Excluding the Impossible for Open Vocabulary Semantic Segmentation

Shiyuan Zhao[1], **Baodi Liu**[1*], **Yu Bai**[1], **Weifeng Liu**[1], **Shuai Shao**[2*]

[1] China University of Petroleum (East China)
[2]Zhejiang Lab
zhaoshiyuan@s.upc.edu.cn, {thu.liubaodi, liuwfxy, shaoshuai0914}@gmail.com, baiyu_upc@163.com

Abstract

Open vocabulary semantic segmentation is a hot topic in research, focusing on segmenting and recognizing a diverse array of categories in varied environments, including those previously unknown, thereby holding significant practical value. Mainstream studies utilize the CLIP model for direct semantic segmentation (denoted as "forward methods"), which often struggles to represent underrepresented categories effectively. To address this issue, this paper introduces a novel approach **E**xcluding the Impossib**L**e **S**emantic S**E**gmentation Network (**ELSE-Net**) based on reverse thinking. By excluding improbable categories, ELSE-Net narrows the selection range for forward methods, significantly reducing the risk of misclassification. In implementation, we initially draw on leading research to design the **G**eneral **P**rocessing Block (**GP-Block**), which generates inclusion probabilities (the likelihood of belonging to a category) by using the CLIP model cooperated with a **M**ask **P**roposal **N**etwork (**MPN**). We then present the **EX**cluding the Im**P**ossible Block (**EXP-Block**), which computes exclusion probabilities (the likelihood of not belonging to a category) through the CLIPN model and a custom-designed **R**everse **R**etrieval Adapter (**R^2-Adapter**). These exclusion probabilities are subsequently used to refine the inclusion probabilities, which are ultimately employed to annotate class-agnostic masks. Moreover, the core component of our EXP-Block is model-agnostic, enabling it to enhance the capabilities of existing frameworks. Experimental results from four benchmark datasets validate the effectiveness of ELSE-Net and underscore the seamless model-agnostic functionality of the EXP-Block.

Code — https://github.com/shishiyuanzhao/ELSE-Net

Introduction

Open vocabulary semantic segmentation research has garnered considerable attention in recent years (Liang et al. 2023; Han et al. 2023a; Qin et al. 2023), aiming to perform pixel-level labeling in diverse environments, including categories that were not encountered during training. This research extends the capabilities of segmentation models beyond known categories, enabling them to adapt to dynamic real-world scenarios. Consequently, it holds significant importance for tasks such as self-driving technology, city planning, and marine surveillance, where new objects of interest continuously emerge.

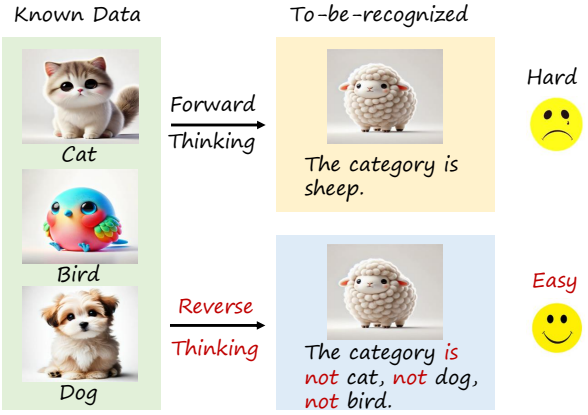

Figure 1: An intuitive example to introduce the concept of "Excluding the Impossible". Predicting unseen categories based solely on data from known categories is challenging. However, it is comparatively straightforward to ascertain that certain data do not belong to any known categories. By using this reverse thinking strategy to exclude impossible categories as prior information for forward methods, we can effectively narrow the decision space of these methods, thereby effectively lowering the risk of misclassification.

The prevailing mainstream approaches leverage the exceptional zero-shot capabilities of CLIP (Radford et al. 2021) to develop "forward methods" that directly enable semantic segmentation, achieving notable progress (Xu et al. 2023; Cho et al. 2024; Xie et al. 2024; Ghiasi et al. 2022). However, as the CLIP model suffers from the class-imbalance-problem (Chuang et al. 2023; Parashar et al. 2024), this issue inevitably affects CLIP-based segmentation models as well, causing them to be biased towards recognizing seldom-seen or novel categories as common ones.

To address this issue, we introduce an innovative concept from a reverse perspective by excluding improbable categories to narrow the selection range of forward models. This reverse concept is straightforward: although it is challeng-

*Corresponding author.

ing to identify unseen categories based solely on data from known categories, we can readily ascertain that data from new categories do not belong to any known ones (an example is shown in Fig.1).

In this context, we present the **E**xcluding the Impossib**L**e **S**emantic S**E**gmentation Network (**ELSE-Net**) (see Fig. 2 for more details), comprising two principal components: (1) The **G**eneral **P**rocessing Block (**GP-Block**) follows traditional forward methods, initially generating class-agnostic masks using the **M**ask **P**roposal **N**etwork (**MPN**) and subsequently calculating the inclusion probability (the likelihood of belonging to a category). (2) The **EX**cluding the Im**P**ossible Block (**EXP-Block**) takes a reverse perspective, starting with encoding the reverse prompt using the CLIPN text encoder (Wang et al. 2023). This is followed by the design of a **R**everse **R**etrieval Adapter (**R^2-Adapter**) that generates a high-quality exclusion probability (the likelihood of not belonging to a category). Ultimately, this exclusion probability is employed to correct the inclusion probability and precisely annotate the class-agnostic masks using the adjusted probability. Within ELSE-Net, the CLIP model remains frozen, while the MPN, CLIPN text encoder, and R^2-Adapter require ongoing parameter updates. Notably, the EXP-Block is a model-agnostic module.

Our main contributions are summarized as follows:

- We present a novel method named ELSE-Net that leverages the "Excluding the Impossible" concept to narrow the model's selection scope.
- The core EXP-Block is a model-agnostic module tailored for compatibility with the majority of CLIP-based approaches. It comprises the CLIPN model, which necessitates fine-tuning, and the R^2-Adapter, engineered to optimize the functionality of "excluding the impossible".
- Our method has been evaluated across 4 benchmark datasets. The results have evaluated the efficiency and the model-agnostic functionality of our method.

Related Work

Vision-Language Models

Recent advancements in the field of vision-language models (VLMs) have sparked significant interest. In our research, we use two of them: CLIP (Radford et al. 2021) employs a vision-language contrastive pretraining approach that facilitates joint understanding of images and text. By aligning images with corresponding textual descriptions, CLIP excels in zero-shot image classification and other downstream visual tasks. Its capabilities highlight the potential of cross-modal learning within the realm of artificial intelligence. As an evolution of CLIP, CLIPN (Wang et al. 2023) maintains the core architecture and training approaches of its predecessor but incorporates a distinctive feature that differentiates it. By adjusting the prompt template during training, CLIPN modifies the interaction between images and text. This adjustment allows CLIPN to make more nuanced predictions regarding whether a sample belongs to a specific class, thus increasing its effectiveness in complex challenges.

Open Vocabulary Semantic Segmentation

Open vocabulary semantic segmentation research has been critical recently. Current methodologies are primarily categorized based on "image-text feature contrast" into pixel-level and region-level approaches. Pixel-level methods (Li et al. 2022) treat each pixel as a discrete unit for annotation, whereas region-level methods (Ghiasi et al. 2022; Ding et al. 2022; Liang et al. 2023) involve initially masking the image and subsequently annotating the masked regions. The latter approach is more prevalent in mainstream techniques.

Region-level methods vary in their implementation and can be classified into two types: (1) Single-stage methods (Xu et al. 2023; Yu et al. 2024): These are designed as end-to-end networks that efficiently generate and annotate masks in a single step, benefiting from a reduced parameter count. (2) Two-stage methods (Qin et al. 2023; Han et al. 2023a; Liang et al. 2023): These methods decouple the segmentation process into two phases: generating class-agnostic masks and then classifying these masks. These approaches can effectively utilize the strengths of VLMs.

In this paper, we consider the above methods as forward approaches. While, the EXP-Block, innovatively designed using a reverse thinking strategy, can be seamlessly integrated into these methods to enhance their functionality.

Reverse Thinking Methods

Reverse thinking, when effectively applied, can significantly enhance forward methods and lead to breakthroughs. However, in the field of computer vision, the adoption of this concept remains relatively limited. One notable example is CLIPN (Wang et al. 2023), which leverages a reverse text encoder to address out-of-distribution detection task even when no samples are available. Another instance, DeIL (Shao et al. 2024), utilizes this concept to overcome challenges in few-shot classification within an open-world setting. This paper is the first research to integrate the reverse concept into the open vocabulary semantic segmentation.

Problem Setup

Open vocabulary semantic segmentation is designed to segment the image $\mathcal{X} \in \mathbb{R}^{H\times W\times 3}$ into K masks with corresponding semantic labels. We formulate this step as:

$$\{y_i\}_{i=1}^{K} = \{m_i, c_i, t_i\}_{i=1}^{K}. \tag{1}$$

where $m_i \in \{0,1\}^{H\times W}$ represents the ground truth mask, and each mask is associated with a unique ground truth class label c_i. t_i corresponds to c_i's category name.

During the training phase, a specific set of class labels $\mathcal{C}_{train}$ is employed, whereas a different set, $\mathcal{C}_{test}$, is to-be-predicted during inference. The sets $\mathcal{T}_{train}$ and $\mathcal{T}_{test}$ correspond to the category names of $\mathcal{C}_{train}$ and $\mathcal{C}_{test}$, respectively. In the open vocabulary scenario, $\mathcal{C}_{test}$ may include categories not encountered during training, *i.e.*, $\mathcal{C}_{train} \neq \mathcal{C}_{test}$. Besides, consistent with prior studies(Qin et al. 2023; Xu et al. 2023), we assume access to all categories ($\mathcal{C}$) and the corresponded category names ($\mathcal{T}$) in the world during the inference phase, *i.e.*, $\{\mathcal{C}_{train} \cup \mathcal{C}_{test}\} \in \mathcal{C}$ and $\{\mathcal{T}_{train} \cup \mathcal{T}_{test}\} \in \mathcal{T}$.

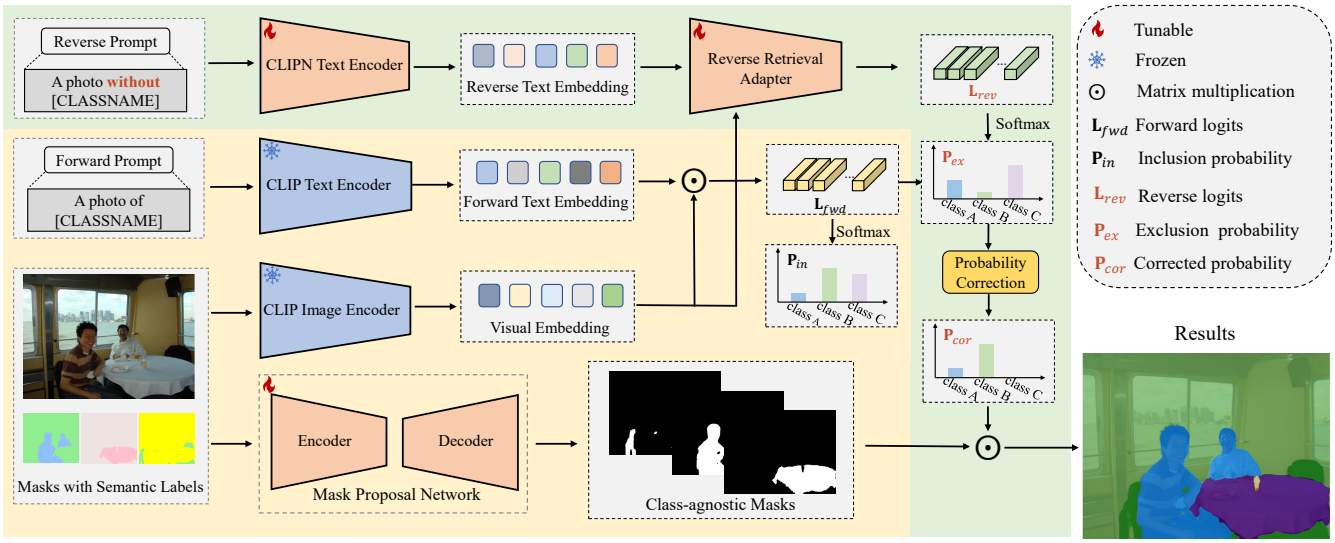

Figure 2: The framework of **E**xcluding the Impossib**L**e **S**emantic S**E**gmentation Network (**ELSE-Net**), which includes two primary components. (1) **G**eneral **P**rocessing Block (**GP-Block**) on the light yellow background: The process initiates with the generation of visual embedding (*i.e.*, $\mathbf{E}_{vis}$) using the CLIP image encoder. Simultaneously, the image and the visual embedding are then sent into the **M**ask **P**roposal **N**etwork (**MPN**) to generate N class-agnostic masks. Following this, the CLIP text encoder processes the forward prompt to produce forward text embeddings (*i.e.*, $\mathbf{E}_{txt_{fwd}}$). Finally, the inclusion probability (*i.e.*, $\mathbf{P}_{in}$, denotes the likelihood of belonging to a certain category) for each mask is calculated. (2) **EX**cluding the Im**P**ossible Block (**EXP-Block**) on the light green background: It begins with encoding the reverse prompt through the CLIPN Text Encoder to generate reverse text embeddings (*i.e.*, $\mathbf{E}_{txt_{rev}}$). Then a **R**everse **R**etrieval Adapter (**R^2-Adapter**) is deployed to calculate the exclusion probability (*i.e.*, $\mathbf{P}_{ex}$, denotes the likelihood of not belonging to a certain category). This exclusion probability is used as a prior to correct the inclusion probability, which is subsequently applied to annotate the class-agnostic masks.

Methodology

Overview

In this paper, we propose the **E**xcluding the Impossib**L**e **S**emantic S**E**gmentation Network (**ELSE-Net**). The fundamental principle behind ELSE-Net is to "exclude the impossible", thereby narrowing the range of choices, which reduces the risk of misclassification and improves accuracy. The overall framework of ELSE-Net is shown in Fig. 2, which includes two blocks:

- **G**eneral **P**rocessing Block (**GP-Block**), referring to top-tier studies (Xu et al. 2024; Han et al. 2023a), generates class-agnostic masks and predicts inclusion probability. For more details, please see the light yellow part in Fig. 2. GP-Block comprises four steps: (1) The image is input into the frozen CLIP Image Encoder to produce visual embedding. (2) Simultaneously, the image and the visual embedding are processed through the Mask Proposal Network (MPN) to produce class-agnostic masks. (3) The forward prompt is input into the frozen CLIP text encoder to generate the forward text embedding. (4) The inclusion probability is then calculated from the masks and forward text embeddings, indicating the likelihood that a sample belongs to a specific category.
- **EX**cluding the Im**P**ossible Block (**EXP-Block**) is a model-agnostic auxiliary module we developed. It is designed to enhance classification performance by selectively excluding improbable categories. For more details, please refer to the light green part in Fig. 2. EXP-Block operates in three key steps: (1) The reverse prompt is fed into the tunable CLIPN text encoder to generate the reverse text embedding. (2) The masks and reverse text embeddings are then processed through our custom-designed **R**everse **R**etrieval Adapter (**R^2-Adapter**), which computes the exclusion probability, assessing the likelihood that a sample does not belong to a particular category. (3) This exclusion probability is utilized to adjust the inclusion probability, leading to a corrected probability. This adjusted probability is then used to accurately annotate the class-agnostic masks.

General Processing Block

We drew inspiration from top-tier studies (Han et al. 2023a; Ghiasi et al. 2022; Qin et al. 2023; Liang et al. 2023; Xu et al. 2022) to design the General Processing Block (GP-Block). For more details, please see the light yellow part in Fig.2. The primary objective of this block is to generate class-agnostic masks and predict the masks' inclusion probability through CLIP. The GP-Block comprises four steps:

Visual Embedding Initially, we crop the image and input them into the CLIP image encoder, yielding visual embedding, which can be formulated as:

$$\mathbf{E}_{vis} = \mathcal{F}_{clip_{img}}(\mathcal{X}). \tag{2}$$

where $\mathbf{E}_{vis}$ denotes the visual embedding; $\mathcal{F}_{clip_{vis}}$ denotes the frozen CLIP image encoder.

Class-agnostic Masks Subsequently, the image and the visual embedding are fed into the Mask Proposal Network (MPN) to generate N class-agnostic masks, which can be formulated as:

$$\{\hat{m}_i\}_{i=1}^{N} = \mathcal{F}_{mpn}\left(\mathcal{X}, \mathbf{E}_{vis}\right). \tag{3}$$

where $\mathcal{F}_{mpn}$ denotes the to-be-trained MPN, which is composed of eight transformer layers; $\mathcal{X}$ is the training image; $\{\hat{m}_i\}_{i=1}^{N}$ is the to-be-predicted class-agnostic masks.

Forward Text Embedding Afterward, we utilize a forward prompt template to carefully construct descriptions for each category. The template is formatted as sentences like "`A photo of [CLASSNAME]`", which can be formulated as $Template_{fwd}\left(t_i\right)$. where t_i is the c_i's category name. Following this, we employ the frozen CLIP text encoder to process these descriptions, generating informative text representations for each class.

$$\mathbf{E}_{txt_{fwd}} = \mathcal{F}_{clip_{txt}}\left(Template_{fwd}\left(\mathcal{T}\right)\right). \tag{4}$$

where $\mathbf{E}_{txt_{fwd}} \in \mathbb{R}^{dim \times |\mathcal{C}|}$ denotes the text embedding and forward text embedding; $|\mathcal{C}|$ indicates the number of categories; $\mathcal{F}_{clip_{txt}}$ denotes the frozen CLIP text encoder.

Inclusion Probability Finally, we gauge the similarity between the forward text features and the masks by:

$$\mathbf{L}_{fwd} = \left(\{\hat{m}_i\}_{i=1}^{N}\right)^{T} \mathbf{E}_{txt_{fwd}} \tag{5}$$

$$\mathbf{P}_{in} = Softmax\left(\mathbf{L}_{fwd}\right) \tag{6}$$

where $\mathbf{L}_{fwd} \in \mathbb{R}^{N \times |\mathcal{C}|}$, $\mathbf{P}_{in} \in \mathbb{R}^{N \times |\mathcal{C}|}$ denote the forward logits and the inclusion probability, the element $\mathbf{P}_{in}(i, c)$ represents the probability that the i-th sample belongs to the c-th class.

Excluding the Impossible Block

This section introduces the Excluding the Impossible Block (EXP-Block), which integrates the "Excluding the Impossible" concept into the framework via reverse thinking to improve segmentation performance. For further details, please refer to the light green part in Fig. 2. It is noteworthy that this module does not affect the information generation within the GP-Block, rendering the EXP-Block effectively model-agnostic. The module consists of three primary components, which are elaborated upon below.

Reverse Text Embedding Contrary to the Forward Text Processing, we employ a reverse prompt template here to construct descriptions for each category. The template can be formatted as sentences like "`A photo without [CLASSNAME]`". We denote the reverse prompt for categories as $Template_{rev}\left(t_i\right)$. Subsequently, we deploy the text encoder of the CLIPN model (Wang et al. 2023) to generate reverse text embeddings to capture intricate reverse textual details. It is crucial to note that the CLIPN model necessitates fine-tuning. This step plays a crucial role in achieving the goal of "excluding the impossible". We formulate this process as:

$$\mathbf{E}_{txt_{rev}} = \mathcal{F}_{clipn_{txt}}\left(Template_{rev}\left(\mathcal{T}\right)\right). \tag{7}$$

where $\mathbf{E}_{txt_{rev}} \in \mathbb{R}^{dim \times |\mathcal{C}|}$ denotes the reverse text embedding.

Reverse Retrieval Adapter for Exclusion Probability To enhance the retrieval of reverse text features alongside visual features, we developed the Reverse Retrieval Adapter (R^2-Adapter). Initially, masks undergo processing via a straightforward multi-layer perceptron (MLP) to refine their quality. Subsequently, the refined masks are utilized to match with the reverse text embeddings, thereby facilitating the computation of reverse logits. We formulate these steps as:

$$\mathbf{L}_{rev} = \mathcal{F}_{R^2}\left(\{\hat{m}_i\}_{i=1}^{N}, \mathbf{E}_{txt_{rev}}\right) \tag{8}$$

$$\mathbf{P}_{ex} = Softmax\left(\mathbf{L}_{rev}\right). \tag{9}$$

where $\mathcal{F}_{R^2}$ denotes the R^2-Adapter; $\mathbf{L}_{rev} \in \mathbb{R}^{N \times |\mathcal{C}|}$ and $\mathbf{P}_{ex} \in \mathbb{R}^{N \times |\mathcal{C}|}$ denote the reverse logits and exclusion probability, the element $\mathbf{P}_{ex}(i, c)$ represents the probability that the i-th sample **does not** belong to the c-th class; α is the hyperparameter.

Probability Correction Here, we use the exclusion probability ($\mathbf{P}_{ex}$) obtained through the reverse process, to adjust the inclusion probability ($\mathbf{P}_{in}$) generated from the forward process. Specifically, we set a threshold (ϵ) for correction. If $\mathbf{P}_{ex}(i, c) \geq \epsilon$, it signifies that the i-th sample **does not** belong to the c-th class, and we update the $\mathbf{P}_{in}(i, c)$ to 0. We formulate this process as:

$$\mathbf{P}_{cor}(i, c) = \begin{cases} 0 & \text{if } \mathbf{P}_{ex}(i, c) \geq \epsilon \\ \mathbf{P}_{in}(i, c) & \text{otherwise.} \end{cases} \tag{10}$$

where $\mathbf{P}_{cor} \in \mathbb{R}^{N \times |\mathcal{C}|}$ is the corrected probability, the element $\mathbf{P}_{cor}(i, c)$ represents the probability that the i-th sample belongs to the c-th class.

Semantic Segmentation and Loss Function

Finally, we use the corrected probability to label the class-agnostic masks by $\left(\{\hat{m}_i\}_{i=1}^{N} \odot \mathbf{P}_{cor}\right)$.

Throughout the entire process, three types of losses are involved: dice loss (Cheng et al. 2022) and binary-entropy (BE) loss (Jadon 2020) are employed during the update of the MPN, and the cross-entropy (CE) loss (Xu et al. 2023) is used during the final classification stage.

$$loss = loss_{dice} + loss_{be} + loss_{ce}. \tag{11}$$

Experiments

In this section, we present the experiments to address the four questions:

Q1. How does ELSE-Net perform in comparison to current SOTAs? (A1. See Tab. 1)

Q2. Is the EXP-Block genuinely model-agnostic, and what impact does it have? (A2. See Tab. 2)

Q3. How do different components influence the outcomes? (A3. See Tab. 3 and Fig. 4)

Q4. What is the extra consumption of the EXP-Block for the existing method? (A4. See Tab. 4)

Method	Backbone	Training Dataset	Testing Dateset			
			ADE-847	ADE-150	PC-59	VOC
LSeg+ (ICLR'22) (Ghiasi et al. 2022)	ALIGN NB-B7	COCO	3.8	18.0	46.5	/
Simseg (ECCV'22) (Xu et al. 2022)	CLIP ViT-B/16	COCO	7.0	20.5	47.7	88.4
FreeSeg (CVPR'23) (Qin et al. 2023)	CLIP ViT-B/16	COCO	/	**28.6**	/	82.6
DeOp (CVPR'23) (Han et al. 2023a)	CLIP ViT-B/16	COCO	7.1	22.9	48.8	91.7
OVSeg (CVPR'23) (Liang et al. 2023)	CLIP ViT-B/16	COCO	7.1	24.8	53.3	92.6
SAN (CVPR'23) (Xu et al. 2023)	CLIP ViT-B/16	COCO	<u>10.1</u>	27.5	<u>53.8</u>	<u>94.0</u>
GKC (ICCV'23) (Han et al. 2023b)	CLIP ViT-B/16	COCO Panoptic	3.5	18.8	45.2	83.2
MaskCLIP (ICML'23) (Ding, Wang, and Tu 2023)	CLIP ViT-L/14	COCO Panoptic	8.2	23.7	45.9	/
SPT-SEG (AAAI'24) (Xu et al. 2024)	CLIP ViT-B/16	COCO	/	/	/	93.2
ELSE-Net (Ours)	CLIP ViT-B/16	COCO	**10.3** (+0.2)	<u>27.9</u> (-0.7)	**54.1** (+0.3)	**95.1** (+1.1)

Table 1: Quantitative evaluation on standard benchmarks. The best-performing results are presented in bold, while the second-best results are underlined. Improvements and decreases over the previous SOTA are highlighted in red and green.

Settings

Datasets and Evaluation Metric We adhere to standard settings (Xu et al. 2023), training our model on the COCO-Stuff dataset (Caesar, Uijlings, and Ferrari 2018) and evaluating it across several datasets: Pascal VOC (VOC) (Everingham et al. 2015), ADE20K-847 (ADE-847) (Zhou et al. 2017), ADE20K-150 (ADE-150) (Zhou et al. 2017), and Pascal Context-59 (PC-59) (Mottaghi et al. 2014). The categories for training and testing are not fully overlapping; specifically, the label-set similarity (Xu et al. 2023) between VOC, ADE-847, ADE-150, PC-59, and COCO is 0.91, 0.57, 0.73, and 0.86, respectively.

Besides, we follow (Xu et al. 2022; Qin et al. 2023; Han et al. 2023a; Liang et al. 2023) to adopt the mean of Interaction over Union (mIOU) to evaluate the open vocabulary semantic segmentation performance.

Implementation All models were trained on the COCO Stuff dataset using the AdamW optimizer, which was configured with an initial learning rate of 1e-4 and a weight decay of 1e-4. For both the CLIP and CLIPN models, the backbone architecture employed was ViT-B/16, leveraging its strong capabilities in visual feature representation. Additionally, the R^2-Adapter was designed with 2 layers and a feature dimension of 2048, ensuring adequate flexibility and capacity for feature adaptation. The adapter weight parameter α was fixed at 0.2 to balance the contribution of the adapter to the overall learning process. In the probability correction procedure, a threshold value of 0.2 was used, enabling effective filtering and adjustment of predictions. Further details regarding these configurations and their impact are provided in the ablation study section. All experiments were executed on a high-performance system featuring an NVIDIA RTX 4090 GPU with 24GB of memory, paired with a 16 vCPU Intel (R) Xeon (R) Gold 6430 processor. The batch size for training was set to 4 to accommodate the computational requirements while ensuring stability during optimization.

Comparison with SOTAs

The comparative results on four benchmark datasets are presented in Tab. 1.

(1) On the Pascal VOC dataset: Due to the high label similarity with the COCO dataset, all methods tested exhibit superior performance. Among these, our ELSE-Net consistently outperformed the others by a margin of at least 1.1. This significant improvement primarily results from the integration of our model-agnostic EXP-Block, which will be further elaborated upon in subsequent sections. Furthermore, to present a more intuitive understanding of our model's performance, we showcase its results on the VOC, where it achieved excellent outcomes in the annotated categories. Parts of visualization results are displayed in Fig.3.

(2) On other datasets: Unlike the Pascal VOC dataset, the ADE-847, ADE-150, and PC-59 datasets exhibit less similarity to the COCO dataset, generally leading to lower performance across these benchmarks. Nonetheless, our ELSE-Net demonstrated superior effectiveness, achieving the best experimental results on ADE-847 and PC-59 with a performance improvement of +0.2 mIOU and +0.3 mIOU respectively. On ADE-150 , ELSE-Net achieved the second-best result over previous state-of-the-art model. Overall, this experiment conclusively demonstrates the effectiveness of our ELSE-Net.

Model-Agnostic Functionality of the EXP-Block

As discussed earlier, our EXP-Block is designed to be model-agnostic, making it compatible with most existing methods. To validate this claim and further illustrate the effectiveness of our approach, we selected three classic open-source methods as baseline models. First, we reproduced these baseline methods to ensure consistency, and then we integrated the EXP-Block into each of them. The corresponding experimental results are summarized in Tab.2. The observations reveal that, in the vast majority of cases, the inclusion of the EXP-Block leads to a notable improvement in the performance of the original methods, with gains ranging from 0.0 to 4.4. These results highlight the versatility and

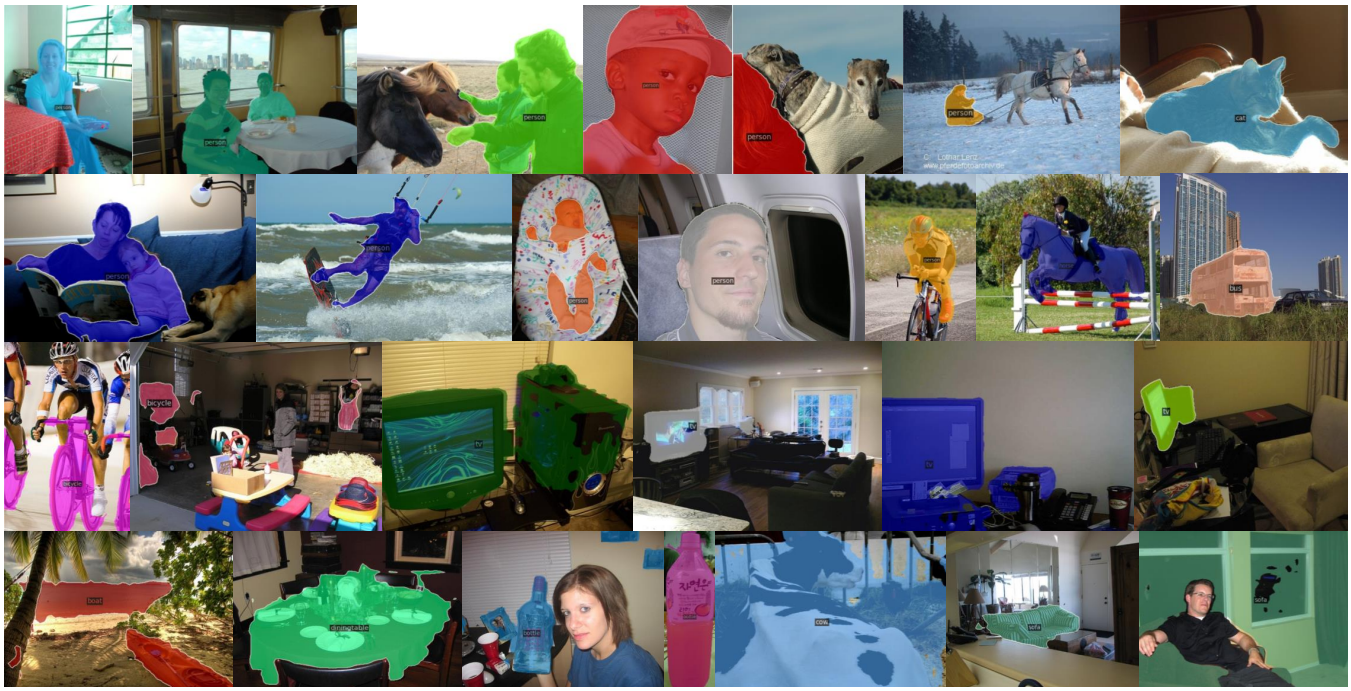

Figure 3: Visualization results of segmentation on Pascal VOC. We only visualize the mask of the annotated category.

practicality of the EXP-Block as an auxiliary module. In summary, this experiment provides strong evidence of the EXP-Block's effectiveness and adaptability, confirming its utility as a general-purpose enhancement component for a wide range of methods.

Ablation Study

Overall We conduct comprehensive ablation studies to evaluate the efficiency of the EXP-Block. The results are detailed in Tab. 3, 4. First, let us assess the overall impact of the EXP-Block on our methodology. A comparison between the rows ① and ⑤ of Tab.3 clearly shows that the EXP-Block substantially boosts the performance of the ELSE-Net, yielding an improvement of approximately 5.5 points. Subsequently, we will analyze the individual contributions of each component.

CLIPN Text Encoder The CLIPN Text Encoder is pivotal within the EXP-Block, serving as the foundation for reverse thinking. Here, we focus on the impacts of freezing versus fine-tuning on the outcomes. A comparison between rows ② and ⑤ in Tab.3 clearly demonstrates that fine-tuning significantly enhances the mIOU by 2.2 points compared to freezing, underscoring the benefits of fine-tuning.

Reverse Retrieval Adapter The R^2-Adapter is specifically designed to enhance the integration of reverse text embedding with visual embedding, consequently producing high-quality exclusion probabilities. A comparison of rows ③ and ⑤ in Tab.3 clearly shows that the R^2-Adapter is essential to the process, yielding improvements of 0.2 mIOU.

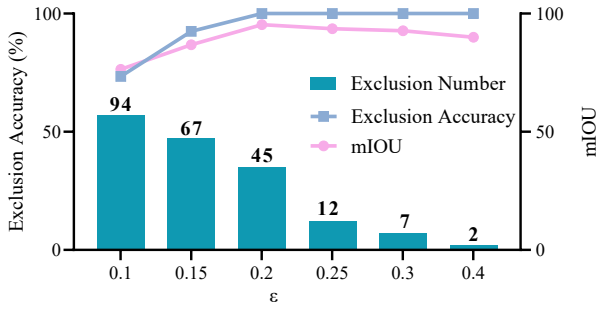

Figure 4: Ablation study to evaluate the threshold ϵ on VOC.

Probability Correction The probability correction process is the final step toward achieving "excluding the impossible". The corresponding experiments are detailed in rows ④ and ⑤ of Tab.3. Row ④ depicts the scenario without the probability correction process, utilizing the most naive approach where the final probability is computed as $\mathbf{P}_{in} \times (1 - \mathbf{P}_{ex})$. Observations reveal that the application of the probability correction process enhances the model's performance by 1.3 points.

In our method, threshold ϵ is critical; it dictates the number of samples to be excluded within the EXP-Block, the exclusion accuracy (the accuracy that the excluded categories are not the ground truth categories), and its overall impact on the mIOU. As demonstrated in Fig. 4, increasing the ϵ results in fewer exclusions, yet the precision of these exclusions improves. Nevertheless, the effect on the final outcome

Method		ADE-847	ADE-150	PC-59	VOC
SAN (CVPR'2023) (Xu et al. 2023)	Original	10.1	27.5	53.8	94.0
	Reproduced	7.9	26.9	53.2	94.2
	Reproduced + EXP-Block	**10.1** (+2.2)	**27.4** (+0.5)	**53.7** (+0.5)	**94.2** (+0.0)
Freeseg (CVPR'2023) (Qin et al. 2023)	Original	/	**28.6**	/	82.6
	Reproduced	/	25.6	/	80.2
	Reproduced + EXP-Block	/ /	27.6 (+2.0)	/ /	**82.7** (+2.5)
OVSeg (CVPR'2023) (Liang et al. 2023)	Original	7.1	24.8	53.3	92.6
	Reproduced	5.5	25.6	49.8	93.2
	Reproduced + EXP-Block	**8.3** (+2.8)	**26.4** (+0.8)	**54.2** (+4.4)	**93.4** (+0.2)

Table 2: Evaluation of the model-agnostic functionality of the EXP-Block. *Original* represents the baseline results as reported in the original papers. *Reproduced* denotes our reproduction of the baseline results. *+ EXP* illustrates the enhanced baseline results with the addition of the EXP-Block. Improvements over the reproduction are highlighted in red.

	CLIPN Text Encoder		R^2-Adapter	Prob Correction	mIOU
	Frozen	Tunable			
①					89.6
②	✓		✓	✓	92.9
③		✓		✓	94.9
④		✓	✓		93.8
⑤		✓	✓	✓	**95.1**

Table 3: Ablation studies of different components in EXP-Block on VOC. ① denotes the ELSE-Net w/o EXP-Block.

initially rises before declining. All experiments were conducted on the VOC dataset using both the tunable CLIPN text encoder and the R^2-Adapter.

Extra Consumption

Intuitively, adding an extra EXP-Block will inevitably increase the consumption of the original method, but these additional costs are entirely within an acceptable range, as detailed in Tab.4. It takes the OVSeg (Liang et al. 2023) as an example, trained on the COCO dataset. The OVSeg takes 32.5 hours with 147.2M parameters; with the EXP-Block added, it takes 35 hours with 211.2M parameters. In the inference phase, the time for OVSeg to process a single image is 1.25s, whereas OVSeg+EXP-Block takes 1.8s. With such a modest increase in resource usage, the EXP-Block can significantly enhance model performance by 0.2 to 4.4 points, further proving the method's practicality.

Conclusion

To tackle the open vocabulary semantic segmentation challenge, we introduce the "Excluding the Impossible" concept. Based on this concept, we developed the ELSE-Net,

Method	Param (M)	Training Time	Inference Time
OVSeg	147.2	32.5h	1.25s
OVSeg + EXP	211.2	35h	1.8s

Table 4: Experiments of consumption on one 4090 GPU. The training time specified covers all training data. Inference time is calculated for each test image.

which consists of the GP-Block and the model-agnostic EXP-Block. Experiments across four benchmark datasets have assessed its efficiency. Looking ahead, our research will pivot to two crucial areas: (1) We have noticed that our exclusion method occasionally leads to ineffective exclusions, which offer limited benefits to the forward process. Consequently, developing a more effective and innovative exclusion strategy is one of our key research priorities moving forward. (2) Our implementation of "Excluding the Impossible" utilizes the existing CLIPN, which is not fully compatible with the forward model. Thus, we aim to retrain a tailored model for this specific task.

Acknowledgments

This work was supported by Shandong Natural Science Foundation under Grants ZR2024MF102, ZR2023MF008, National Natural Science Foundation of China under Grant 62372468, Major Basic Research Projects in Shandong Province under Grant ZR2023ZD32, Qingdao Natural Science Foundation under Grant 23-2-1-161-zyyd-jch.

References

Caesar, H.; Uijlings, J.; and Ferrari, V. 2018. Coco-stuff: Thing and stuff classes in context. In *CVPR*, 1209–1218.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 1290–1299.

Cho, S.; Shin, H.; Hong, S.; Arnab, A.; Seo, P. H.; and Kim, S. 2024. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *CVPR*, 4113–4123.

Chuang, C.-Y.; Jampani, V.; Li, Y.; Torralba, A.; and Jegelka, S. 2023. Debiasing vision-language models via biased prompts. *arXiv preprint arXiv:2302.00070*.

Ding, J.; Xue, N.; Xia, G.-S.; and Dai, D. 2022. Decoupling zero-shot semantic segmentation. In *CVPR*, 11583–11592.

Ding, Z.; Wang, J.; and Tu, Z. 2023. Open-vocabulary universal image segmentation with MaskCLIP. In *ICML*, 8090–8102.

Everingham, M.; Eslami, S. A.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2015. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111: 98–136.

Ghiasi, G.; Gu, X.; Cui, Y.; and Lin, T.-Y. 2022. Scaling open-vocabulary image segmentation with image-level labels. In *ECCV*, 540–557. Springer.

Han, C.; Zhong, Y.; Li, D.; Han, K.; and Ma, L. 2023a. Open-vocabulary semantic segmentation with decoupled one-pass network. In *CVPR*, 1086–1096.

Han, K.; Liu, Y.; Liew, J. H.; Ding, H.; Liu, J.; Wang, Y.; Tang, Y.; Yang, Y.; Feng, J.; Zhao, Y.; et al. 2023b. Global knowledge calibration for fast open-vocabulary segmentation. In *ICCV*, 797–807.

Jadon, S. 2020. A survey of loss functions for semantic segmentation. In *2020 IEEE conference on computational intelligence in bioinformatics and computational biology (CIBCB)*, 1–7. IEEE.

Li, B.; Weinberger, K. Q.; Belongie, S.; Koltun, V.; and Ranftl, R. 2022. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*.

Liang, F.; Wu, B.; Dai, X.; Li, K.; Zhao, Y.; Zhang, H.; Zhang, P.; Vajda, P.; and Marculescu, D. 2023. Open-vocabulary semantic segmentation with mask-adapted clip. In *CVPR*, 7061–7070.

Mottaghi, R.; Chen, X.; Liu, X.; Cho, N.-G.; Lee, S.-W.; Fidler, S.; Urtasun, R.; and Yuille, A. 2014. The role of context for object detection and semantic segmentation in the wild. In *CVPR*, 891–898.

Parashar, S.; Lin, Z.; Liu, T.; Dong, X.; Li, Y.; Ramanan, D.; Caverlee, J.; and Kong, S. 2024. The Neglected Tails in Vision-Language Models. In *CVPR*, 12988–12997.

Qin, J.; Wu, J.; Yan, P.; Li, M.; Yuxi, R.; Xiao, X.; Wang, Y.; Wang, R.; Wen, S.; Pan, X.; et al. 2023. Freeseg: Unified, universal and open-vocabulary image segmentation. In *CVPR*, 19446–19455.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763. PMLR.

Shao, S.; Bai, Y.; Wang, Y.; Liu, B.; and Zhou, Y. 2024. DeIL: Direct-and-Inverse CLIP for Open-World Few-Shot Learning. In *CVPR*, 28505–28514.

Wang, H.; Li, Y.; Yao, H.; and Li, X. 2023. Clipn for zero-shot ood detection: Teaching clip to say no. In *ICCV*, 1802–1812.

Xie, B.; Cao, J.; Xie, J.; Khan, F. S.; and Pang, Y. 2024. Sed: A simple encoder-decoder for open-vocabulary semantic segmentation. In *CVPR*, 3426–3436.

Xu, M.; Zhang, Z.; Wei, F.; Hu, H.; and Bai, X. 2023. Side adapter network for open-vocabulary semantic segmentation. In *CVPR*, 2945–2954.

Xu, M.; Zhang, Z.; Wei, F.; Lin, Y.; Cao, Y.; Hu, H.; and Bai, X. 2022. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *ECCV*, 736–753. Springer.

Xu, W.; Xu, R.; Wang, C.; Xu, S.; Guo, L.; Zhang, M.; and Zhang, X. 2024. Spectral prompt tuning: Unveiling unseen classes for zero-shot semantic segmentation. In *AAAI*, 6369–6377.

Yu, Q.; He, J.; Deng, X.; Shen, X.; and Chen, L.-C. 2024. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. *Neurips*, 36.

Zhou, B.; Zhao, H.; Puig, X.; Fidler, S.; Barriuso, A.; and Torralba, A. 2017. Scene parsing through ade20k dataset. In *CVPR*, 633–641.