Effective SAM Combination for Open-Vocabulary Semantic Segmentation

Minhyeok Lee[1] Suhwan Cho[1] Jungho Lee[1] Sunghun Yang[1]
Heeseung Choi[2] Ig-Jae Kim[2] Sangyoun Lee[1]
[1]Yonsei University
[2]Korea Institute of Science and Technology (KIST)
{hydragon516, chosuhwan, 2015142131, sunghun98, syleee}@yonsei.ac.kr
{hschoi, drjay}@kist.re.kr

- Problem / objective
  - Open-vocabulary semantic segmentation
- Contribution / Key idea
  - **ESC-NET:** Effective SAM Combination NETwork
    - One-stage open-vocabulary semantic segmentation model

- ## **Limitations of previous research**

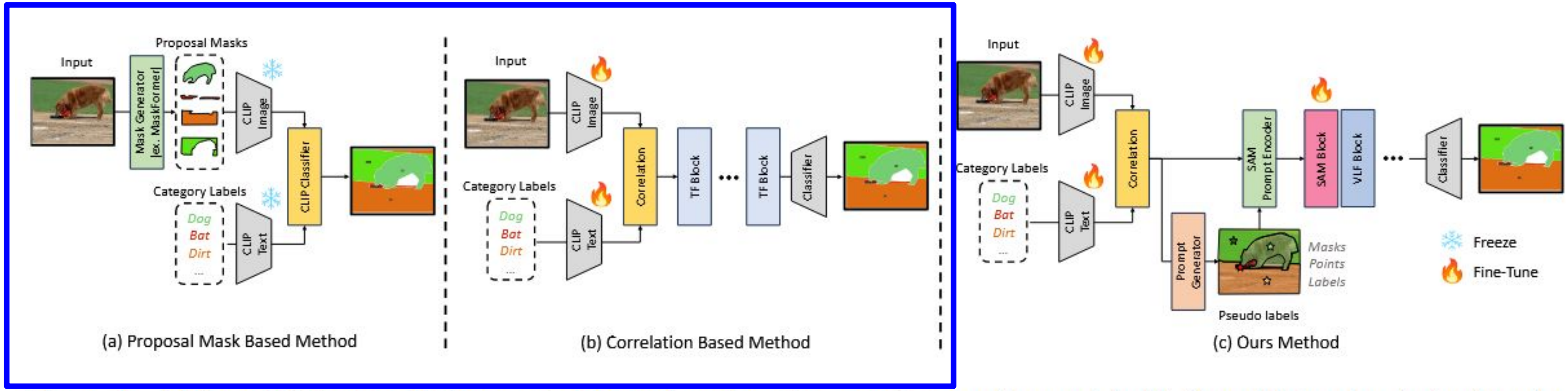

Figure 1. (a) A model structure that generates proposal masks using a mask generation model. (b) A model structure that refines the correlation between image and text. (c) The structure of the proposed ESC-Net. Our ESC-Net efficiently models the relationship between images and text by combining a pre-trained SAM block with pseudo prompts instead of an inefficient mask generation model. This approach enables much denser mask prediction compared to conventional correlation-based methods.

(a):
1. 높은 계산 비용, 비효율적 메모리 사용 (∵ Two-stage pipeline)
2. 사전학습된 CLIP 모델과 마스크 영역 간에 도메인 차이로 인한 낮은 정확도 (∵ CLIP's image-level learning)

(b):
1. 경계 불안정한 디테일하지 못한 마스크 생성 (∵ Low-resolution correlation due to CLIP's global representation learning)

- **Ours**

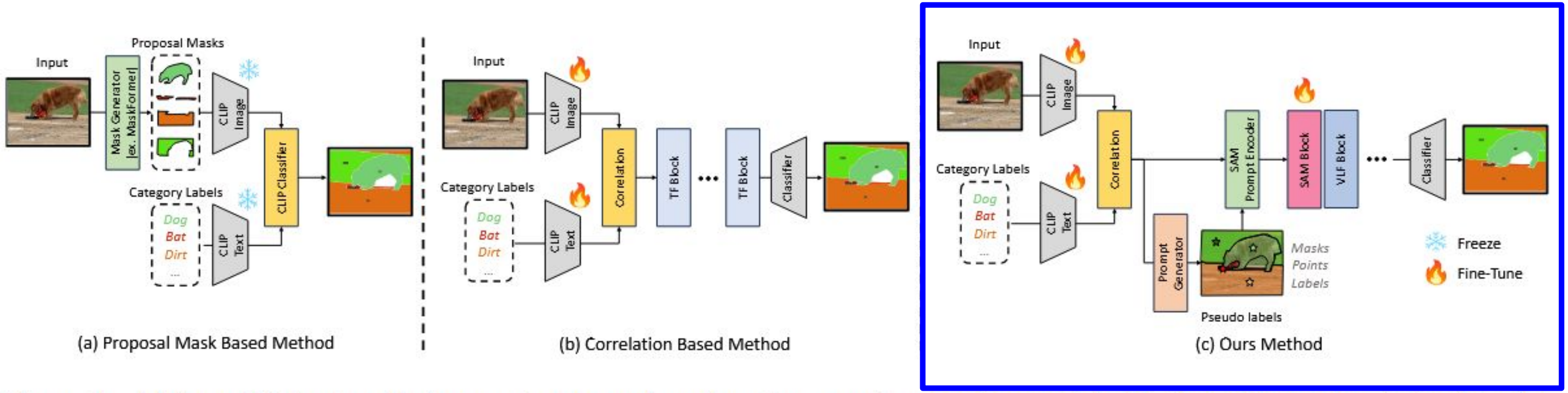

Figure 1. (a) A model structure that generates proposal masks using a mask generation model. (b) A model structure that refines the correlation between image and text. (c) The structure of the proposed ESC-Net. Our ESC-Net efficiently models the relationship between images and text by combining a pre-trained SAM block with pseudo prompts instead of an inefficient mask generation model. This approach enables much denser mask prediction compared to conventional correlation-based methods.

(c): SAM 사용해서 위 문제 극복 (구체적으로, SAM의 prompt encoder 및 mask decoder 사용)

- **Overview**

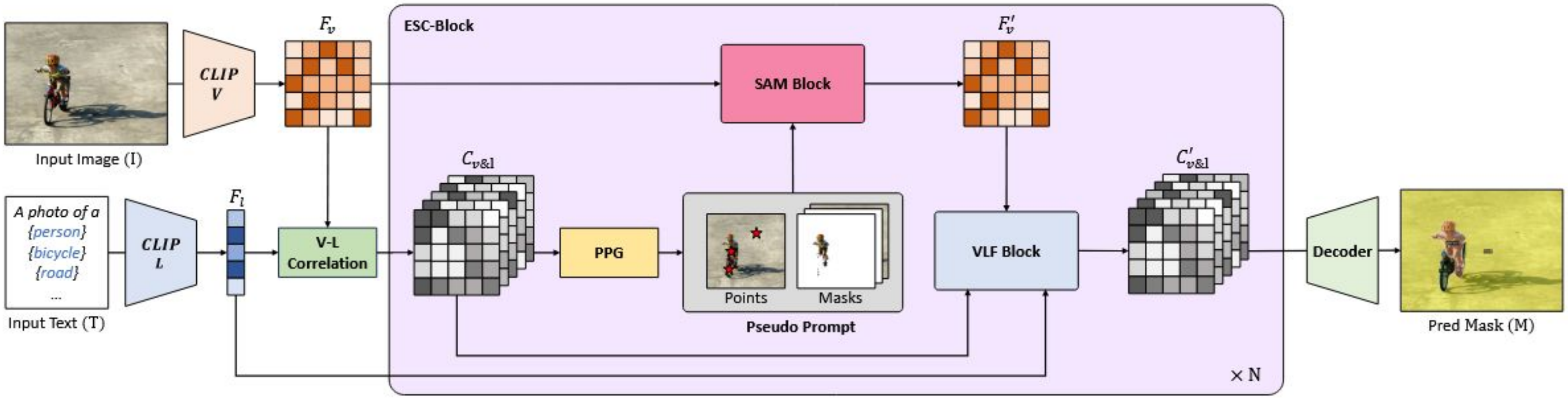

Figure 2. The proposed ESC-Net consists of the CLIP vision and language encoders, N consecutive ESCBlocks, and a decoder. Each ESC-Block generates a pseudo prompt from the image-text correlation map and uses it as input to the SAM block. The SAM block aggregates the CLIP image features. The VLF block models the image-text correlation using image features and text features, refining the correlation map through this process.

- # Vision-Language Correlation

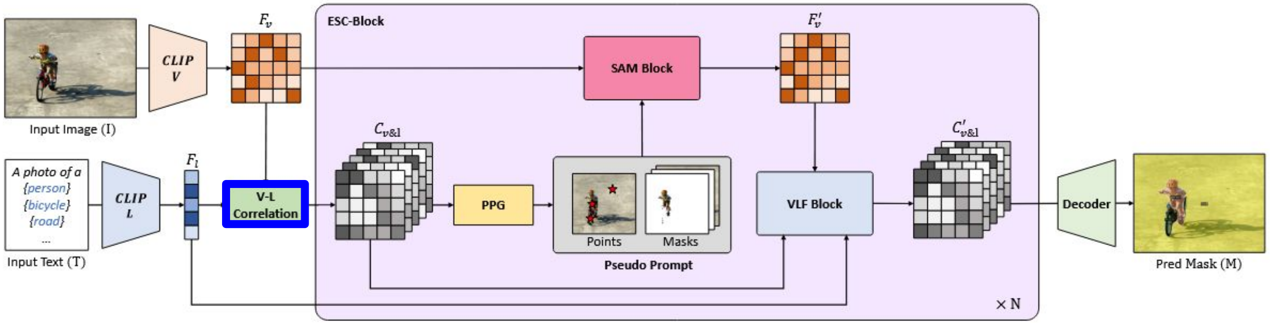

Figure 2. The proposed ESC-Net consists of the CLIP vision and language encoders, N consecutive ESCBlocks, and a decoder. Each ESC-Block generates a pseudo prompt from the image-text correlation map and uses it as input to the SAM block. The SAM block aggregates the CLIP image features. The VLF block models the image-text correlation using image features and text features, refining the correlation map through this process.

Cosine similarity로 vision-language correlation map 생성.

$$F_v \in \mathbb{R}^{C\times H\times W}$$
$$F_l \in \mathbb{R}^{C\times N_c}$$
$$C_{v\&l} \in \mathbb{R}^{N_c\times H\times W}$$

클래스

$$C^n_{v\&l}(i) = \frac{F_v(i)\cdot F^n_l}{\|F_v(i)\|\,\|F^n_l\|}. \quad (1)$$

[Effective SAM Combination for Open-Vocabulary Semantic Segmentation]

CVPR 2025

- **Pseudo Prompt Generator**

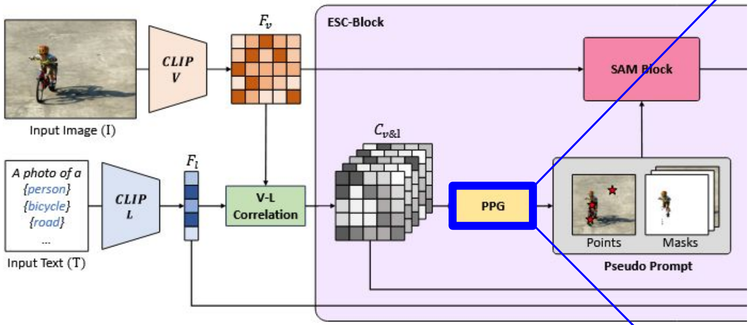

Figure 2. The proposed ESC-Net consists of the CLIP vision and language encode
Block generates a pseudo prompt from the image-text correlation map and uses i
the CLIP image features. The VLF block models the image-text correlation using
map through this process.

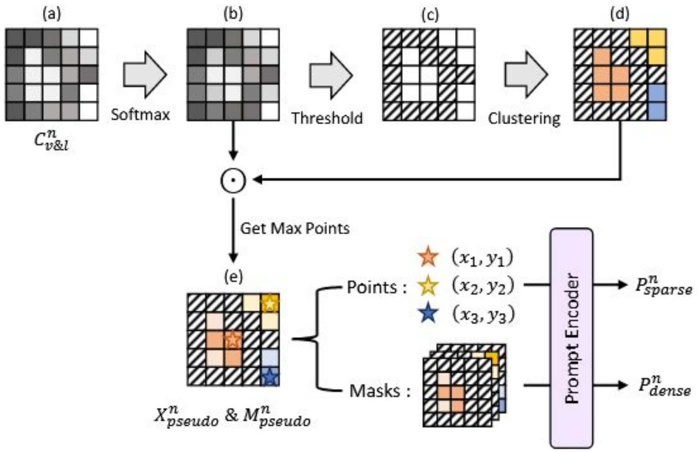

Figure 3. The process of the proposed Pseudo Prompt Generator (PPG). PPG aims to generate class-specific pseudo prompts from image-text correlation maps. For efficiency, all processes are computed in batch-wise parallelization over all classes.

Class-specific pseudo-prompt 생성.

(a): 앞서 생성한 vision-language correlation map $C^n_{v\&l} \in \mathbb{R}^{1\times H\times W}$
(b): Softmax 취해 생성한 probability mask
(c): Thresholding 통해 생성한 binary mask (1: 해당 클래스에 해당하는 객체 있다, 0: 없다.)
(d): K-means clustering 통해 생성한 여러 객체 구분된 clustered mask region map
(e): 앞서 생성한 probability mask와 clustered mask region map을 곱하여 생성한 filtered probability map
∴ Filtered probability map에서 각 region을 pseudo-mask로, 각 region에서 확률 제일 높은 픽셀을 pseudo-point로 결정.

N_o개의 Pseudo-points & Pseudo-masks → SAM's prompt encoder 통과 → Sparse prompt features, Dense prompt features 생성.

전유진

- **SAM Block**

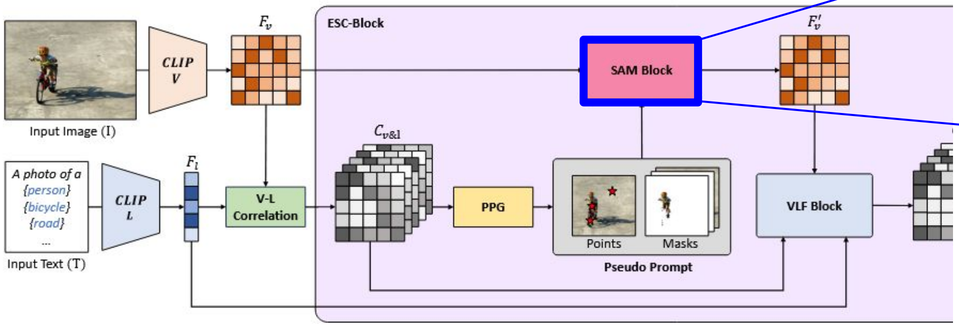

Figure 2. The proposed ESC-Net consists of the CLIP vision and language encoders, N consecutive ESC
Block generates a pseudo prompt from the image-text correlation map and uses it as input to the SAM
the CLIP image features. The VLF block models the image-text correlation using image features and te
map through this process.

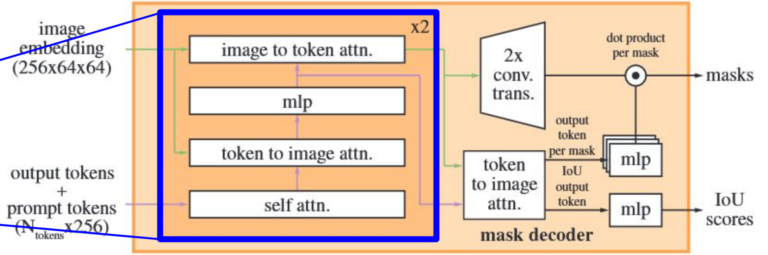

Figure 14: Details of the lightweight mask decoder. A two-layer decoder updates both the image embedding and prompt tokens via cross-attention. Then the image embedding is upscaled, from which the updated output tokens are used to dynamically predict masks. (Not illustrated for figure clarity: At every attention layer, positional encodings are added to the image embedding, and the entire original prompt token (including position encoding) is re-added to the token queries and keys.)

앞서 생성한 pseudo-prompt 정보 사용하여 pretrained SAM decoder TF block 통해 image feature 개선.

$$
\begin{aligned}
(F_v^n)' &= BCA\left(SA\left(F_l^n\right), F_v\right), \\
F_v' &= \mathrm{Conv}\left(\left[F_v^0; F_v^1; \ldots; F_v^{N_c}\right]\right),
\end{aligned}
\quad (2)
$$

$$F_v' \in \mathbb{R}^{C \times H \times W}$$

SA: Self Attention, BCA: Bidirectional Cross Attention

Effective SAM Combination for Open-Vocabulary Semantic Segmentation

- **Vision-Language Fusion Module**

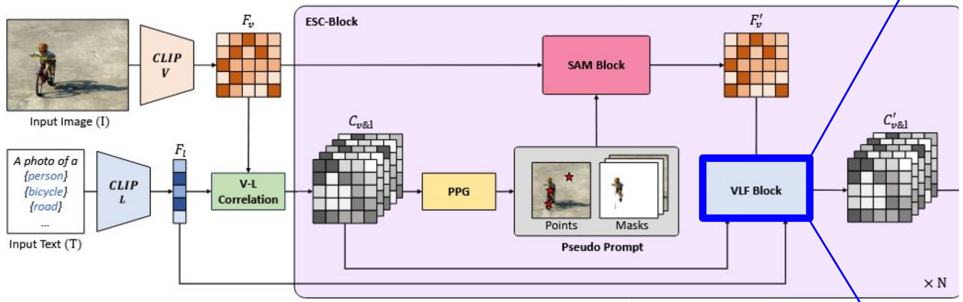

Figure 2. The proposed ESC-Net consists of the CLIP vision and language encoders, N consecutive ESCBlocks, a
Block generates a pseudo prompt from the image-text correlation map and uses it as input to the SAM block. Th
the CLIP image features. The VLF block models the image-text correlation using image features and text feature
map through this process.

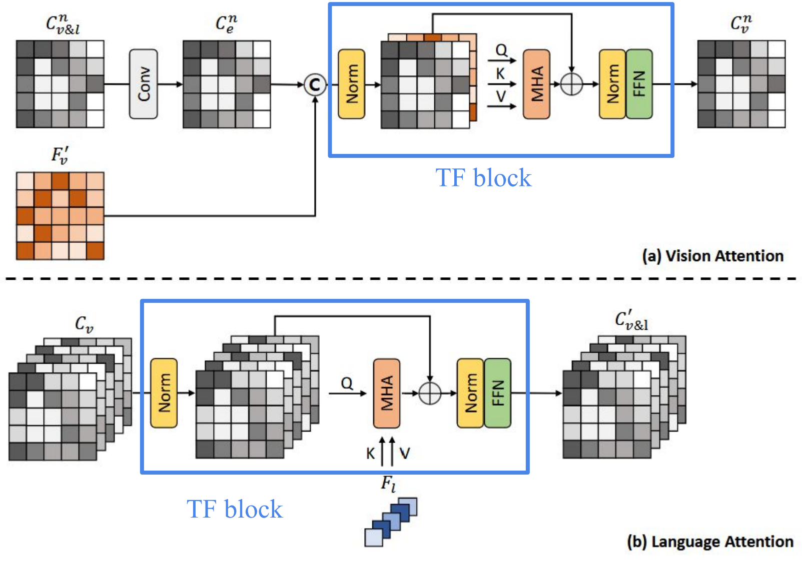

Figure 4. The structure of the proposed Vision-Language Fusion (VLF) block. VLF sequentially applies image and text guidance to the correlation map to refine it.

앞서 개선한 image feature 사용하여 vision-language correlation map 개선.

(a) Vision correlation map 생성

n-th class에 대한 Correlation map: $C_{v\&l}^n \in \mathbb{R}^{1\times H\times W}$

Embedded class correlation map: $C_e^n \in \mathbb{R}^{C\times H\times W}$

∴ Vision correlation map: $C_v^n \in \mathbb{R}^{C\times H\times W}$

(b) 개선된 correlation map 생성 $C_{v\&l}'$

전유진

- **Mask Prediction Decoder**

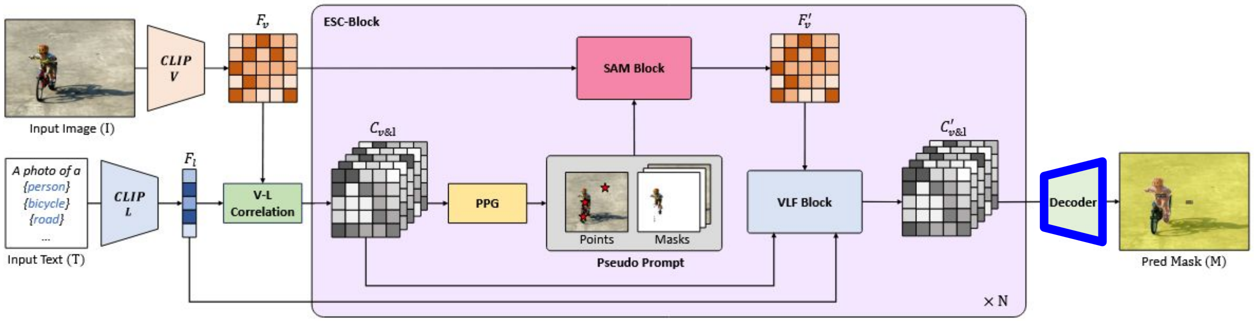

Figure 2. The proposed ESC-Net consists of the CLIP vision and language encoders, N consecutive ESCBlocks, and a decoder. Each ESC-Block generates a pseudo prompt from the image-text correlation map and uses it as input to the SAM block. The SAM block aggregates the CLIP image features. The VLF block models the image-text correlation using image features and text features, refining the correlation map through this process.

U-Net's upsampling layers 사용
CLIP image feature를 skip connection으로 사용

- # Experiments

Model	Publication	VLM	Additional Backbone	Training Dataset	Additional Dataset	A-847	PC-459	A-150	PC-59	PAS-20	PAS-20^b
SPNet [32]	CVPR'19	-	ResNet-101	PASCAL VOC	✗	-	-	-	24.3	18.3	-
ZS3Net [1]	NeurIPS'19	-	ResNet-101	PASCAL VOC	✗	-	-	-	19.4	38.3	-
LSeg [20]	ICLR'22	CLIP ViT-B/32	ResNet-101	PASCAL VOC-15	✗	-	-	-	-	47.4	-
LSeg+ [10]	ECCV'22	ALIGN	ResNet-101	COCO-Stuff	✗	2.5	5.2	13.0	36.0	-	59.0
ZegFormer [6]	CVPR'22	CLIP ViT-B/16	ResNet-101	COCO-Stuff-156	✗	4.9	9.1	16.9	42.8	86.2	62.7
ZSseg [35]	ECCV'22	CLIP ViT-B/16	ResNet-101	COCO-Stuff	✗	7.0	-	20.5	47.7	88.4	-
OpenSeg [10]	ECCV'22	ALIGN	ResNet-101	COCO Panoptic	✓	4.4	7.9	17.5	40.1	-	63.8
OVSeg [21]	CVPR'23	CLIP ViT-B/16	ResNet-101c	COCO-Stuff	✓	7.1	11.0	24.8	53.3	92.6	-
ZegCLIP [40]	CVPR'23	CLIP ViT-B/16	-	COCO-Stuff-156	✗	-	-	-	41.2	93.6	-
SAN [36]	CVPR'23	CLIP ViT-B/16	-	COCO-Stuff	✗	10.1	12.6	27.5	53.8	94.0	-
DeOP [11]	ICCV'23	CLIP ViT-B/16	ResNet-101c	COCO-Stuff-156	✗	7.1	9.4	22.9	48.8	91.7	-
SCAN [22]	CVPR'24	CLIP ViT-B/16	Swin-B	COCO-Stuff	✗	10.8	13.2	30.8	<u>58.4</u>	<u>97.0</u>	-
EBSeg [30]	CVPR'24	CLIP ViT-B/16	SAM ViT-B	COCO-Stuff	✗	11.1	17.3	30.0	56.7	94.6	-
SED [33]	CVPR'24	ConvNeXt-B	-	COCO-Stuff	✗	11.4	18.6	31.6	57.3	94.4	-
CAT-Seg [5]	CVPR'24	CLIP ViT-B/16	-	COCO-Stuff	✗	<u>12.0</u>	<u>19.0</u>	<u>31.8</u>	57.5	94.6	<u>77.3</u>
ESC-Net (ours)	-	CLIP ViT-B/16	-	COCO-Stuff	✗	**13.3** (+1.3)	**21.1** (+2.1)	**35.6** (+3.8)	**59.0** (+0.6)	**97.3** (+0.3)	**80.1** (+2.8)
LSeg [20]	ICLR'22	CLIP ViT-B/32	ViT-L/16	PASCAL VOC-15	✗	-	-	-	-	52.3	-
OpenSeg [10]	ECCV'22	ALIGN	EfficientNet-B7	COCO Panoptic	✓	8.1	11.5	26.4	44.8	-	70.2
OVSeg [21]	CVPR'23	CLIP ViT-L/14	Swin-B	COCO-Stuff	✓	9.0	12.4	29.6	55.7	94.5	-
SAN [36]	CVPR'23	CLIP ViT-L/14	-	COCO-Stuff	✗	12.4	15.7	32.1	57.7	94.6	-
ODISE [34]	CVPR'23	CLIP ViT-L/14	Stable Diffusion	COCO-Stuff	✗	11.1	14.5	29.9	57.3	-	-
FC-CLIP [37]	NeurIPS'23	ConvNeXt-L	-	COCO Panoptic	✗	11.2	12.7	26.6	42.4	89.5	-
MAFT [15]	NeurIPS'23	CLIP ViT-L/14	-	COCO-Stuff	✗	12.7	16.2	33.0	59.0	92.1	-
USE [31]	CVPR'24	CLIP ViT-L/14	DINOv2, SAM	COCO-Stuff	✓	13.4	15.0	37.1	58.0	-	-
SCAN [22]	CVPR'24	CLIP ViT-L/14	Swin-B	COCO-Stuff	✗	14.0	16.7	33.5	59.3	97.0	-
EBSeg [30]	CVPR'24	CLIP ViT-L/14	SAM ViT-B	COCO-Stuff	✗	13.7	21.0	32.8	60.2	<u>97.2</u>	-
SED [33]	CVPR'24	ConvNeXt-L	-	COCO-Stuff	✗	13.9	22.6	35.2	60.6	96.1	-
CAT-Seg [5]	CVPR'24	CLIP ViT-L/14	-	COCO-Stuff	✗	<u>16.0</u>	<u>23.8</u>	<u>37.9</u>	<u>63.3</u>	97.0	<u>82.5</u>
MAFT+ [17]	ECCV'24	ConvNeXt-L	-	COCO-Stuff	✗	15.1	21.6	36.1	59.4	96.5	-
ESC-Net (ours)	-	CLIP ViT-L/14	-	COCO-Stuff	✗	**18.1** (+2.1)	**27.0** (+3.2)	**41.8** (+3.9)	**65.6** (+2.3)	**98.3** (+1.1)	**86.3** (+3.8)

Table 1. Quantitative evaluation on open-vocabulary segmentation benchmarks. The best-performing results are presented in bold, while the second-best results are underlined. Improvements over the second-best are highlighted in red.

전유진

- **Experiments**

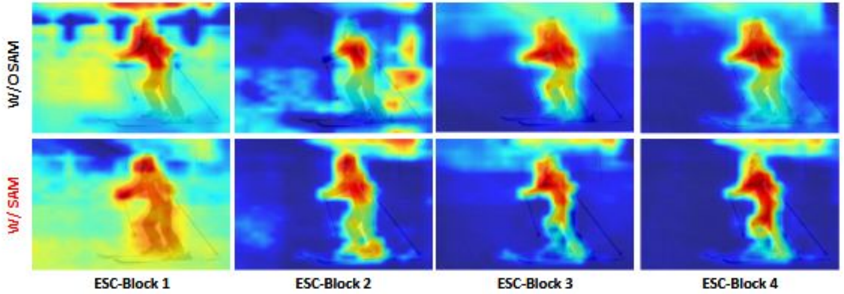

Figure 6. Visualization of image-text correlation maps with and without the SAM block. We visualize the model activation maps for the "Person" class for each ESC-Block. The proposed SAM-based method enables more accurate and dense object localization compared to the baseline.

- **Experiments**

Figure 5. Qualitative comparison of CAT-Seg and our ESC-Net across various datasets. Our model is capable of generating more accurate and robust masks compared to existing correlation-based state-of-the-art method.