Exploring CLIP's Dense Knowledge for Weakly Supervised Semantic Segmentation

Zhiwei Yang[1,2] Yucong Meng[2,3]
Kexue Fu[4] Feilong Tang[1] Shuo Wang[2,3*] Zhijian Song[1,2,3*]
[1]Academy for Engineering and Technology, Fudan University, Shanghai 200433, China
[2]Shanghai Key Laboratory of Medical Image Computing and Computer Assisted Intervention
[3]Digital Medical Research Center, School of Basic Medical Sciences, Fudan University, China
[4]Shandong Computer Science Center (National Supercomputer Center in Jinan)

- Problem / objective
  - Weakly Supervised Semantic Segmentation (WSSS) via CLIP
- Contribution / Key idea
  - Text Semantic Enrichment module (**TSE**)
  - Visual Calibration module
    - Static Visual Calibration (**SVC**)
    - Learnable Visual Calibration (**LVC**)

Weakly Supervised Semantic Segmentation (WSSS)

Definition

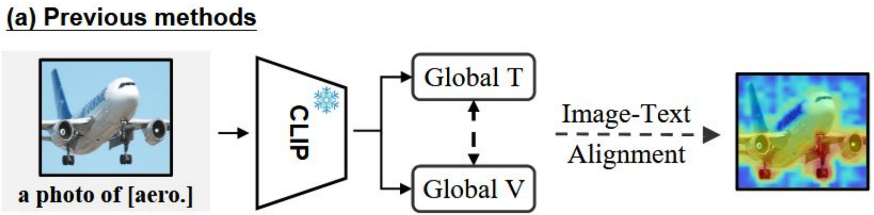
- Generate pixel-level predictions using weak annotations like points, scribbles, bounding boxes, or **image-level labels**

WSSS 3-stage Pipeline

1. Generate Class Activation Maps (**CAMs**) by training a classification network
2. Refine CAMs into pseudo labels (**PL**)
3. Use these labels to **train** a segmentation model

Motivation

- Current methods treating **'WSSS via CLIP'** primarily focus on CLIP's global **image-text alignment**, as shown in Fig. 1 (a). CLIP's dense knowledge with **patch-text alignment** still remains under-explored in WSSS.



- **Motivation**

❑ **Two key challenges:**

1. Semantic sparsity in **textual prompts**

   : The template 'a photo of [CLASS]' only indicates object presence but lacks knowledge for localization.

2. Fine-grained insufficiency in **visual features**

   : CLIP prioritizes global representation due to its image-text pairing nature.

❑ **Our proposed solution:**

1. Text Semantic Enrichment (**TSE**) module
2. Visual Calibration (**VC**) module

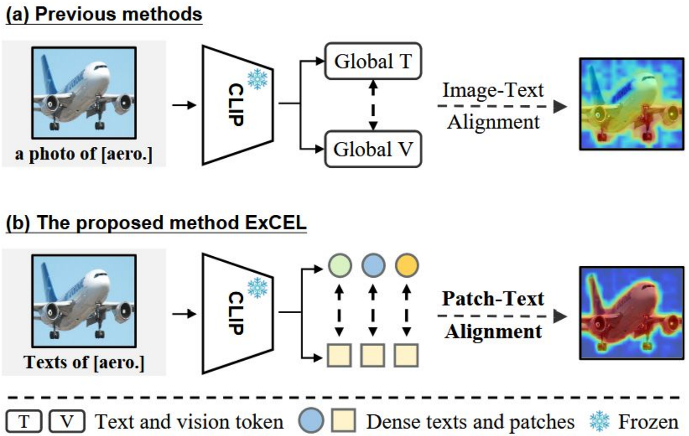


Figure 1. Our motivation. (a) Previous methods leverage CLIP to generate CAMs with global image-text alignment, leaving CLIP's dense knowledge unexplored. (b) The proposed ExCEL explores CLIP's dense knowledge via a novel patch-text alignment paradigm, which generates better CAMs with less training cost.

YANG, Zhiwei, et al. Exploring CLIP's Dense Knowledge for Weakly Supervised Semantic Segmentation. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025. p. 20223-20232.

CVPR 2025

- **Overview**

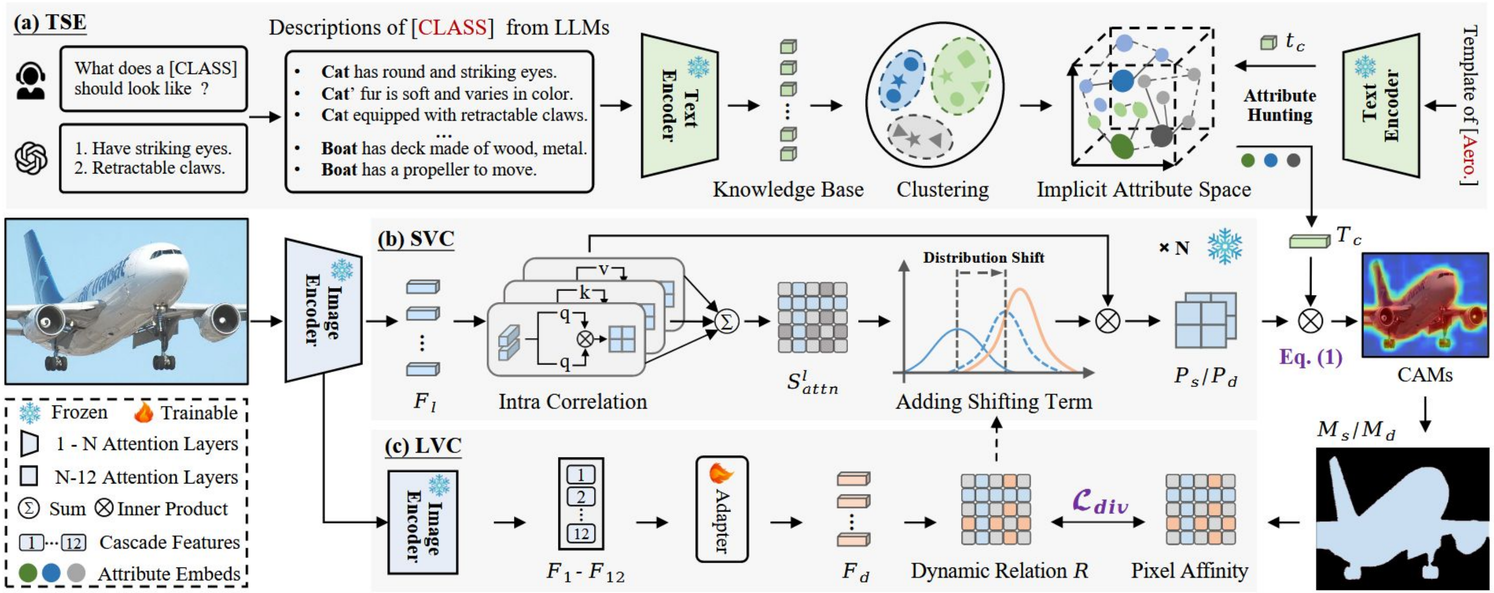


Figure 2. ExCEL Architecture. We explore CLIP's dense knowledge with Text Semantic Enrichment (TSE) and Visual Calibration (VC). (a) TSE uses LLMs to build a knowledge base and clusters it into an implicit attribute space. The final text representation T_c is enhanced by hunting for relevant attributes. For vision modality, (b) we introduce Static Visual Calibration (SVC) to calibrate visual features using the Inter-correlation operation across N intermediate layers. It generates static CAMs with T_c and calibrated features P_s. (c) Learnable Visual Calibration (LVC) designs a learnable adapter to add a dynamic shift R to SVC. It generates optimized features P_d based on static CAMs guidance, creating dynamic CAMs from P_d and T_c. Dynamic CAMs are refined for segmentation supervision. Details are in Sec. 3.1.

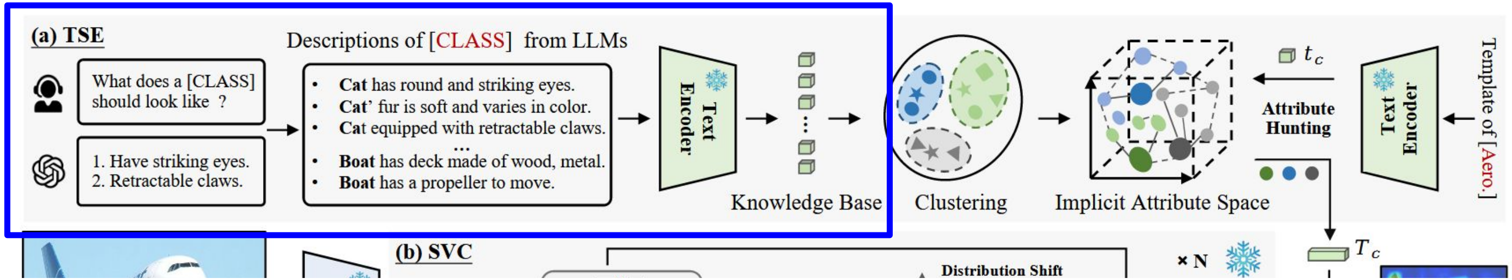
전유진

● Text Semantic Enrichment

❑ Knowledge Base Construction

- Global text template E_c: *'a clean origami of [CLASS]'*
- Instructions for GPT: *"List n descriptions with key properties to describe the [CLASS] in terms of appearance, color, shape, size, or material, etc. These descriptions will help visually distinguish the [CLASS] from other classes in the dataset. Each description should follow the format: 'a clean origami [CLASS]. it + descriptive contexts.'"*
- GPT generate n detailed descriptions for each class, which are subsequently encoded into a dataset-wide knowledge base with CLIP's text encoder.
- **Knowledge base**: $\mathcal{T} = \{\Phi(e_i)\}_{i=1}^{n\times C}$

Knowledge Base Construction



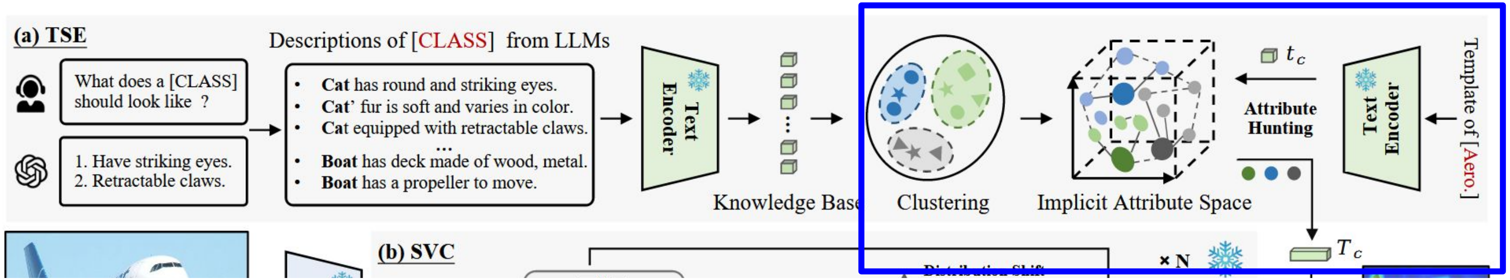
• Text Semantic Enrichment

❑ Implicit Attribute Hunting

- Cluster this knowledge into generalized attributes and treat text prompting as an implicit attribute-hunting process
- Each cluster centroid is viewed as the implicit attribute that represents a group of descriptions sharing similar properties
- Attribute feature space: $A = \mathrm{Kmeans}(\mathcal{T}, B) = \{a_i\}_{i=1}^{B}$, (2)
- Global text template, Global text embedding: E_c, $t_c \in \mathbb{R}^{D\times 1}$
- Top-K Attribute neighbors: $A_c = \{a_j : j \in \mathrm{argmax}_{\mathrm{TOPK}} \{t_c^T a_j\}_{j=1}^{B}\}$. (3)
- **Final text representation**:

$$T_c = t_c + \lambda \sum_{j=1}^{K} \mathrm{softmax}\left(t_c^T\ A_c\right) a_j, \quad (4)$$

Implicit Attribute Hunting

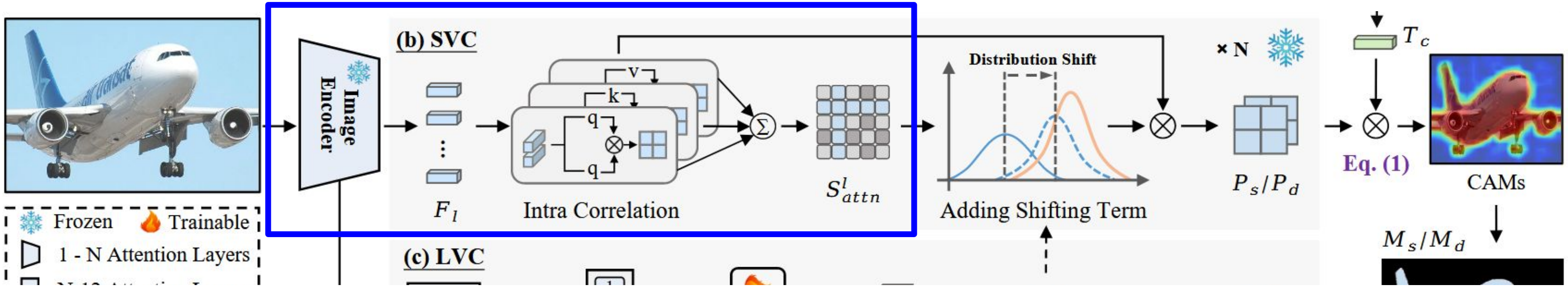


Visual Calibrations

Static Visual Calibration

- Input image, features from l-th layer of CLIP: $X \in \mathbb{R}^{3\times\mathcal{H}\times\mathcal{W}}, \quad F_l \in \mathbb{R}^{D_s\times hw}$
- Original attention map: $\mathrm{SA}(q,k) = \mathrm{sofmax}\left(q^T k/\sqrt{D_s}\right),$ (5)
- Limitation: The original q-k attention produces overly uniform attention maps, homogenizing diverse tokens from v to capture broad semantics for global image representation, due to the inherent image-text alignment of CLIP.
- Ours: **Intra-correlation** calculates the **attention within each space of {q, k, v} across intermediate layers**, instead of generating q-k correlation
- Attention map from l-th SVC layer: $S^l_{attn} = \sum w_i \,\mathrm{SA}\left(O^l_i, O^l_i\right), O^l_i \in \{q^l, k^l, v^l\},$ (6) $\quad S^l_{attn} \in \mathbb{R}^{hw\times hw} \quad l \in \{12-N, \ldots, 12\}$
- Calibrated features from the last layer of SVC: $P_s \in \mathbb{R}^{D\times h\times w}$
- **Static CAM** is generated by calibrated visual features from the last layer P_s and text embedding T_c: CAM_s

$$\mathrm{CAM} = \mathrm{Norm}\left(\cos\left(P, T\right)\right), \qquad (1)$$



유진

YANG, Zhiwei, et al. Exploring CLIP's Dense Knowledge for Weakly Supervised Semantic Segmentation. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025. p. 20223-20232.
CVPR 2025

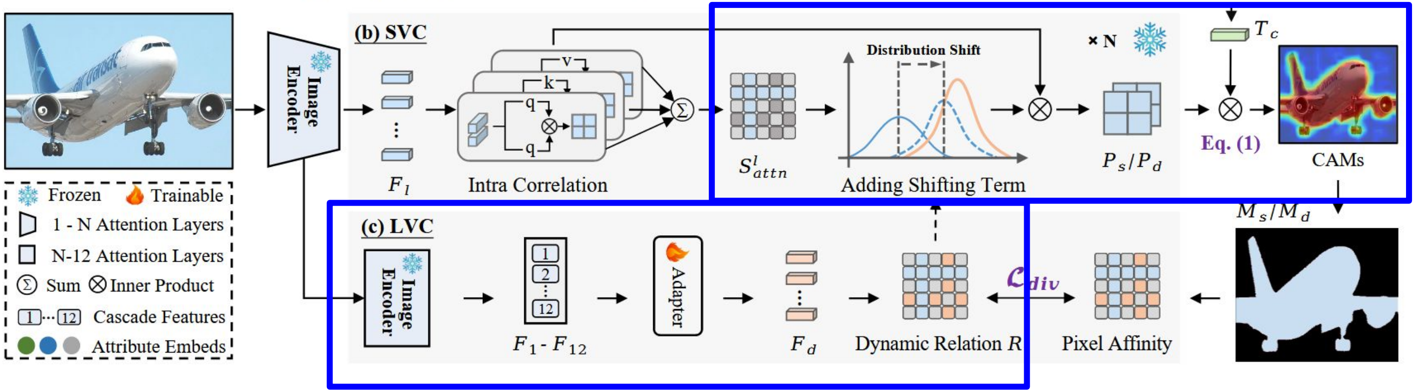
Visual Calibrations

Learnable Visual Calibration

- Limitation: Although ExCEL generates comparable CAMs without training, its performance is still limited by the fixed features in CLIP.
- Ours: We design a **lightweight adapter**, which only incorporates a distribution shift to calibrate the fixed features, to dynamically calibrate the visual features with diverse details.
- Frozen features from 1-12th layer of CLIP: $F_l \in \mathbb{R}^{D_s \times hw}$
- Dynamic feature: $F_d \in \mathbb{R}^{D_d \times hw}$ $\quad F_d = \mathrm{Conv}(\mathrm{Concate}\,[\delta_l\,(F_l)]_{l=1}^{12}), \quad (7)$
- Dynamic token relations: $r \in \mathbb{R}^{hw \times hw}$ $\quad r = \alpha(\cos(F_d, F_d) - \beta\overline{\cos(F_d, F_d)}), \quad (8)$
- Dynamic relations: $R \in \mathbb{R}^{hw \times hw}$

$$R_{ij} = \begin{cases} r_{ij}, & \text{if } r_{ij} \geq 0 \\ -inf, & \text{else} \end{cases}. \quad (9)$$

- Optimized attention map: $L_{attn}^l \in \mathbb{R}^{hw \times hw}$ $\quad L_{attn}^l = S_{attn}^l + \mathrm{softmax}(R). \quad (10)$
- Dynamically calibrated features from the last layer of LVC: $P_d \in \mathbb{R}^{D \times h \times w}$
- **Dynamic CAM**: $\mathrm{CAM} = \mathrm{Norm}\left(\cos\left(P, T\right)\right), \quad (1)$



전유진

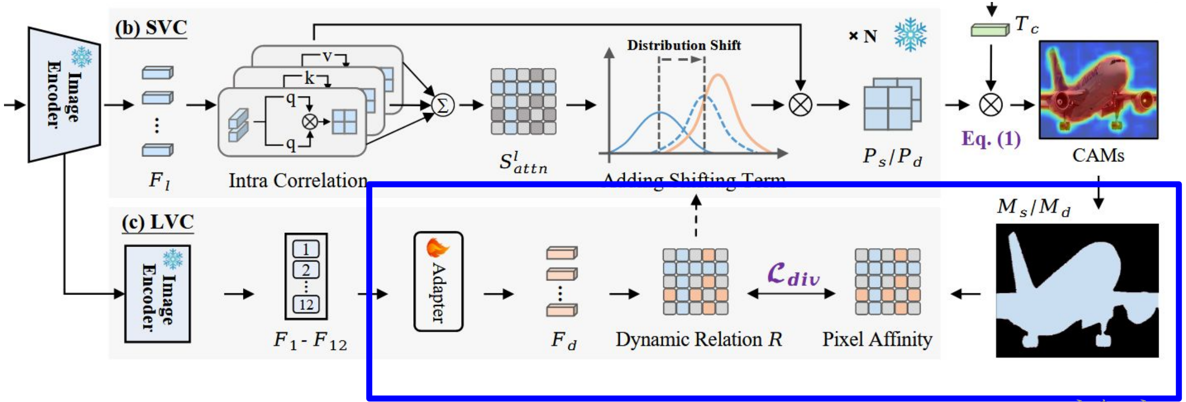
- **Training Objectives**

  - **Diversity Loss**
    - Objective: To supervise the learning of F_d in LVC module
    - Token correlations of F_d : $\hat{\mathcal{R}} \in \mathbb{R}^{hw \times hw}$ $\hat{\mathcal{R}} = \text{sigmoid}(\cos(F_d, F_d))$
    - Static pseudo-labels: M_s
    - **Diversity loss**: $$\mathcal{L}_{\text{div}} = \frac{1}{N^+} \sum_{u^+ \in \hat{\mathcal{R}}^+} (1 - u^+) + \frac{1}{N^-} \sum_{u^- \in \hat{\mathcal{R}}^-} u^-, \quad (11)$$

  - **Cross-Entropy Loss**
    - Objective: To supervise lightweight transformer-based segmentation head from WeCLIP [1]
    - Dynamic pseudo-labels: M_d
    - **Cross-entropy loss**: $\mathcal{L}_{seg}$

  - **Final Loss**
    - Adapter + Segmentation Head 학습

$$\mathcal{L}_{\text{ExCEL}} = \mathcal{L}_{seg} + \gamma \mathcal{L}_{\text{div}}, \quad (12)$$



[1] ZHANG, Bingfeng, et al. Frozen clip: A strong backbone for weakly supervised semantic segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024. p. 3796-3806.

YANG, Zhiwei, et al. Exploring CLIP's Dense Knowledge for Weakly Supervised Semantic Segmentation. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025. p. 20223-20232.

- **Experiments**

Table 1. Segmentation comparisons on VOC and COCO. Net. is the backbone for segmentation. Sup. is the supervision type. $\mathcal{I}$: image-level labels. $\mathcal{SA}$: saliency maps. $\mathcal{L}$: language.

Method	Sup.	Net.	VOC Val	VOC Test	COCO Val
Multi-stage WSSS methods.					
L2G [14] CVPR'2022	$\mathcal{I}+\mathcal{SA}$	RN101	72.1	71.7	44.2
RCA [49] CVPR'2023	$\mathcal{I}+\mathcal{SA}$	RN38	72.2	72.8	36.8
OCR [7] CVPR'2023	$\mathcal{I}$	RN38	72.7	72.0	42.5
BECO [26] CVPR'2023	$\mathcal{I}$	RN101	73.7	73.5	45.1
MCTformer+ [38] TPAMI'2024	$\mathcal{I}$	RN38	74.0	73.6	45.2
CTI [41] CVPR'2024	$\mathcal{I}$	RN101	74.1	73.2	45.4
CLIMS [36] CVPR'2022	$\mathcal{I}+\mathcal{L}$	RN101	70.4	70.0	-
CLIP-ES [20] CVPR'2023	$\mathcal{I}+\mathcal{L}$	RN101	72.2	72.8	45.4
PSDPM [45] CVPR'2024	$\mathcal{I}+\mathcal{L}$	RN101	74.1	74.9	47.2
CPAL [31] CVPR'2024	$\mathcal{I}+\mathcal{L}$	RN101	74.5	74.7	46.8
Single-stage WSSS methods.					
AFA [28] CVPR'2022	$\mathcal{I}$	MiT-B1	66.0	66.3	38.9
ViT-PCM [27] ECCV'2022	$\mathcal{I}$	ViT-B	70.3	70.9	-
ToCo [29] CVPR'2023	$\mathcal{I}$	ViT-B	71.1	72.2	42.3
DuPL [35] CVPR'2024	$\mathcal{I}$	ViT-B	73.3	72.8	44.6
SeCo [39] CVPR'2024	$\mathcal{I}$	ViT-B	74.0	73.8	46.7
DIAL [13] ECCV'2024	$\mathcal{I}+\mathcal{L}$	ViT-B	74.5	74.9	44.4
WeCLIP [43] CVPR'2024	$\mathcal{I}+\mathcal{L}$	ViT-B	76.4	77.2	47.1
ExCEL(w/o CRF)	$\mathcal{I}+\mathcal{L}$	**ViT-B**	**77.2**	**77.3**	**49.3**
ExCEL (Ours)	$\mathcal{I}+\mathcal{L}$	**ViT-B**	**78.4**	**78.5**	**50.3**

Table 2. CAM seed comparisons on VOC train set. $\mathcal{M}$: multi-stage methods. $\mathcal{S}$: single-stage methods. †: our reproduction following official codes. ExCEL*: ExCEL in a training-free manner.

Method	Type	Sup.	Net.	VOC Train
Training-free WSSS methods.				
CLIP-ES [20] CVPR'2023	$\mathcal{M}$	$\mathcal{I}+\mathcal{L}$	ViT-B	70.8
ExCEL* (Ours)	$\mathcal{S}$	$\mathcal{I}+\mathcal{L}$	**ViT-B**	**74.6**
Training-required WSSS methods.				
ReCAM [6] CVPR'2022	$\mathcal{M}$	$\mathcal{I}$	RN101	54.8
FPR [3] CVPR'2023	$\mathcal{M}$	$\mathcal{I}$	RN101	63.8
LPCAM [5] CVPR'2023	$\mathcal{M}$	$\mathcal{I}$	RN50	65.3
MCTformer+ [38] TPAMI'2024	$\mathcal{M}$	$\mathcal{I}$	RN38	68.8
SFC [44] AAAI'2024	$\mathcal{M}$	$\mathcal{I}$	RN101	64.7
CTI [41] CVPR'2024	$\mathcal{M}$	$\mathcal{I}$	RN101	69.5
AFA [28] CVPR'2022	$\mathcal{S}$	$\mathcal{I}$	MiT-B1	65.0
ViT-PCM [27] ECCV'2022	$\mathcal{S}$	$\mathcal{I}$	ViT-B	67.7
†ToCo [29] CVPR'2023	$\mathcal{S}$	$\mathcal{I}$	ViT-B	71.6
†DuPL [35] CVPR'2024	$\mathcal{S}$	$\mathcal{I}$	ViT-B	75.0
SeCo [39] CVPR'2024	$\mathcal{S}$	$\mathcal{I}$	ViT-B	74.8
CLIMS [36] CVPR'2022	$\mathcal{M}$	$\mathcal{I}+\mathcal{L}$	RN101	56.6
POLE [22] WACV'2023	$\mathcal{M}$	$\mathcal{I}+\mathcal{L}$	RN50	59.0
CPAL [31] CVPR'2024	$\mathcal{M}$	$\mathcal{I}+\mathcal{L}$	RN101	71.9
DIAL [13] ECCV'2024	$\mathcal{S}$	$\mathcal{I}+\mathcal{L}$	ViT-B	75.2
†WeCLIP [43] CVPR'2024	$\mathcal{S}$	$\mathcal{I}+\mathcal{L}$	ViT-B	75.4
ExCEL (Ours)	$\mathcal{S}$	$\mathcal{I}+\mathcal{L}$	**ViT-B**	**78.0**

전유진

YANG, Zhiwei, et al. Exploring CLIP's Dense Knowledge for Weakly Supervised Semantic Segmentation. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025. p. 20223-20232.

- **Experiments**

Table 3. Ablation study of ExCEL on VOC val set.

Conditions	SVC	TSE	LVC	Precision	Recall	mIoU
Baseline (CLIP)				18.8	21.3	12.1
w/ SVC	✓			81.2	86.2	72.5
w/o LVC	✓	✓		80.7	89.8	74.7
w/o TSE	✓		✓	83.7	86.3	75.1
ExCEL	✓	✓	✓	**85.0**	**88.4**	**77.2**

Table 4. Ablation study of attribute number B on VOC val set.

Number of Attr	None	32	64	**112**	144	196
mIoU	75.1	75.8	76.2	**77.2**	77.0	76.5

Table 5. Ablation study of VC module on VOC train set.

Conditions	q-k	v	I.C.	M.C.	LVC	Precision	Recall	mIoU
Baseline (CLIP)	✓					18.0	21.8	11.2
MaskCLIP		✓				77.1	80.9	65.8
w/ I.C.			✓			79.1	84.7	69.7
SVC			✓	✓		82.2	88.2	74.6
ExCEL			✓	✓	✓	**86.6**	**87.9**	**78.0**

- **Experiments**

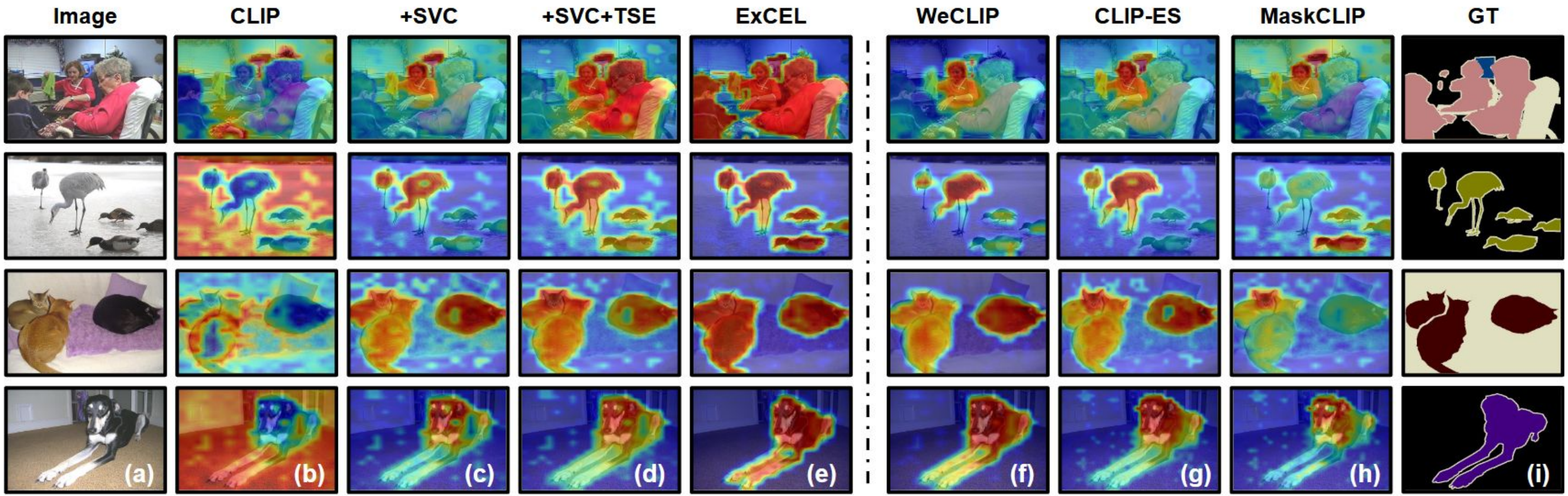


Figure 4. CAM visualizations on VOC train set. (a) Image. (b-e) Ablative visualizations of proposed modules. (e-h) Qualitative comparisons of (e) ExCEL and recent CLIP-based methods, i.e., (f) WeCLIP [43], (g) CLIP-ES [20] and (h) MaskCLIP [47]. (i) Ground truth.

- **Experiments**

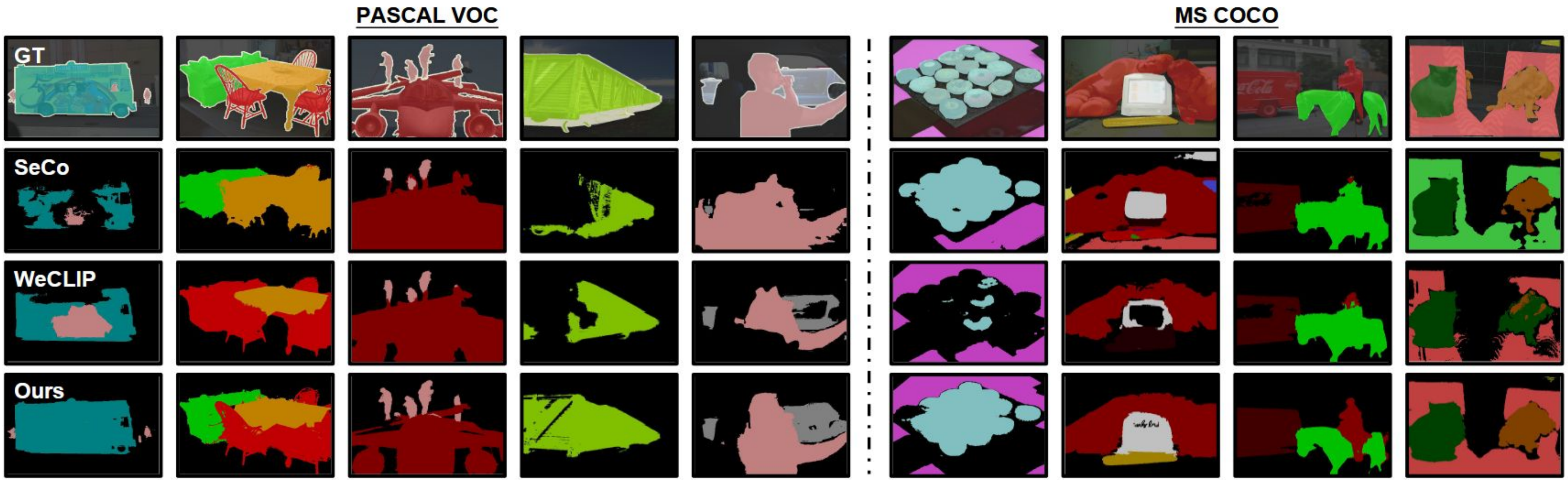


Figure 3. Segmentation visualizations of SeCo [39], WeCLIP [43] and ours on VOC and COCO. ExCEL segments objects more precisely.