Supplementary Material

A Impact of the Hyperparameters

Table 5: Impact of hyperparameters. The results are evaluated on PASCAL VOC validation set.

Temp. factor τ	**CAM**	**Seg**
0.2	69.6	69.2
0.3	68.8	66.6
0.5	68.9	68.9
0.7	69.5	68.0
1.0	**70.2**	**71.4**

(a) Temperature factor.

Background β	**CAM**	**Seg**
0.3	67.0	68.7
0.4	68.3	69.8
0.5	**70.2**	**71.4**
0.6	68.5	70.1
0.7	67.8	68.9

(b) Background threshold.

Block	**CAM**	**Seg**
#7	59.8	57.6
#8	61.4	60.0
#9	65.2	65.2
#10	**70.2**	**71.4**
#11	54.9	54.0

(c) Intermediate block.

Temperature Factor. The experimental results, shown in Tab. 5a, depend on the temperature factor τ in Eq. (6). The outcomes of the experiment show that optimal results are obtained when τ is configured to 1.0.
Background Threshold. In Tab. 5b, we present the findings on the impact of the background thresholds. The parameter β is utilized to generate the intermediate masks and CAMs. Our findings indicate that the best semantic segmentation results are obtained when β is 0.5.
Intermediate Block. The influence of extracting intermediate features from different blocks is depicted in Tab. 5c. While deep blocks in the model include high-level semantics and extract features corresponding to specific object details, shallow blocks in the model primarily capture low-level semantics and extract more general features. Experimental results indicate that the 10-th block produces the most optimal results for semantic segmentation.

Table 6: Impact of the different loss weights. The results are evaluated on the PASCAL VOC validation set.

λ	**0.5**	**0.1**	**0.01**	**0.005**	**0.001**
Seg	68.3	67.2	67.8	71.1	**71.4**
λ_i	**0.2**	**0.4**	**0.5**	**0.7**	**1.0**
Seg	66.3	67.9	66.5	66.8	**71.4**
λ_e	**0.2**	**0.4**	**0.5**	**0.7**	**1.0**
Seg	70.2	68.2	68.8	70.4	**71.4**

Loss Weights. As detailed in Tab. 6, we show the segmentation results using different weight factors for the loss terms on the PASCAL VOC validation set. While λ_i and λ_e are hyperparameters associated with the global implicit and local explicit alignments in the total loss Eq. (7), the hyperparameter λ in Eq. (6) is related to the background. The optimal combination for achieving the best semantic segmentation results is $(\lambda = 0.001, \lambda_i = 1.0, \lambda_e = 1.0)$.

B Further Analysis

Table 7: Efficiency comparison on PASCAL VOC with RTX 3090.

Method	CAM	Refine	Decoder	Val	Test
Multi-stage WSSS method					
CLIMS [59]	101 mins	332 mins	635 mins	70.4	70.0
Single-stage WSSS method					
AFA [48]		554 mins		66.0	66.3
ToCo [49]		506 mins		71.1	72.2
Ours		444 mins		**74.5**	**74.9**

Fig. 7: Visualization of similarity coherence between image features and text embedding on PASCAL VOC validation set by class, including the background class.

Training Efficiency. Our DALNet is a single-stage framework that allows for end-to-end learning, as compared to multi-stage methods. We compare the training efficiency of our proposed method, and the results are presented in Tab. 7. CLIMS [59], which utilizes text supervision, is divided into three parts. This requires training multiple models for different purposes, thus making the training

pipeline complex and demanding more computational resources. In comparison to CLIMS, DALNet takes 444 minutes to finish training and performs significantly more effectively in dense localization. Furthermore, as compared to other existing single-stage frameworks [48, 49], we observe improved training efficiency and performance.

Similarity Coherence. By using both the specified class prompts and the "`background`" prompt, we align the background text embedding with areas outside the foreground in images. Unlike previous methods that relied on external models or pre-defined terms, our approach efficiently represents the "`background`" prompt, which refers to areas outside image objects. As shown in Fig. 7, our results validate that background prompts effectively incorporate information from these areas and mitigate distractions. We observe that background coherence captures over-activated regions in the foreground as well as the boundary between objects and background in the image. This result indicates the effectiveness and outstanding performance of the proposed cross-contrastive learning approach.

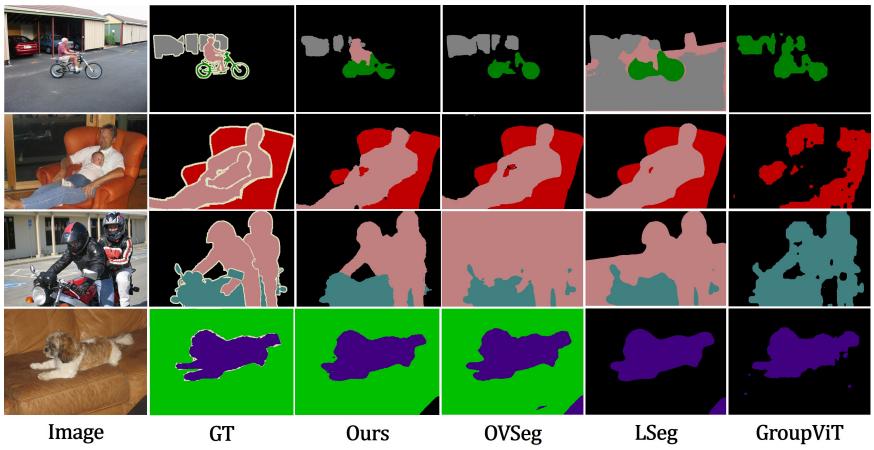

Fig. 8: Visualization of semantic segmentation masks on the PASCAL VOC validation set. We generated semantic masks using OVSeg [36], LSeg [33] and GroupViT [61] in open-vocabulary setting.

Comparison with Open-vocabulary based Methods. We focus on the WSSS task, using class categories defined in the dataset. Recently, the progress in vision-language pre-training has broadened the approaches for semantic segmentation to an open-vocabulary setting, encompassing diverse and intricate natural language expressions. However, these methods often depend on additional fine-tuning modules or require segmentation annotations and large external training datasets. Although it is not a completely fair comparison, we evaluate the se-

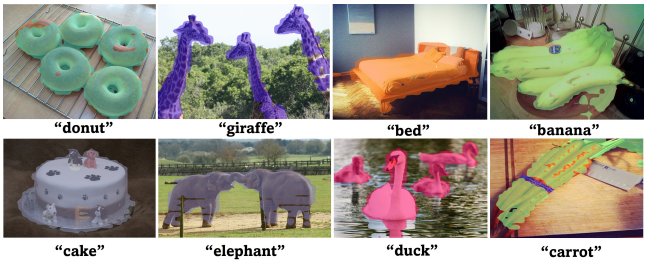

Fig. 9: Visualization of similarity masks on unseen category on the PASCAL VOC.

Table 8: The comparison to the fully-supervised counterparts on PASCAL VOC validation set.

Method	Backbone	Val	Ratio
DeepLabV2 [6]	ResNet101	77.7	-
WideResNet [57]	ResNet38	80.8	-
Segformer [58]	MiT-B1	78.7	-
DeepLabV2 [6]	ViT-B/16	82.3	-
Multi-stage WSSS method			
W-OoD [30]CVPR'22	ResNet38	70.7	87.5%
CLIMS [59]CVPR'22	ResNet101	69.3	89.2%
MCTformer [62]CVPR'22	ResNet38	71.9	89.0%
CLIP-ES [38]CVPR'23	ViT-B/16	71.1	86.4%
Single-stage WSSS method			
1Stage [4]CVPR'20	ResNet38	62.7	77.6%
AFA [48]CVPR'22	MiT-B1	66.0	83.9%
SLRNet [42]IJCV'22	ResNet38	69.3	85.8%
ViT-PCM [46]ECCV'22	ViT-B/16	70.3	85.4%
ToCo [49]CVPR'23	ViT-B/16	71.1	86.4%
Ours	ViT-B/16	**74.5**	**90.5%**

mantic segmentation results of these methods using the class labels from the PASCAL VOC dataset. The visualization of the semantic segmentation results is shown in Fig. 8. These findings demonstrate that our proposed DALNet effectively achieves dense alignment between text embeddings and image features. Furthermore, Fig. 9 shows the potential of DALNet for generating relevant object masks even for unseen categories, highlighting its generalization capability. For future exploration, we intend to investigate learning methods that rely solely on text embeddings and image features. This approach aims to reduce the reliance on additional annotation mask supervision or fine-tuning modules, thereby enhancing the model's generalization capabilities.

Fully-supervised Counterparts. Tab. 2 demonstrates that various WSSS methods employ different backbones. To ensure a fair comparison, we present their upper bound performance on the PASCAL VOC validation set, which is the performance of their fully-supervised counterparts, in Tab. 8. Our method

achieves 74.5 mIoU and 90.5% relative to its fully-supervised counterpart. These outcomes exceed the performance of previous single-stage WSSS methods using the ViT backbone, as well as multi-stage WSSS methods. Furthermore, in comparison to CLIMS [59] and CLIP-ES [38], which employ text supervision, our method proves to be more effective.

C Additional Quantitative/Qualitative Results

Table 9: Per-class segmentation results (mIoU) on PASCAL VOC validation set. † indicates the use of ImageNet-21k [45] pre-trained parameters.

Method	Bkg	Aero	Bicycle	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Table	Dog	Horse	Mbk	Person	Plant	Sheep	Sofa	Train	TV	mIoU
Ours	91.1	85.5	40.4	82.8	49.5	77.3	86.2	80.9	89.0	35.2	81.5	50.2	86.0	87.6	77.5	74.7	67.8	87.3	49.7	56.9	62.9	**71.4**
Ours†	92.4	86.5	44.7	84.9	71.5	73.0	86.4	82.1	87.7	42.2	89.0	58.5	84.5	86.1	75.4	81.8	64.1	87.7	56.8	63.4	65.4	**74.5**

We present additional similarity coherence for text embeddings and image features in Fig. 10. We note a high similarity between visual features corresponding to objects and foreground text embeddings. This observation emphasizes the efficacy of our proposed DALNet in achieving dense localization by capturing the significant similarity between relevant visual features and foreground text embeddings. Additionally, in Tab. 9, we present class-specific qualitative segmentation results for the PASCAL VOC dataset. We notice variations in segmentation performance among different classes. These differences can be attributed to limitations of the dataset, such as the presence of small, intricate objects or an imbalanced number of samples per-class. However, as shown in Fig. 11 and Fig. 12, our quantitative results demonstrate that our proposed method achieves effective segmentation outcomes across various classes.

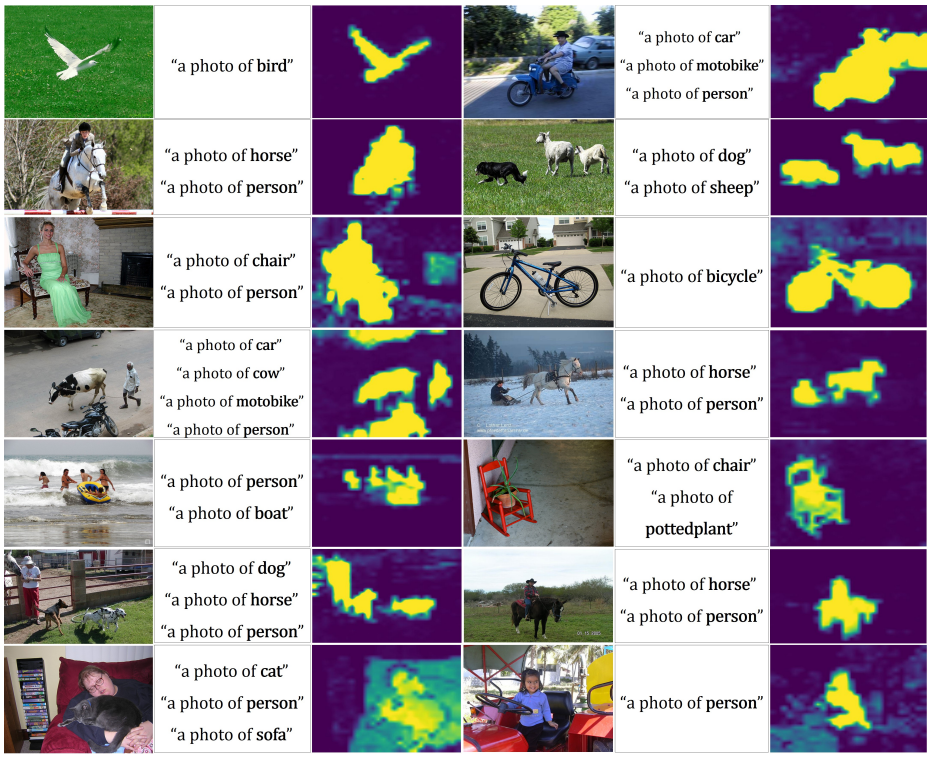

Fig. 10: Visualization of similarity coherence between image features and text embeddings on the PASCAL VOC validation set. The text prompt "a photo of `background`" is omitted for clarity.

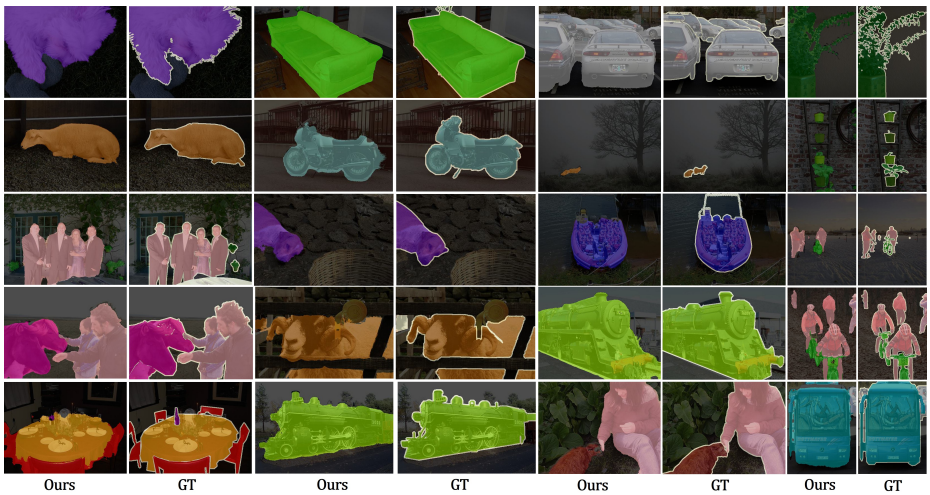

Fig. 11: Visualization results of semantic segmentation on the PASCAL VOC validation set.

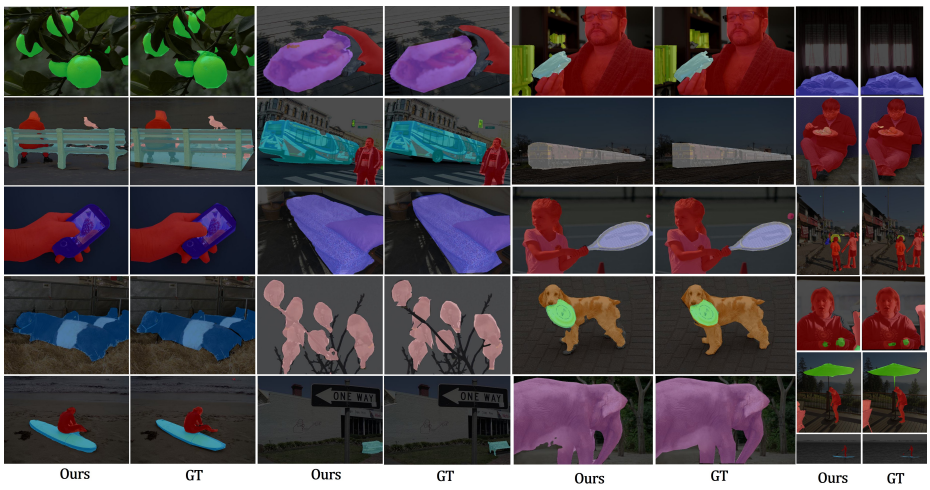

Fig. 12: Visualization results of semantic segmentation on the MS COCO validation set.