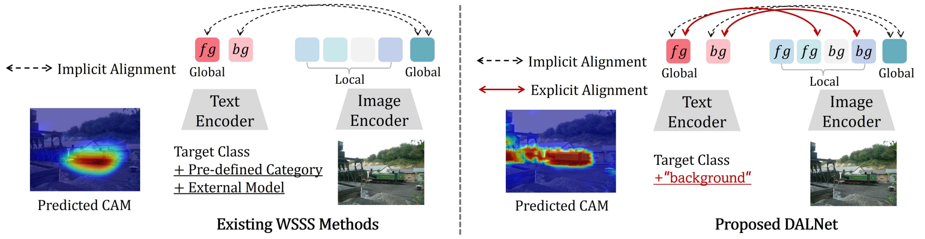

Fig. 1: Comparison of **(Left)** existing WSSS methods and **(Right)** proposed DALNet. **(Left)** Existing methods for implicit alignment depend on global image features, potentially missing local details within the image. **(Right)** In contrast, the proposed DALNet integrates global and local features to preserve spatial details and facilitate explicit alignment. It distinguishes between foreground and background in image patches and text, addressing various levels of granularity. The dual alignment mechanism captures diverse object regions without any pre-defined category or external model.

struggle to accurately identify object regions, primarily relying on convolutional neural networks (CNNs) with inherent limitations in contextual understanding, hindering the activation of object regions in images [23, 32, 34]. Several studies have proposed improving CAMs using vision transformer (ViT) architecture [15], recognized for capturing global relationships, offering a potential improvement over CAMs in addressing WSSS challenges [48, 49, 62].

Despite these advancements, relying solely on image-level class labels for WSSS can lead to incomplete object activation in CAMs. To address this issue, several language-guided models [43, 63] have emerged, leveraging natural language to emphasize visual relationships and object categories. Recent studies on WSSS [38, 59, 63] have incorporated text supervision to address the challenges posed by limited guidance. These approaches have adopted contrastive language-image pre-training (CLIP) [43] to extract text representations from linguistic descriptions. CLIP employs a contrastive learning strategy to align images with their corresponding texts, distinguishing them from other texts. However, existing WSSS methods often depend on implicit alignment, aggregating representations such as global image features, and neglecting the spatial details, as illustrated in Fig. 1 (Left) [38, 59]. These methods only implicitly align visual and textual concepts at the same level of granularity [10], limiting the localization performance [64]. In addition, these methods often rely on predefined sets of background categories or external models to extract background text representations.

To overcome such limitations, we propose **Dense Alignment Learning Network (DALNet)**, which employs a dual-level alignment mechanism for vision-language representation learning in WSSS. As shown in Fig. 1 (Right), we consider various levels of granularity by differentiating between global and local attributes using class and patch tokens, further dividing them into foreground and background for both image and text. Firstly, we implement **Global Implicit Alignment (GIA)** to capture the rich semantics of the image from a global perspective. This approach measures the similarity between the class token, which aggregates the global features of the image, and text embeddings that distinguish between the foreground target class and the background. Secondly, we propose **Local Explicit Alignment (LEA)** to enhance object localization by utilizing spatial details within patch tokens. We identify visual features for foreground and background patch tokens using an object-aware mask and then evaluate their similarity to corresponding text embeddings. This method facilitates explicit alignment in the vision-language embedding space through image-text cross-contrastive learning. Unlike existing methods, our novel cross-contrastive learning approach uniquely considers positive and negative pairs across both foreground and background representations. It adjusts the foreground in images to correspond with target classes, thereby mitigating over-activation in irrelevant areas and preventing the over-activation of excessively specific areas. Leveraging this granular information improves comprehension and object localization, while also activating missing regions and reducing distractions. Our method efficiently learns foreground and background representations using only class information and the "`background`" term, providing practical advantages without relying on additional prior knowledge.

Furthermore, unlike most existing WSSS approaches that involve multi-stage processes [21, 23, 59, 69], we integrate DALNet into a single-stage WSSS that shares the encoder for classification and segmentation networks, enabling end-to-end training. The main contributions of this paper are summarized as follows:

- We introduce Dense Alignment Learning Network (DALNet) for WSSS, which integrates global and local visual features through a dual-level alignment strategy. Global Implicit Alignment (GIA) and Local Explicit Alignment (LEA) are designed to enhance dense localization and comprehensive understanding of various objects within the image.
- We propose a novel cross-contrastive learning approach that aligns visual features with text embeddings by employing contrastive learning across an image's foreground and background at various levels of granularity. Using only the target class and the "`background`" term, this approach effectively activates relevant regions while suppressing irrelevant objects.
- Experimental results on the PASCAL VOC and MS COCO datasets demonstrate the effectiveness of our proposed method, surpassing state-of-the-art single-stage WSSS methods and exhibiting competitive performance comparable to multi-stage methods.

2 Related Work

2.1 Weakly Supervised Semantic Segmentation

Several WSSS methods have employed CNNs to generate CAMs as initial seeds [1, 2, 8, 21, 27, 29, 54]. However, recent research suggests that CNN-based architectures are limited in capturing global information effectively due to their restricted receptive fields [14,26,53]. Various approaches have been proposed to address the locality problem of the CNN architecture [23,25,25,34,56]. Recently, WSSS methods have adopted ViT as a backbone for generating localization maps [48,49,62]. By leveraging the multi-head self-attention mechanism, ViT effectively models long-range dependencies by gathering information from diverse regions, surpassing CNNs [44,52]. MCTFormer uses a data-efficient variant of ViT, DeiT-S [51], as its backbone to model interactions between multiple class tokens and patch tokens [62]. Moreover, L2G makes use of both global and local contexts to obtain more integral object attention [23].

To improve efficiency, various approaches have focused on single-stage WSSS. AFA learns reliable semantic affinity and refines the initial pseudo labels with low-level image appearance [48]. In addition, ToCo addresses over-smoothing and enhances semantic consistency by contrasting class tokens between the local and global regions of ViT [49]. Moreover, ViT-PCM preserves the locality characteristics of ViT and performs effective mapping between multi-label classification and semantic segmentation [46].

2.2 Vision-language Pre-training

Recently, visual-language pre-training (VLP) models addressing large-scale image-text pairs have demonstrated robust performance in downstream tasks [3,13,22,35,39,43]. Additionally, several approaches have focused on learning image representations with language supervision for dense prediction tasks [61, 65, 66]. Open-vocabulary semantic segmentation has been explored, which segments images based on arbitrary categories described through texts [40,55,61]. However, these methods often depend on fine-tuning modules or require segmentation annotations and extensive external training datasets. Some recent studies have integrated visual-language models and language supervision into WSSS [38,59,63]. CLIMS employs the visual-language model [43] and language-guided supervision to effectively identify object regions while suppressing background regions [59]. Additionally, class representation capabilities have been enhanced through the exploitation of language priors and class-specific tokens [63]. CLIP-ES employs text-driven strategies by modifying text input to capitalize on the strengths of CLIP [38]. Similarly, we introduce DALNet based on language-guided supervision. In contrast to other approaches, the proposed method can generate text embeddings without requiring additional models or pre-defining the related text. Furthermore, we employ GIA and LEA to address the limitations of aggregated features in existing methods, leading to poor object localization.

2.3 Contrastive Learning for Segmentation Tasks

Contrastive learning aims to train the model to bring similar samples close together and dissimilar samples far apart in the given input data without explicit labels for training [9,18,20]. This approach involves using pairs of identical samples as positives and pairs of different samples as negatives to learn the similarity between the data. Recently, several researches have aimed to enhance localization performance by leveraging the alignment between visual and text embeddings during segmentation [36,61]. GroupViT groups similar visual concepts by applying contrastive loss between pooled image tokens and text representations [61]. In addition, research has been conducted on open-vocabulary segmentation by contrasting masked image regions with nouns using CLIP [36]. Contrastive learning has been introduced into WSSS to bridge the information gap between weak labels and target labels [16, 28, 60, 69]. An approach has been introduced for contrasting pixels and prototypes to improve the performance of the localization map [16]. C2AM disentangles foreground and background attributes within an image by leveraging contrastive learning [60]. RCA employs contrastive loss to increase similarity in the same class for region-aware representation while decreasing similarity for different classes [69]. SMA leverages contrastive learning to separate distinct features in order to synthesize diverse combinations of object-background representations [28]. Unlike existing methods, the proposed method employs cross-contrastive learning between images and texts, considering positive and negative pairs across various granularities. This approach our proposed DALNet to activate in missing regions while suppressing distractions and focusing on the target objects.

3 Methodology

3.1 Method Overview

The overview of the proposed method is illustrated in Fig. 2. Visual and textual features are extracted from the image encoder and text encoder, respectively. Notably, we employ simple prompt templates, "a photo of `[CLS]`" and "a photo of `background`," to obtain text embeddings associated with foreground and background, respectively. We utilize intermediate features from the image encoder to generate an object-aware mask, which offers a rough localization of the object and background regions. Then, multiplying the object-aware mask by reshaped patch tokens generates foreground and background representations rich in regional information. We leverage cross-contrastive learning between the representations of the two modalities and employ LEA, preserving local information. Additionally, we perform GIA by maximizing the similarity between the class token obtained from the image encoder and the foreground text embeddings while minimizing the similarity with the background embeddings. Finally, a classifier produces the final CAMs to generate the final pseudo labels.

3.2 Revisiting Class Activation Maps

The primary objective of CAMs is to identify specific regions within images activated during the prediction process in the classification network. Due to their simplicity and efficiency, CAMs have gained extensive application as a method of generating initial seeds in WSSS. Given an image, feature maps $F \in \mathbb{R}^{HW \times D}$ are extracted from a classification network, where HW denotes the size of the spatial resolution and D represents channel dimensions. Within the classification

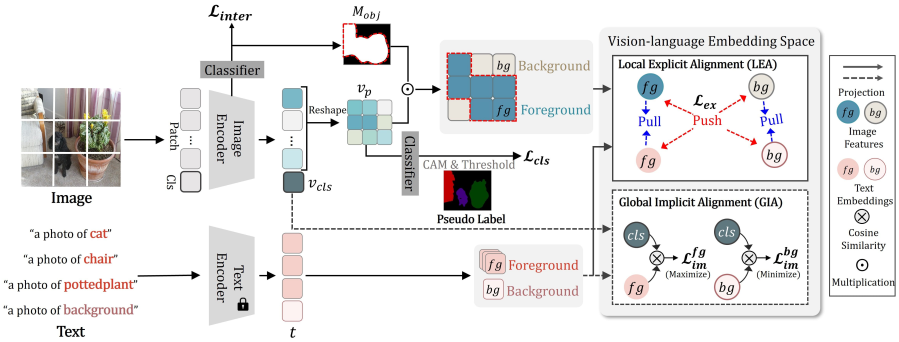

Fig. 2: Overview of the proposed DALNet. This approach employs a visual encoder to extract features and an object-aware mask M_{obj} to distinguish foreground and background. Text prompts are fed into a text encoder to generate embeddings for target objects and the background. Cross-contrastive learning aligns representations from both modalities, associating each token with either the foreground or background. GIA contrasts the class token with text embeddings to incorporate global information, while LEA leverages patch tokens and text embeddings for precise localization.

network, classifier weights $W \in \mathbb{R}^{C \times D}$ are used to weight and sum the feature maps F, where C denotes the number of classes. Negative activations can be mitigated by scaling CAMs to the range of $[0, 1]$ using the max normalization and the `ReLU` function. Thus, CAMs for c-th class are determined as follows:

$$\texttt{CAM}_c(F, W) = \frac{\texttt{ReLU}(A_c)}{\max(\texttt{ReLU}(A_c))}, \quad A_c = \sum_i W_{c,i} F_{:,i}. \tag{1}$$

In general, a background threshold β is used to differentiate between the foreground and background regions within the image.

3.3 Image and Text Representations

Visual Features. We employ ViT as a visual encoder, generating a class token that aggregates global semantics for predicting the output class, along with patch tokens that encode rich information about their corresponding local image patches [24]. To generate patch tokens $x_p \in \mathbb{R}^{N^2 \times d_v}$, we initially split the input image into $N \times N$ patches, which are then embedded into a sequence of N^2 patch tokens. Here, d_v represents the channel dimension. Subsequently, we use a learnable class token $x_{cls} \in \mathbb{R}^{1 \times d_v}$ as aggregated representations of the entire image. We prepend the class token x_{cls} to the sequence of patch tokens x_p, apply positional embedding, and then use them as input for the visual encoder. The resulting visual features $V = \{v_p, v_{cls}\}$ from the visual encoder comprise the patch and class tokens, denoted as $v_p \in \mathbb{R}^{N^2 \times d_v}$ and $v_{cls} \in \mathbb{R}^{1 \times d_v}$, respectively.

Furthermore, we leverage information from the intermediate layers to aggregate foreground and background attributes from visual features. To preserve semantic diversity, we leverage intermediate features that encode general class attributes rather than specific object details. Using these intermediate features, we construct an object-aware mask. Visual features F_{inter} are obtained through intermediate block from the visual encoder. A global max pooling operation is applied to aggregate intermediate patch tokens, following [48]. Following Eq. (1), we compute the intermediate CAMs, employing the visual feature F_{inter} and convolutional layer parameters θ_{inter}. The object-aware mask $M_{obj} \in \{0, 1\}^{N \times N}$ is obtained by applying the background threshold β to intermediate CAMs. M_{obj} is a binary map where the foreground region has a value of 1, and the background region has a value of 0, representing a coarse localization map of the input image. Then, we multiply the reshaped patch tokens $v_p \in \mathbb{R}^{N \times N \times d_v}$ by the object-aware mask M_{obj} as follows:

$$v_p^{fg} = v_p \odot M_{obj}, \quad v_p^{bg} = v_p \odot (1 - M_{obj}), \tag{2}$$

where $\odot$ denotes element-wise multiplication. In addition, v_p^{fg} and v_p^{bg} contain information about the foreground and background features, respectively. These features are used for dense alignment between image and text representations.

Text Embeddings. We employ the text encoder of the CLIP model to emphasize visual relationship contexts and object categories guided by natural language expressions. CLIP introduces a novel approach to conveying visual concepts. Text embeddings of the corresponding image are obtained using an intuitively simple prompt. For each class, we employ the template "a photo of `[CLS]`," where `[CLS]` represents the class label. Additionally, to represent text excluding the region of the target object in the image, we generate a background text embedding using the prompt "a photo of `background`". Through these prompts, we generate text embeddings $T = \left\{t^{fg}, t^{bg}\right\}$ using the CLIP text encoder, where $t^{fg} \in \mathbb{R}^{C \times d_t}$ and $t^{bg} \in \mathbb{R}^{1 \times d_t}$ denote foreground and background text embeddings, respectively. Further, d_t denotes channel dimension, and C represents the number of classes.

3.4 Dense Alignment between Image and Text Representations

In two modalities—the corresponding visual features and text embeddings—we expect semantic coherence at various levels of granularity. Following the aforementioned process, we obtain representations associated with the foreground and background of each modality. For visual features related to objects, the similarity to the foreground text embedding should be high, whereas the similarity with the background text embedding should be low. This principle also applies conversely, especially regarding background visual features. Building on this interaction, we employ a dual-level alignment strategy. To align the representations, we project the visual features from the visual encoder and the text embeddings from the text encoder into a unified vision-language embedding space.

Global Implicit Alignment. We provide comprehensive textual guidance by aligning visual features with text embeddings. Global Implicit Alignment (GIA) is achieved through cross-contrastive learning between the class token of the projected visual features and the projected text embeddings. The cosine similarity $\texttt{sim}(\cdot)$ is computed between the class token $v_{cls} \in \mathbb{R}^{1 \times d}$ and the foreground and background text embeddings, denoted as $t^{fg} \in \mathbb{R}^{C\times d}$ and $t^{bg} \in \mathbb{R}^{1\times d}$, respectively, where d denotes the projected feature dimension. As the class token v_{cls} aggregates global semantics, the objective is to maximize the similarity to t^{fg} and minimize the similarity to t^{bg}. The objective of GIA is represented as follows:

$$\mathcal{L}_{im} = -\sum_{c=1}^{C} y_c \cdot \log(\texttt{sim}(v_{cls}, t_c^{fg})) - \log(1 - \texttt{sim}(v_{cls}, t^{bg})), \tag{3}$$

where y_c denotes the image-level label for c-th class. Although class tokens have different semantics, GIA enables class tokens to effectively aggregate related visual global representations by incorporating background text embeddings. However, depending on GIA, it may neglect localization details, as it might not fully capture the entire object region due to limited interaction with the class token.

Local Explicit Alignment. To capture the entire object region and enhance localization details, we introduce Local Explicit Alignment (LEA). In contrast to GIA using the class token that aggregates semantic attributes, we leverage patch tokens that effectively preserve local information. For cross-contrastive learning, we explore various levels of granularity by utilizing foreground and background features from both images and texts.

We project flattened patch tokens, $v_p^{fg} \in \mathbb{R}^{N^2\times d}$ and $v_p^{bg} \in \mathbb{R}^{N^2\times d}$, obtained by applying the object-aware mask M_{obj}. We consider t^{fg} as the positive pair and t^{bg} as the negative pair for v_p^{fg}. Conversely, for v_p^{bg}, we match t^{bg} as the positive pair and t^{fg} as the negative pair.

$$s_+^{fg} = \sum_{c=1}^{C} y_c \cdot \texttt{sim}(v_p^{fg}, t_c^{fg}), \quad s_-^{fg} = \texttt{sim}(v_p^{fg}, t^{bg}), \tag{4}$$

$$s_+^{bg} = \texttt{sim}(v_p^{bg}, t^{bg}), \quad s_-^{bg} = \sum_{c=1}^{C} y_c \cdot \texttt{sim}(v_p^{bg}, t_c^{fg}), \tag{5}$$

where y_c denotes the image-level label for c-th class. Similar to other studies [16, 49, 60], we adopt the InfoNCE [41] loss as the objective.

$$\begin{aligned}\mathcal{L}_{ex} = &-\frac{1}{N^2}\sum_{i=1}^{N^2} \log \frac{\exp(s_{+_i}^{fg}/\tau)}{\exp(s_{+_i}^{fg}/\tau) + \exp(s_{-_i}^{fg}/\tau)} \\ &-\lambda \frac{1}{N^2}\sum_{i=1}^{N^2} \log \frac{\exp(s_{+_i}^{bg}/\tau)}{\exp(s_{+_i}^{bg}/\tau) + \exp(s_{-_i}^{bg}/\tau)},\end{aligned} \tag{6}$$

where λ is the hyperparameter for balancing, and τ is the temperature factor. We empirically set them to 0.001 and 1.0, respectively. Explicitly cross-aligning visual features and text embeddings across foreground and background representations enables the identification of specific regions based on the target class embedding. Additionally, by separating patch tokens based on the foreground and background and aligning them with the text embeddings, the classification network can more precisely distinguish objects and backgrounds in the image.

3.5 Network Training

We apply the global max pooling operation to the visual features obtained from the classifier of the classification network and the intermediate CAMs to compute the logits. Subsequently, we compute the multi-label soft margin loss, represented as $\mathcal{L}_{cls}$ and $\mathcal{L}_{inter}$, for each logit. The overall loss in the proposed method includes $\mathcal{L}_{im}$, $\mathcal{L}_{ex}$, $\mathcal{L}_{cls}$, and $\mathcal{L}_{inter}$. To address the problem of over-smoothing in the ViT, we incorporate the patch tokens contrast loss $\mathcal{L}_{ptc}$ following [49]. The loss for training the model is defined as follows:

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{inter} + \lambda_i \mathcal{L}_{im} + \lambda_e \mathcal{L}_{ex} + \lambda_p \mathcal{L}_{ptc}, \tag{7}$$

where λ_i, λ_e, and λ_p are hyperparameters. Furthermore, DALNet is designed as a single-stage WSSS framework. Additionally, to improve boundary alignment, we integrate the pixel-adaptive refinement (PAR) module [48]. This module refines the pseudo labels generated by DALNet, and these refined pseudo labels then supervise the segmentation head. For semantic segmentation, we employ the widely used cross-entropy loss, denoted by $\mathcal{L}_{seg}$. Similar to previous single-stage WSSS studies [42,47,49], we integrate an additional regularization loss [50] $\mathcal{L}_{reg}$ to ensure the spatial consistency of the predicted segmentation masks. The total loss for semantic segmentation is represented as follows:

$$\mathcal{L}_{total} = \mathcal{L} + \mathcal{L}_{seg} + \mathcal{L}_{reg}. \tag{8}$$

4 Experiments

4.1 Dataset and Evaluation Metric

We evaluate the proposed method on the PASCAL VOC 2012 [17] segmentation benchmark, consisting of 20 foreground classes and one background class. Following common practice in semantic segmentation, we take additional annotations from SBD [19]. The official dataset split contains 10,582 images for training, 1,449 for validation, and 1,456 for testing. We also employ the MS COCO 2014 [37] dataset, with 81 classes, 82k training, and 40k validation images. During the training stage, we only use image-level labels. To evaluate segmentation results, we employ the mean intersection over union (mIoU) metric.

4.2 Experimental Settings

Network Architectures. We use ViT-base (ViT-B) [15] as the visual encoder initiated from pre-trained weights obtained from ImageNet [45]. We employ bilinear interpolation to adapt the positional embedding, ensuring compatibility with the input size and enabling the handling of different-sized images. We utilize the frozen text encoder from the CLIP model. The segmentation head is constructed with a 1×1 prediction layer and two 3×3 convolutional layers.
Implementation Details. We use an AdamW optimizer. The learning rate

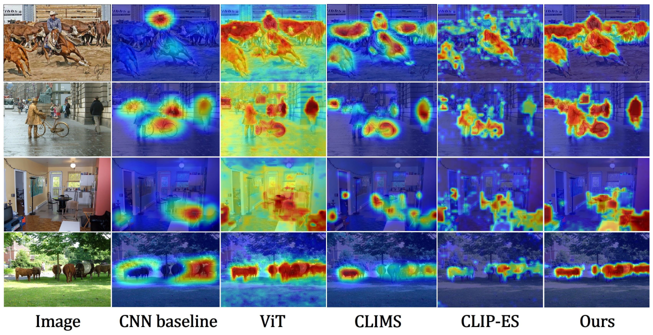

Fig. 3: Visualization results of CAMs. We generate CAMs using CNN baseline, ViT, CLIMS [59], CLIP-ES [38] and our proposed DALNet.

Table 1: Performance comparison of the initial CAMs on the PASCAL VOC dataset [17]. † denotes pre-trained parameters from ImageNet-21k [45].

Method	Backbone	Train	Val
RRM [67]AAAI'20	WR38	-	65.4
1Stage [4]CVPR'20	WR38	66.9	65.3
SLRNet [42]IJCV'22	WR38	67.1	66.2
AFA [48]CVPR'22	MiT-B1	68.7	66.5
ViT-PCM [46]ECCV'22	ViT-B†	67.7	66.0
CLIMS [59]CVPR'22	R50	56.6	-
MTCformer [62]CVPR'22	ViT-B	61.7	-
AMN [31]CVPR'22	R50	62.1	-
MMCST [63]CVPR'23	ViT-B†	66.3	-
CLIP-ES [38]CVPR'23	ViT-B†	70.8	-
ToCo [49]CVPR'23	ViT-B†	73.6	72.3
Ours	ViT-B	71.9	70.2
Ours†	ViT-B†	**75.2**	**73.1**

gradually increases to 6e-5 for the first 1,500 iterations and then decays with a polynomial scheduler. The warm-up rate is set to 1e-6, and the decay rate is 0.9. In the experiments conducted on the PASCAL VOC dataset, we set the batch size to 4 and the total number of iterations to 20k. On the MS COCO dataset, the network is trained for 80k iterations with a batch size of 8. To generate object-aware masks, we utilize the features of transformer's 10-th block in Sec. 3.3. The background threshold β is set at 0.5. We use (1.0, 1.0, 0.2) for the loss weight factors (λ_i, λ_e, λ_p) in Eq. (7). Additionally, we employ the data augmentation and multi-crop approaches described in [5]. Following common practices in semantic segmentation [6], we apply dense conditional random fields (CRF) processing and multi-scale testing during the inference stage.

Table 2: Comparison of the semantic segmentation results on the PASCAL VOC validation and testing sets. Sup. indicates the type of supervision ($\mathcal{I}$: image-level label; $\mathcal{S}$: saliency map; $\mathcal{L}$: image-level language). Net. denotes the backbone network (for single-stage methods) and the semantic segmentation network (for multi-stage methods). † denotes the use of ImageNet-21k [45] pre-trained parameters.

	Sup.	**Net.**	**Val**	**Test**
Multi-stage WSSS methods				
RIB [19]NeurIPS'21	$\mathcal{I}+\mathcal{S}$	DL-V2	70.2	70.0
EPS [32]CVPR'21	$\mathcal{I}+\mathcal{S}$	DL-V2	71.0	71.8
L2G [23]CVPR'22	$\mathcal{I}+\mathcal{S}$	DL-V2	72.1	71.7
RCA [69]CVPR'22	$\mathcal{I}+\mathcal{S}$	DL-V2	72.2	72.8
Du *et al.* [16]CVPR'22	$\mathcal{I}+\mathcal{S}$	DL-V2	72.6	73.6
ReCAM [12]CVPR'22	$\mathcal{I}$	DL-V2	68.4	68.2
VWL [47]IJCV'22	$\mathcal{I}$	DL-V2	69.2	69.2
W-OoD [30]CVPR'22	$\mathcal{I}$	WR38	70.7	70.1
MCTformer [62]CVPR'22	$\mathcal{I}$	WR38	71.9	71.6
ESOL [34]NeurIPS'22	$\mathcal{I}$	DL-V2	69.9	69.3
LPCAM [11]CVPR'23	$\mathcal{I}$	DL-V2	70.1	70.4
FPR [7]ICCV'23	$\mathcal{I}$	DL-V2	70.3	70.1
CLIMS [59]CVPR'22	$\mathcal{I}+\mathcal{L}$	DL-V2	69.3	68.2
CLIP-ES [38]CVPR'23	$\mathcal{I}+\mathcal{L}$	DL-V2	71.1	71.4
MMCST [63]CVPR'23	$\mathcal{I}+\mathcal{L}$	WR38	72.2	72.2
Single-stage WSSS methods				
RRM [67]AAAI'20	$\mathcal{I}$	WR38	62.6	62.9
1Stage [4]CVPR'20	$\mathcal{I}$	WR38	62.7	64.3
AFA [48]CVPR'22	$\mathcal{I}$	MiT-B1	66.0	66.3
SLRNet [42]IJCV'22	$\mathcal{I}$	WR38	67.2	67.6
ViT-PCM [46]ECCV'22	$\mathcal{I}$	ViT-B	70.3	70.9
ToCo [49]CVPR'23	$\mathcal{I}$	ViT-B†	71.1	72.2
Ours	$\mathcal{I}+\mathcal{L}$	ViT-B	**71.4**	**71.4**
Ours†	$\mathcal{I}+\mathcal{L}$	ViT-B†	**74.5**	**74.9**

4.3 Experimental Results

Quality of Pseudo Labels. We visualize CAMs using the proposed method, which activates multiple objects within the image, as shown in Fig. 3. Our approach offers more precise object localization across the entire object region compared to recent methods like CLIMS [59] and CLIP-ES [38], which rely on text guidance. We conduct a quantitative evaluation of the training and validation sets of the PASCAL VOC dataset. We compare our DALNet with existing WSSS methods in Tab. 1. To ensure a fair comparison, we include results using pre-trained ImageNet-1k weights (DeiT [51]), considering the initial pre-training of ViT on ImageNet-21k. Our findings demonstrate that our method generates CAMs with mIoU comparable to or exceeding those of recent methods.

Semantic Segmentation Results. The semantic segmentation outcomes are detailed in Tab. 2 and Tab. 3 from experiments conducted on the PASCAL

Table 3: Comparison of the semantic segmentation results on the MS COCO [37] validation set. Net. denotes the backbone network (for single-stage methods) and the semantic segmentation network (for multi-stage methods). † indicates the use of ImageNet-21k [45] pre-trained parameters.

	Net.	**Val**		**Net.**	**Val**
Multi-stage WSSS methods			**Single-stage WSSS methods**		
SEAM [54]$_{\text{CVPR'20}}$	DL-V1	31.9	AFA [48]$_{\text{CVPR'22}}$	MiT-B1	38.9
RIB [19]$_{\text{NeurIPS'21}}$	DL-V2	43.8	SLRNet [42]$_{\text{ECCV'22}}$	WR38	35.0
RCA [69]$_{\text{CVPR'22}}$	DL-V2	36.8	ToCo [49]$_{\text{CVPR'23}}$	ViT-B†	42.3
SIPE [8]$_{\text{CVPR'22}}$	DL-V2	40.6	Ours	ViT-B	**42.7**
MCTformer [62]$_{\text{CVPR'22}}$	WR38	42.0	Ours†	ViT-B†	**44.4**

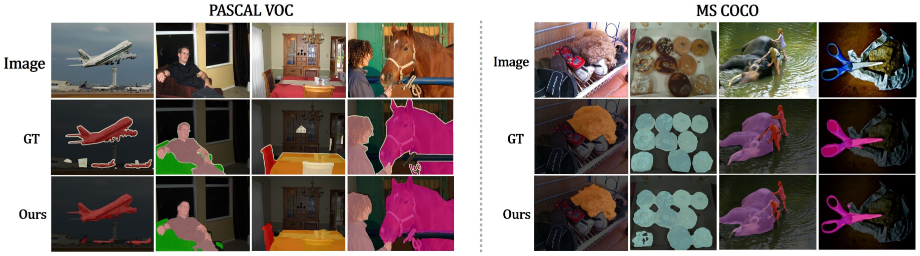

Fig. 4: Visualization results of semantic segmentation on the PASCAL VOC and MS COCO datasets.

VOC and MS COCO datasets. Notably, our proposed method significantly outperforms on the PASCAL VOC dataset, achieving mIoU of 74.5% and 74.9% on the validation and testing sets, respectively. Furthermore, we achieve a mIoU of 44.4% on the MS COCO validation dataset, demonstrating effective performance compared to multi-stage methods and superior results compared to single-stage methods. We emphasize that our study is dedicated to advancing single-stage WSSS with a unified objective. As shown in Fig. 4, the visualized semantic segmentation masks demonstrate the activation of various objects, providing a comprehensive understanding of the images. This suggests that incorporating textual information as guidance enhances object localization.

4.4 Ablation and Analysis

Configurations of Loss Function. In our ablation study, we explore the impact of the proposed method. The results are detailed in Tab. 4, including the results of initial CAMs on the PASCAL VOC validation set. For dense alignment between images and texts, we employ $\mathcal{L}_{im}$ (GIA) and $\mathcal{L}_{ex}$ (LEA). Without cross-contrastive learning, the mIoU of CAMs is 64.1% when only $\mathcal{L}_{im}$ is applied. Applying only $\mathcal{L}_{ex}$ results in CAMs mIoU of 65.1%. Combining $\mathcal{L}_{im}$ and $\mathcal{L}_{ex}$ significantly improves the mIoU of CAMs to 66.9%, indicating the best performance. As shown in Fig. 5 (Left), relying solely on global ($\mathcal{L}_{im}$) or local ($\mathcal{L}_{ex}$)

Table 4: Comparison of initial CAMs (mIoU) with different loss function configurations and cross-contrastive learning (CCL) on the PASCAL VOC validation set.

Method	$\mathcal{L}_{im}$**(GIA)**	$\mathcal{L}_{ex}$**(LEA)**	**w/o CCL**	**w/ CCL**
ViT	✓		64.1	67.9
		✓	65.1	68.3
	✓	✓	66.9	70.2

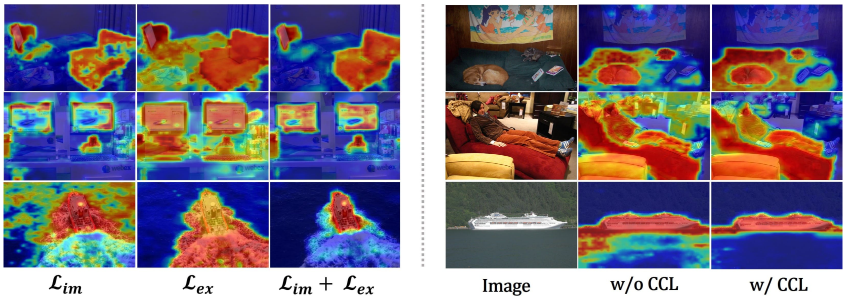

Fig. 5: (Left) Visualization of initial CAMs with different loss function configurations. **(Right)** Visualization results for the CAMs on the PASCAL VOC dataset without and with cross-contrastive learning (CCL).

information either fails to activate the integral object regions or results in over-activation. Our dense alignment approach ($\mathcal{L}_{im}+\mathcal{L}_{ex}$) effectively enhances dense localization attained by integrating global and local information.

Analysis of Cross-contrastive Learning. We facilitate cross-contrastive learning between visual features and text embeddings. In Tab. 4, we compare the performance of initial CAMs using cross-contrastive learning (CCL). CCL pairs visual foreground with target class embeddings as positives and background text embeddings as negatives. Simultaneously, it pairs the visual background with background text embeddings as positives and target class embeddings as negatives. Using only the foreground as positives (w/o CCL), the mIoU of the CAMs is 66.9%. By incorporating the foreground and background (w/ CCL), the mIoU of the CAMs increases to 70.2%. These results demonstrate the effectiveness of our cross-contrastive learning approach. It indicates that incorporating background visual features and text embeddings can enhance performance by addressing missing regions and mitigating distractions, ultimately improving focus on the target object. By utilizing specific classes and the background prompts, we align the background text embedding with areas outside the foreground in images. The "`background`" term is broadly defined to encompass diverse elements beyond image objects. Notably, it can produce promising outcomes by eliminating the need for specific background categorization and additional prior efforts. Additionally, Fig. 5 (Right) presents a visualization of CAMs results, demonstrating the effectiveness of our proposed method in suppressing over-activated background areas and compensating for the under-activation in object

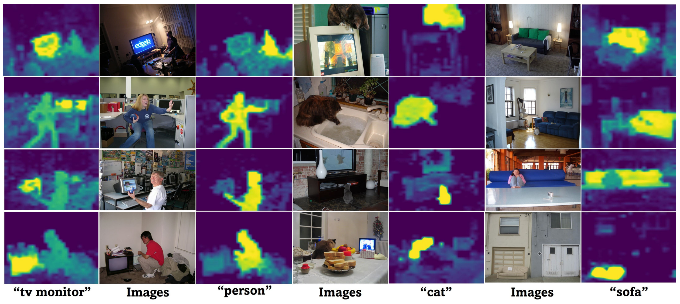

Fig. 6: Visualization of similarity coherence between image and text representations for each class.

areas within the CAMs.

Similarity Coherence. To validate the effectiveness of DALNet, which combines GIA and LEA, we present similarity coherence between image patch tokens and categorical text embeddings. Visualization results using the PASCAL VOC training set are shown in Fig. 6. Our proposed method enables the identification of specific regions based on the target class embedding. Specifically, local contrastive learning with object-ware mask can enhance the similarity of patch tokens corresponding to the target class text embedding. Our focus on WSSS leverages the categories defined within the dataset. However, there are approaches in VLP that use open-vocabulary categories to learn the similarity coherence between text and image [40, 55, 61]. Unlike previous studies that depend on additional annotation masks or fine-tuning modules, learning processes based solely on text embeddings and image features provide insights for improving generalization in our future exploration.

5 Conclusion

In this paper, we address the challenges of discriminative CAMs and WSSS that rely on image-level labels. We present a novel approach, Dense Alignment Learning Network (DALNet), which leverages text guidance at different levels of granularity. DALNet consists of two strategies: Global Implicit Alignment (GIA) for semantic information and Local Explicit Alignment (LEA) for spatial details. Based on this dual-level alignment strategy, DALNet aligns foregrounds across modalities and distinguishes them from their backgrounds through cross-contrastive learning. Extensive experiments show that DALNet improves dense localization by combining global and local information. Furthermore, we demonstrate that by utilizing straightforward prompts with specific class labels and the "`background`" term, DALNet enhances focus on target objects by simultaneously activating missing regions and suppressing distractions. Our experimental findings validate its effectiveness, achieving outstanding performance on the PASCAL VOC and MS COCO datasets, even as a single-stage WSSS framework.

Acknowledgements

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2022R1C1C1008534, Contribution Rate: 25%); the Institute for Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2021-0-01341, Artificial Intelligence Graduate School Program, Chung-Ang University, Contribution Rate: 25%); and the Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2024 (Project Name: Developing Professionals for R&D in Contents Production Based on Generative AI and Cloud, Project Number: RS-2024-00352578, Contribution Rate: 50%).

References

1. Ahn, J., Cho, S., Kwak, S.: Weakly supervised learning of instance segmentation with inter-pixel relations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2209–2218 (2019)
2. Ahn, J., Kwak, S.: 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. pp. 4981–4990 (2018)
3. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems **35**, 23716–23736 (2022)
4. Araslanov, N., Roth, S.: Single-stage semantic segmentation from image labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4253–4262 (2020)
5. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
6. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: 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)
7. Chen, L., Lei, C., Li, R., Li, S., Zhang, Z., Zhang, L.: Fpr: False positive rectification for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1108–1118 (2023)
8. Chen, Q., Yang, L., Lai, J.H., Xie, X.: Self-supervised image-specific prototype exploration for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4288–4298 (2022)
9. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: Simclr: A simple framework for contrastive learning of visual representations. In: International Conference on Learning Representations. vol. 2 (2020)
10. Chen, Y., Yuan, J., Tian, Y., Geng, S., Li, X., Zhou, D., Metaxas, D.N., Yang, H.: Revisiting multimodal representation in contrastive learning: from patch and token

embeddings to finite discrete tokens. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15095–15104 (2023)

11. Chen, Z., Sun, Q.: Extracting class activation maps from non-discriminative features as well. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3135–3144 (2023)
12. Chen, Z., Wang, T., Wu, X., Hua, X.S., Zhang, H., Sun, Q.: Class re-activation maps for weakly-supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 969–978 (2022)
13. Desai, K., Johnson, J.: Virtex: Learning visual representations from textual annotations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11162–11173 (2021)
14. Ding, X., Zhang, X., Han, J., Ding, G.: Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11963–11975 (2022)
15. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
16. Du, Y., Fu, Z., Liu, Q., Wang, Y.: Weakly supervised semantic segmentation by pixel-to-prototype contrast. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4320–4329 (2022)
17. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**, 303–338 (2010)
18. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems **33**, 21271–21284 (2020)
19. Hariharan, B., Arbeláez, P., Bourdev, L., Maji, S., Malik, J.: Semantic contours from inverse detectors. In: 2011 international conference on computer vision. pp. 991–998. IEEE (2011)
20. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
21. Jang, S., Kwon, J., Jin, K., Kim, Y.: Weakly supervised semantic segmentation via graph recalibration with scaling weight unit. Engineering Applications of Artificial Intelligence **119**, 105706 (2023)
22. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021)
23. Jiang, P.T., Yang, Y., Hou, Q., Wei, Y.: L2g: A simple local-to-global knowledge transfer framework for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16886–16896 (2022)
24. Jiang, Z.H., Hou, Q., Yuan, L., Zhou, D., Shi, Y., Jin, X., Wang, A., Feng, J.: All tokens matter: Token labeling for training better vision transformers. Advances in neural information processing systems **34**, 18590–18602 (2021)

25. Kim, B., Han, S., Kim, J.: Discriminative region suppression for weakly-supervised semantic segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 1754–1761 (2021)
26. Kim, B.J., Choi, H., Jang, H., Lee, D.G., Jeong, W., Kim, S.W.: Dead pixel test using effective receptive field. Pattern Recognition Letters **167**, 149–156 (2023)
27. Kweon, H., Yoon, S.H., Yoon, K.J.: Weakly supervised semantic segmentation via adversarial learning of classifier and reconstructor. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11329–11339 (2023)
28. Kwon, J., Lee, E., Cho, Y., Kim, Y.: Learning to detour: Shortcut mitigating augmentation for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 819–828 (2024)
29. Lee, J., Kim, E., Yoon, S.: Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4071–4080 (2021)
30. Lee, J., Oh, S.J., Yun, S., Choe, J., Kim, E., Yoon, S.: Weakly supervised semantic segmentation using out-of-distribution data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16897–16906 (2022)
31. Lee, M., Kim, D., Shim, H.: 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. pp. 4330–4339 (2022)
32. Lee, S., Lee, M., Lee, J., Shim, H.: 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. pp. 5495–5505 (2021)
33. Li, B., Weinberger, K.Q., Belongie, S., Koltun, V., Ranftl, R.: Language-driven semantic segmentation. arXiv preprint arXiv:2201.03546 (2022)
34. Li, J., Jie, Z., Wang, X., Wei, X., Ma, L.: Expansion and shrinkage of localization for weakly-supervised semantic segmentation. Advances in Neural Information Processing Systems **35**, 16037–16051 (2022)
35. Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W.: Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557 (2019)
36. Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., Zhang, H., Zhang, P., Vajda, P., Marculescu, D.: Open-vocabulary semantic segmentation with mask-adapted clip. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7061–7070 (2023)
37. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
38. Lin, Y., Chen, M., Wang, W., Wu, B., Li, K., Lin, B., Liu, H., He, X.: Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15305–15314 (2023)
39. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in neural information processing systems **32** (2019)

40. Mukhoti, J., Lin, T.Y., Poursaeed, O., Wang, R., Shah, A., Torr, P.H., Lim, S.N.: Open vocabulary semantic segmentation with patch aligned contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19413–19423 (2023)
41. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
42. Pan, J., Zhu, P., Zhang, K., Cao, B., Wang, Y., Zhang, D., Han, J., Hu, Q.: Learning self-supervised low-rank network for single-stage weakly and semi-supervised semantic segmentation. International Journal of Computer Vision **130**(5), 1181–1195 (2022)
43. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
44. Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., Dosovitskiy, A.: Do vision transformers see like convolutional neural networks? Advances in Neural Information Processing Systems **34**, 12116–12128 (2021)
45. Ridnik, T., Ben-Baruch, E., Noy, A., Zelnik-Manor, L.: Imagenet-21k pretraining for the masses. arXiv preprint arXiv:2104.10972 (2021)
46. Rossetti, S., Zappia, D., Sanzari, M., Schaerf, M., Pirri, F.: Max pooling with vision transformers reconciles class and shape in weakly supervised semantic segmentation. In: European Conference on Computer Vision. pp. 446–463. Springer (2022)
47. Ru, L., Du, B., Zhan, Y., Wu, C.: Weakly-supervised semantic segmentation with visual words learning and hybrid pooling. International Journal of Computer Vision **130**(4), 1127–1144 (2022)
48. Ru, L., Zhan, Y., Yu, B., Du, B.: 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. pp. 16846–16855 (2022)
49. Ru, L., Zheng, H., Zhan, Y., Du, B.: Token contrast for weakly-supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3093–3102 (2023)
50. Tang, M., Perazzi, F., Djelouah, A., Ben Ayed, I., Schroers, C., Boykov, Y.: On regularized losses for weakly-supervised cnn segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 507–522 (2018)
51. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International conference on machine learning. pp. 10347–10357. PMLR (2021)
52. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
53. Veit, A., Wilber, M.J., Belongie, S.: Residual networks behave like ensembles of relatively shallow networks. Advances in neural information processing systems **29** (2016)
54. Wang, Y., Zhang, J., Kan, M., Shan, S., Chen, X.: Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12275–12284 (2020)

55. Wang, Z., Lu, Y., Li, Q., Tao, X., Guo, Y., Gong, M., Liu, T.: Cris: Clip-driven referring image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11686–11695 (2022)
56. Wei, Y., Xiao, H., Shi, H., Jie, Z., Feng, J., Huang, T.S.: 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. pp. 7268–7277 (2018)
57. Wu, Z., Shen, C., Van Den Hengel, A.: Wider or deeper: Revisiting the resnet model for visual recognition. Pattern recognition **90**, 119–133 (2019)
58. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in Neural Information Processing Systems **34**, 12077–12090 (2021)
59. Xie, J., Hou, X., Ye, K., Shen, L.: Clims: Cross language image matching for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4483–4492 (2022)
60. Xie, J., Xiang, J., Chen, J., Hou, X., Zhao, X., Shen, L.: 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. pp. 989–998 (2022)
61. Xu, J., De Mello, S., Liu, S., Byeon, W., Breuel, T., Kautz, J., Wang, X.: Groupvit: Semantic segmentation emerges from text supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18134–18144 (2022)
62. Xu, L., Ouyang, W., Bennamoun, M., Boussaid, F., Xu, D.: Multi-class token transformer for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4310–4319 (2022)
63. Xu, L., Ouyang, W., Bennamoun, M., Boussaid, F., Xu, D.: Learning multi-modal class-specific tokens for weakly supervised dense object localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19596–19605 (2023)
64. Yang, J., Duan, J., Tran, S., Xu, Y., Chanda, S., Chen, L., Zeng, B., Chilimbi, T., Huang, J.: Vision-language pre-training with triple contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15671–15680 (2022)
65. Yi, M., Cui, Q., Wu, H., Yang, C., Yoshie, O., Lu, H.: A simple framework for text-supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7071–7080 (2023)
66. Yun, S., Park, S.H., Seo, P.H., Shin, J.: Ifseg: Image-free semantic segmentation via vision-language model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2967–2977 (2023)
67. Zhang, B., Xiao, J., Wei, Y., Sun, M., Huang, K.: Reliability does matter: An end-to-end weakly supervised semantic segmentation approach. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 12765–12772 (2020)
68. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2921–2929 (2016)
69. Zhou, T., Zhang, M., Zhao, F., Li, J.: Regional semantic contrast and aggregation for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4299–4309 (2022)