Physically-guided Open Vocabulary Segmentation with Weighted Patched Alignment Loss

Weide Liu[a], Jieming Lou[b], Xingxing Wang[c], Wei Zhou[d], Jun Cheng[c] and Xulei Yang[c]

[a]*Harvard Medical School*
[b]*National University of Singapore*
[c]*Institute for Infocomm Research, Agency for Science, Technology and Research*
[d]*Cardiff University*

ARTICLE INFO

Keywords:
Physics-informed
Open Vocabulary
Segmentation
Patched Alignment Loss

ABSTRACT

Open vocabulary segmentation is a challenging task that aims to segment out the thousands of unseen categories. Directly applying CLIP to open-vocabulary semantic segmentation is challenging due to the granularity gap between its image-level contrastive learning and the pixel-level recognition required for segmentation. To address these challenges, we propose a unified pipeline that leverages physical structure regularization to enhance the generalizability and robustness of open vocabulary segmentation. By incorporating physical structure information, which is independent of the training data, we aim to reduce bias and improve the model's performance on unseen classes. We utilize low-level structures such as edges and keypoints as regularization terms, as they are easier to obtain and strongly correlated with segmentation boundary information. These structures are used as pseudo-ground truth to supervise the model. Furthermore, inspired by the effectiveness of comparative learning in human cognition, we introduce the weighted patched alignment loss. This loss function contrasts similar and dissimilar samples to acquire low-dimensional representations that capture the distinctions between different object classes. By incorporating physical knowledge and leveraging weighted patched alignment loss, we aim to improve the model's generalizability, robustness, and capability to recognize diverse object classes. The experiments on the COCO Stuff, Pascal VOC, Pascal Context-59, Pascal Context-459, ADE20K-150, and ADE20K-847 datasets demonstrate that our proposed method consistently improves baselines and achieves new state-of-the-art in the open vocabulary segmentation task.

1. Introduction

Image segmentation is a fundamental task in computer vision recognition which requests costly dense annotations. Existing deep learning-based semantic segmentation methods heavily rely on large amounts of labeled data. However, currently, the datasets often consist of only tens to hundreds of categories, and the expensive process of data collection and annotation limits our ability to expand the categories further. Furthermore, in practical scenarios, new objects frequently appear, but obtaining sufficient annotations for these novel objects is often impractical and challenging.

The recent development of the large-scale vision-language model, CLIP Radford, Kim, Hallacy, Ramesh, Goh, Agarwal, Sastry, Askell, Mishkin, Clark et al. (2021), has marked a significant advancement in the field of image recognition, specifically in open-vocabulary image classification. This enables recognition of arbitrary categories at the image level, a notable success in the domain. Motivated by this success, researchers are now exploring the potential of adapting such models to semantic segmentation. The goal is to achieve a human-like understanding of scenes, which typically involves recognizing thousands of categories in an open-vocabulary manner.

However, applying the CLIP model to open-vocabulary semantic segmentation Li, Weinberger, Belongie, Koltun and Ranftl (2022); Zhou, Loy and Dai (2021); Ghiasi, Gu, Cui and Lin (2022); Liang, Wu, Dai, Li, Zhao, Zhang, Zhang, Vajda and Marculescu (2022) presents significant challenges. The core issue lies in the fact that CLIP is trained through image-level contrastive learning, which does not inherently provide the pixel-level recognition capability essential for effective semantic segmentation. A proposed solution to overcome this granularity gap is to fine-tune the CLIP model on segmentation datasets. Nonetheless, this approach has its limitations. Segmentation datasets are substantially smaller in size compared to the expansive vision-language pre-training datasets. This size discrepancy often leads to a diminishment in the recognition capabilities of the fine-tuned models when applied to open-vocabulary tasks.

To alleviate directly finetuning the CLIP model to the segmentation task, another solution is to freeze the CLIP features and make the segmentation model adapt to the learned open vocabulary classification model Xu, Zhang, Wei, Lin, Cao, Hu and Bai (2022); Ding, Wang and Tu (2022b). In particular, the two-stage approaches have been proposed to first generate the class-agnostic mask proposals and then leverage pre-trained CLIP for open-vocabulary classification to identify the labels for each pixel. However, the success of these methods relies on two assumptions: (1) the model

✉ weide.liu@childrens.harvard.edu (W. Liu); jieminglou22@u.nus.edu (J. Lou); wangxingxing@outlook.com (X. Wang); zhouw26@cardiff.ac.uk (W. Zhou); cheng_jun@i2r.a-star.edu.sg (J. Cheng); yang_xulei@i2r.a-star.edu.sg (X. Yang)

ORCID(s): 0000-0002-9855-4479 (W. Liu); 0000-0003-3641-1429 (W. Zhou); 0000-0003-1786-6188 (J. Cheng); 0000-0002-7002-4564 (X. Yang)

can generate accurate class-agnostic mask proposals, and (2) pre-trained CLIP can accurately transfer its classification performance to masked image proposals.

However, a study conducted by Ovseg Liang, Wu, Dai, Li, Zhao, Zhang, Zhang, Vajda and Marculescu (2023) highlights a critical limitation in this approach. Their findings reveal that while class-agnostic masks generally succeed in locating objects within images, they often fall short in assigning precise class labels. This issue was examined in two distinct scenarios: in particular, as shown in Figure 1, they utilized an "oracle" mask generator alongside an ordinary CLIP classifier, and the study employed ground-truth masks as region proposals. These were then classified using a pre-trained CLIP model. Under these conditions, the model achieved a mIoU of only 20.1% on the ADE20K-150 dataset, indicating a significant shortfall in classification accuracy. Conversely, when assuming an "oracle" classifier paired with an ordinary mask proposal generator, the model's performance improved markedly. In this setup, masked region proposals were extracted and each region was compared with ground-truth object masks to ascertain the object with the highest overlap, subsequently assigning the corresponding object label to the extracted region. Despite the imperfect nature of the region proposals, this approach yielded a considerably higher mIoU of 66.5%.

This analysis clearly demonstrates that pre-trained CLIP cannot satisfactorily classify masked images, serving as the performance bottleneck for two-stage open-vocabulary segmentation models. This limitation arises due to the purely data-driven nature of current open-vocabulary methods. To solve this challenging problem, Ovseg Liang et al. (2023) proposed to adapt the CLIP by finetuning it on masked images and corresponding text labels through training data collection by mining an existing image-caption dataset. However, training such models can be challenging and often requires a large amount of training data, which contradicts the few-shot setting of both tasks.

Recent progress in deep learning suggests that the backbone network should be capable of extracting general representative features for both seen and unseen class images, regardless of the training data. However, purely data-driven models optimized for current open-vocabulary methods may lead to locally optimal representations that are biased toward the training data. This bias can hinder the model's generalization in accurately recognizing the newly appeared classes, especially those with similar patterns to the base classes. To overcome this limitation, a potential solution is to use additional tasks as constraints to regularize the training.

To address these challenges, we propose to utilize a physical structure of regularized information to regulate the training of the open vocabulary segmentation model. As illustrated in Figure 2, our method includes a physical structure information prediction module through multi-task learning. This module aims to enhance the representation capability of the feature extraction module and reduce bias in the segmentation prediction, particularly for the unseen classes.

We utilize low-level structures such as keypoints as constraints for regularization. These low-level structures are easier to obtain than semantic-text labels and can be considered physical structure information that is independent of the training data for segmentation class prediction, yet strongly correlated with segmentation boundary information.

To overcome the bottleneck issue of identifying classes for masked images, an alternative and intuitive approach is to differentiate between different objects. Extensive research He, Fan, Wu, Xie and Girshick (2020); Grill, Strub, Altché, Tallec, Richemond, Buchatskaya, Doersch, Pires, Guo, Azar et al. (2020b); Chen, Kornblith, Norouzi and Hinton (2020a) has demonstrated that children grasp new concepts more easily when they compare an image containing a dog to other images featuring dogs, enabling them to infer that the target image represents a dog, as opposed to merely reading about animals in a book. This raises the question: What makes this comparative method more effective? The effectiveness of this approach can be attributed to the fact that individuals with limited prior knowledge, like children, find it simpler to learn new things by contrasting similarities and differences, rather than attempting to recognize each item individually. Initially, a child may encounter challenges in identifying a dog. However, over time, the child learns to discern the shared characteristics among dogs, such as the shape of their nose and their body posture.

Inspired by these insightful studies, we propose a weighted patch-aligned contrastive loss that aims to acquire low-dimensional representations of data by contrasting similar and dissimilar samples. This approach mirrors the way a child navigates the process of recognizing a new object. By leveraging the proposed weighted patch-aligned contrastive loss, we aim to enhance the open vocabulary segmentation task by effectively capturing the distinctions between different object classes.

However, the previous CLIP model primarily focused on image-level alignment between categories and the entire image, making it less suitable for segmentation tasks for the following reasons: 1) Image segmentation involves pixel-level classification, where both local and global representations play crucial roles. However, CLIP was designed primarily for image-level representations and does not fully capture the pixel-level details required for segmentation tasks. 2) In typical images, multiple objects coexist, such as a keyboard, desk, and computer. However, the global contrastive loss used in CLIP cannot effectively distinguish between these different objects within the same image. To address these limitations, we propose a novel approach by dividing each image into patches to align it more effectively with text classification. This patch-based approach allows us to achieve a balance of weights within each image by calculating the pixel numbers, which are then used to compute a weighted sum over vision features. This ensures that each patch's contribution is appropriately considered, leading to more accurate segmentation results.

1. We propose physical consistency loss for open vocabulary segmentation to enforce spatial coherence

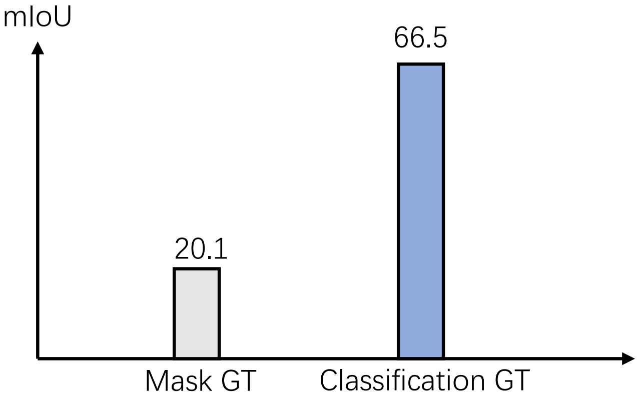

Figure 1: Given the mask ground truth as region proposals and fed them into a pre-trained CLIP for classification. This model achieved only a mean Intersection over Union (mIoU) of 20.1% on the ADE20K-150 dataset. Next, given the model with the classification ground truth, but the mask will be predicted by the model. Despite imperfect region proposals, this model achieved a significantly higher mIoU of 66.5%. This analysis clearly demonstrates that pre-trained CLIP cannot satisfactorily classify masked images, serving as the performance bottleneck for two-stage open-vocabulary segmentation models. This limitation arises due to the purely data-driven nature of current open-vocabulary methods.

and structural consistency in segmentation outputs to reduce the bias due to the limitation of the pure data-driven model.

2. To further enhance discrimination between different object classes, we introduce the weighted patch-aligned contrastive loss, which leverages the comparative approach to improve the model's ability to capture the distinctions between classes. By integrating the weighted patch-aligned contrastive loss, our model achieves efficient discrimination capabilities, leading to more accurate and refined segmentation outputs.
3. Our proposed method improves the open-vocabulary segmentation baselines significantly and achieves new state-of-the-art results on several datasets, including COCO Stuff, ADE20K, Pascal VOC, and Pascal Context.

2. Related Work

2.1. Large Foundation Vision-language Models

The objective of large foundation visual-language is to acquire generic representations of vision and language. Early studies in this domain Su, Zhu, Cao, Li, Lu, Wei and Dai (2019); Lu, Batra, Parikh and Lee (2019); Chen, Li, Yu, El Kholy, Ahmed, Gan, Cheng and Liu (2020c); Li, Yin, Li, Zhang, Hu, Zhang, Wang, Hu, Dong, Wei et al. (2020) primarily followed a two-step approach. Initially, models were pre-trained on visual and language data of moderate size. Subsequently, fine-tuning was conducted on downstream visual-language tasks, such as Visual Question Answering (VQA) Antol, Agrawal, Lu, Mitchell, Batra, Zitnick and Parikh (2015) and image captioning, to assess the advantages of pre-training. However, recent advancements with CLIPRadford et al. (2021) and ALIGN Jia, Yang, Xia, Chen, Parekh, Pham, Le, Sung, Li and Duerig (2021) have demonstrated that visual-language models pre-trained on large-scale noisy text-image pairs possess open-vocabulary recognition capabilities, thus serving as a robust foundation for downstream tasks. Numerous recent studies have verified this observation and achieved remarkable performance in open-vocabulary image recognition Yuan, Chen, Chen, Codella, Dai, Gao, Hu, Huang, Li, Li, Liu, Liu, Liu, Lu, Shi, Wang, Wang, Xiao, Xiao, Yang, Zeng, Zhou and Zhang (2021); Yu, Wang, Vasudevan, Yeung, Seyedhosseini and Wu (2022); Alayrac, Donahue, Luc, Miech, Barr, Hasson, Lenc, Mensch, Millican, Reynolds, Ring, Rutherford, Cabi, Han, Gong, Samangooei, Monteiro, Menick, Borgeaud, Brock, Nematzadeh, Sharifzadeh, Binkowski, Barreira, Vinyals, Zisserman and Simonyan (2022) as well as other related tasks Gu, Lin, Kuo and Cui (2021); Wang, Lu, Li, Tao, Guo, Gong and Liu (2022c); Hessel, Holtzman, Forbes, Bras and Choi (2021); Patashnik, Wu, Shechtman, Cohen-Or and Lischinski (2021). T-MASS Wang, Sun, Wang, Liu, Dianat, Rabbani, Rao and Tao (2024) introduces a stochastic text modelling method, treating text as a stochastic embedding to enhance semantic flexibility and resilience. Meanwhile, LLaVA-Med Li, Wong, Zhang, Usuyama, Liu, Yang, Naumann, Poon and Gao (2024a) presents a cost-effective vision-language conversational assistant, specifically trained to address open-ended questions about biomedical images. Furthermore, E^{2}VPT Han, Wang, Cui, Cao, Wang, Qi and Liu (2023) improves model fine-tuning through the integration of learnable prompts within key model layers, complemented by a prompt pruning procedure that efficiently preserves performance by selectively removing less critical prompts.

In this paper, we also utilize the pre-trained foundation vision-language model for our open-vocabulary segmentation prediction.

2.2. Open-vocabulary Semantic Segmentation

Previous research Zhao, Puig, Zhou, Fidler and Torralba (2017); Xian, Choudhury, He, Schiele and Akata (2019); Bucher, Vu, Cord and Pérez (2019); Mukhoti, Lin, Poursaeed, Wang, Shah, Torr and Lim (2023); Liu, Zhang, Lin and Liu (2020) on open-vocabulary semantic segmentation has focused on learning a joint embedding space that connects image pixels with class names or descriptions. More recently, inspired by the effectiveness of large-scale vision-language pre-training models in open-vocabulary recognition, several approaches have explored their application in the context of open-vocabulary semantic segmentation. Some of these approaches Li et al. (2022); Zhou et al. (2021); Ghiasi et al. (2022); Liang et al. (2022); Qin, Han, Wang, Nie, Yin and Xiankai (2023); Liu, Wu, Zhao, Fang, Foo, Cheng and Lin (2024) involve fine-tuning the vision-language pre-training models. However, this either requires a substantial amount of additional data or compromises

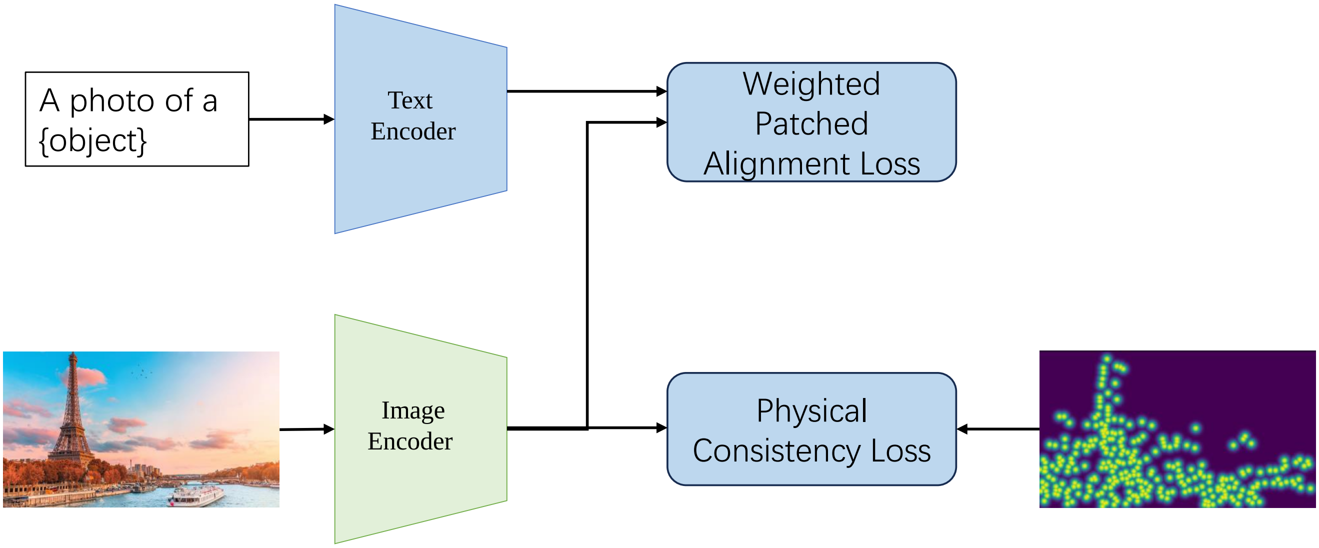

Figure 2: As depicted in the figures, we introduce a physical consistency loss to enhance open vocabulary segmentation. The physical consistency loss enforces spatial coherence and structural consistency in segmentation outputs, mitigating bias arising from pure data-driven models. To augment discrimination among object classes, we introduce the weighted patch-aligned contrastive loss. This novel loss leverages a comparative strategy, enhancing the model's capacity to distinguish between classes. Through the integration of the weighted patch-aligned contrastive loss, our model attains enhanced discrimination abilities, resulting in improved accuracy and refinement of segmentation outputs.

the open-vocabulary capability of the vision-language pre-training model.

An alternative framework called SimSeg Xu et al. (2022) has been proposed, which operates in two stages: first, it generates masked image crops, and then it recognizes these crops using a frozen CLIP model. However, SimSeg relies on a cumbersome mask generator and requires multiple forward passes of CLIP, resulting in inefficiency in terms of both model size and inference speed. Additionally, the mask generator lacks awareness of the CLIP model, further limiting its performance.

OMG-Seg Li, Yuan, Li, Ding, Wu, Zhang, Li, Chen and Loy (2024b) introduces a versatile, transformer-based model capable of handling multiple segmentation tasks, including image and video segmentation, with notably reduced computational demands. Similarly, LEVOS Lu, Zhang, Sun, Guo, Cao, Fei, Yang and Chen (2023) offers an unsupervised video object detection approach utilizing simulated dense labels and motion cues to enhance detection accuracy. Furthermore, SG-Net Liu, Cui, Tan and Chen (2021) presents a unified architecture that combines detection, segmentation, and tracking, promoting enhanced feature sharing and joint optimization across these tasks.

MaskCLIP Ding et al. (2022b) addresses some of the limitations of the two-stage framework by progressively refining the predicted masks using the CLIP encoder and incorporating masks in attention layers, following a similar approach introduced in Cheng, Misra, Schwing, Kirillov and Girdhar (2022). However, MaskCLIP still relies on a heavy mask generator, the initial mask prediction remains CLIP-unaware, and the mask prediction and recognition processes are coupled together. However, the previous CLIP-based models, while effective for image-level classification, fall short in segmentation tasks due to their lack of pixel-level detail and inability to distinguish between multiple objects in an image. To address these issues, we propose a novel patch-based approach that divides images into patches for better alignment with text classification, using pixel counts to compute a weighted sum over vision features for each patch.

2.3. Contrastive Loss

Currently self-supervised contrastive learning methods He et al. (2020); Grill et al. (2020b) have gained popularity for learning feature extractors from unlabeled images as pre-training for downstream tasks. These methods aim to bring the representations of different views of the same image closer while pushing away the representations of different images. For instance, SimCLR Chen et al. (2020a) presents a straightforward self-supervised learning framework by applying contrastive learning to representations of the same image under various data augmentations. On the other hand, MoCo He et al. (2020); Chen, Fan, Girshick and He (2020b) employs a moving average network (momentum encoder) to increase the negative memory bank size.

Beyond direct contrastive loss utilization, some approaches Grill, Strub, Altché, Tallec, Richemond, Buchatskaya, Doersch, Avila Pires, Guo, Gheshlaghi Azar et al. (2020a); Chen and He (2021) have found that learning consistent representations from positive pairings can also lead to reliable representations. BYOL Grill et al. (2020a), for instance, proposes training a network without negative samples. The method employs a Siamese encoder to encode features and minimizes the cosine similarity between the query and key embeddings to achieve good representations. However, BYOL requires a large batch size during training and may face convergence challenges. To address this, Simsiam Chen and He (2021) eliminates the momentum key encoder and adopts a stop-gradient strategy to prevent collapsing issues.

Moreover, in the context of enhancing encoder representation for downstream detection tasks, SCRL Roh, Shin, Kim and Kim (2021) extends consistency loss to the Region of Interest (ROI) in the intersection region of two views. For promoting consistency between embeddings, CO2 Wei, Wang, Shen and Yuille (2020) and RELIC Mitrovic, McWilliams, Walker, Buesing and Blundell (2020) introduce regularization using KL loss on embeddings generated from different data augmentations. These methods aim to refine the representations and improve the performance of the feature extractor for subsequent tasks.

DNC Wang, Han, Zhou and Liu (2022a) employs a non-parametric approach using sub-centroids of training data to represent class distributions, facilitating classification based on proximity in feature space. Wang et al. Wang and Liu (2021) explore the balance between uniformity and temperature t in contrastive learning, highlighting that while uniformity aids in feature distinction, excessive uniformity can compromise the semantic integrity crucial for downstream tasks. Wang et al. Wang, Liang and Liu (2022b) also propose a training framework that improves query-based models by enabling the learning of discriminative query embeddings. Furthermore, Zhao et al. describe a contrastive learning-based training strategy, starting with pretraining using a pixel-wise, label-based contrastive loss, followed by fine-tuning with cross-entropy loss, to enhance pixel classifier effectiveness Zhao, Vemulapalli, Mansfield, Gong, Green, Shapira and Wu (2021).

However, while all these contrastive methods focus on image-level features, they do not take into account the impact of the size of each object inside the image. In this paper, we propose a novel Weighted Patch-Level Contrastive Loss to align object features with open vocabulary category features. This new loss function aims to address the granularity gap by considering the size and importance of individual objects within the image during contrastive learning.

3. Method

As shown in Figure 2, to address the challenge of adapting the image-level pre-trained CLIP model to open-vocabulary semantic segmentation, we freeze the CLIP features and force the decoder to adapt to the learned open vocabulary segmentation model. To achieve this goal, we need accurate class-agnostic mask proposals and accurate transfer of CLIP's classification performance to masked image proposals, which is not always achieved due to the model being biased toward the trained classes and not able to identify the classification. In particular, most of the current models are able to know 'where' is the object but difficult to identify 'what' is the object.

To address these limitations, we propose physical structure regularization. By utilizing low-level structures like keypoints, which are less dependent on segmentation training data, the model's representation capability can be enhanced, and bias in segmentation prediction is reduced, especially for unseen classes. Further elaboration on this approach will be provided in Section 3.1.

Inspired by how humans learn by contrasting similarities and differences, we propose a weighted patched alignment loss to capture distinctions between different object classes and enhance open-vocabulary segmentation. Further elaboration on this approach will be provided in Section 3.2.

3.1. Physical Structure Regularization for Open Vocabulary Segmentation

In Figure 2, we present a unified pipeline designed to enhance the generalizability and robustness of open vocabulary segmentation. This is achieved through the incorporation of physical structure regularization. We treat the open vocabulary segmentation prediction as the primary task, with the regularization term serving as an auxiliary task. Both tasks are jointly trained to optimize the backbone, enhancing its robustness.

As shown in Figure 3, our approach employs a shared-parameter backbone, utilizing the segmentation's feature extractor parameters to generate multi-level features. These features are processed through a 1×1 convolution, encoding them into a 256-dimensional format to ensure spatial consistency and computational efficiency. Features from adjacent network layers, denoted as F_t and F_{t-1}, are fused by concatenation and then normalized. A subsequent 1×1 convolution, coupled with a Sigmoid layer, is used to predict the physical structure information, labeled as s_{est}. During model training, this prediction is compared with the ground truth low-level physical information, e_{gt}.

The calculation of the physical structure regularization loss L^p is defined as follows:

$$L^p = \mathcal{F}_{focal}(s_{gt}, s_{est}), \tag{1}$$

where s_{gt} and s_{est} represent the pseudo ground truth physical structures (e.g., edges or keypoints) and the estimated structures from our network, respectively. The function $\mathcal{F}_{focal}$ is employed to compute the focal loss Lin, Goyal, Girshick, He and Dollár (2017).

3.2. Weighted Patched-level Alignment Contrastive Loss

The weighted patch-level contrastive loss is designed to refine unimodal representations before fusion, serving as an intermediate loss function for image-text representations generated by transformer-based encoders. By utilizing an InfoNCE loss, we constrain positive and negative pairs of the projected vision and text features. As illustrated in Figure 4, we leverage original patch representations obtained from the transformer image encoder, focusing on alignment at the finest-grained level and avoiding reliance on the $[CLS]$

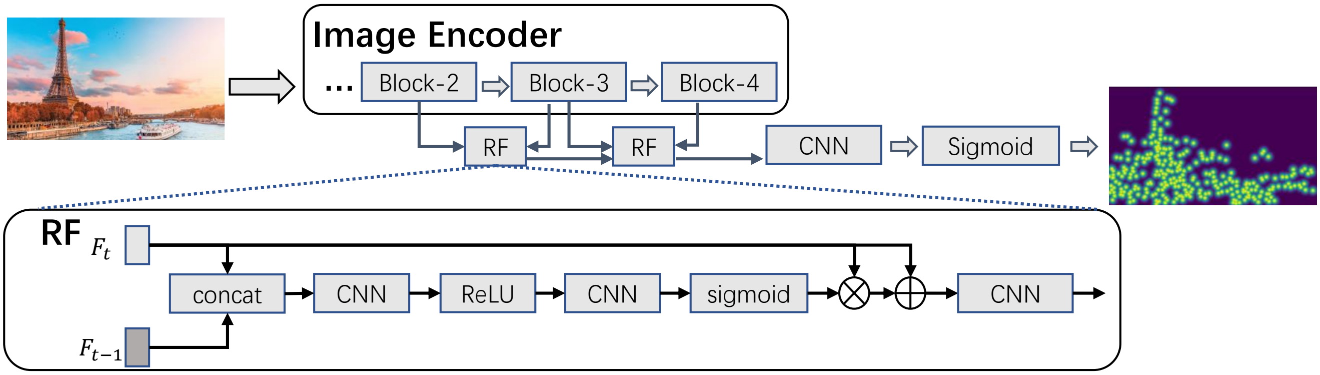

Figure 3: Architecture of the Physical Structure Prediction Network. Our network architecture involves the extraction of multi-level features and then being fed into three Refine modules (RF). Leveraging physical structure regularization, our method enhances feature representation and mitigates bias within the data-driven model, thereby facilitating more robust segmentation prediction.

token representation. Since a single image patch can encompass multiple semantic concepts, we propose contrasting image patches and texts with soft labels, calculated based on pixel counts.

For each image patch representation x_i^{patch} and text representation x_j^{text}, we compute the similarity sim as:

$$\text{sim}(x_i^{patch}, x_j^{text}) = o_{image}(x_i^{patch})^T \cdot o_{text}(x_j^{text}) \qquad (2)$$

where o_{image} and o_{text} are linear projection layers for the image and the text features.

With N image patches features in a batch and M text features of possible classes, the segmentation-aware contrastive loss L^{sac} can be defined as :

$$L^{sac} = \sum_{i=1}^{N} \frac{\sum_{j=1}^{M} w_{i,j} \exp\left(\frac{\text{sim}(x_i^{patch}, x_j^{text})}{\tau}\right)}{\sum_{j=1}^{M} \exp\left(\frac{\text{sim}(x_i^{patch}, x_j^{text})}{\tau}\right)} \qquad (3)$$

where τ is a learnable temperature parameter, and $w_{i,j}$ is the normalized weight based on the number of pixels, which is computed as the number of pixels in class j divided by the total number of pixels of the image patch i.

The proposed patch-level contrastive loss serves two primary objectives: (1) aligning image and text features to facilitate cross-modal learning within the multimodal encoder and (2) enhancing the image encoder's sensitivity to capture the semantic meaning of individual patches, a crucial aspect for tasks involving pixel-level segmentation.

In the realm of open vocabulary segmentation, leveraging vision-language contrastive learning often leads to overly broad textual descriptions. This generality arises from the practice of using templates that merely announce the presence of a certain object in an image without detailing other contextual elements, such as "a photo of an object". Consequently, background patches and those devoid of the object of interest can negatively impact training, disrupting feature learning. Our novel loss function harnesses the feature embeddings of different patches extracted by Vision Transformer (ViT) and computes their similarity with the text description, as illustrated in Equation 2. By doing so, without altering the architecture of ViT-based graph encoders, we introduce a finer level of granularity that effectively mitigates these training challenges. However, we acknowledge that operating at the patch level alone is not entirely sufficient due to the persistence of extraneous noise information within these patches. To address this limitation, we venture further by proposing a weighted approach for these patches. The core idea behind the weighting mechanism revolves around considering the proportion of samples within each patch that genuinely belong to the object class of interest. This strategic weighting allows the model to allocate training emphasis more rationally, thereby enhancing its ability to discern between relevant and irrelevant features during the learning process. The implementation of this notion is encapsulated in Equation 3, our proposed weighted patch-level contrastive loss. Through these detailed explanations and derivations, we aim to clarify the rationale and methodology behind our formulation, reinforcing its significance for advancing open vocabulary segmentation tasks.

Combining with the physical structure regularization loss L^p, the overall loss function used for the final optimization will be:

$$L = L^{sac} + L^p. \qquad (4)$$

4. Experiment

4.1. Datasets and Evaluation Metric

4.1.1. Datasets

For a fair comparison with previous methodsXu et al. (2023); Liang et al. (2023); Cho et al. (2023), our experiments are conducted on 6 datasets: COCO Stuff, ADE20K-150, ADE20K-847, Pascal Context-59, Pascal Context-459, and Pascal VOC. Following the previous methods, all experiments are trained on the training set of COCO Stuff and then directly evaluated on the remaining datasets.

COCO Stuff Lin et al. (2014) comprises 164K images with 171 annotated classes. It is divided into a training set

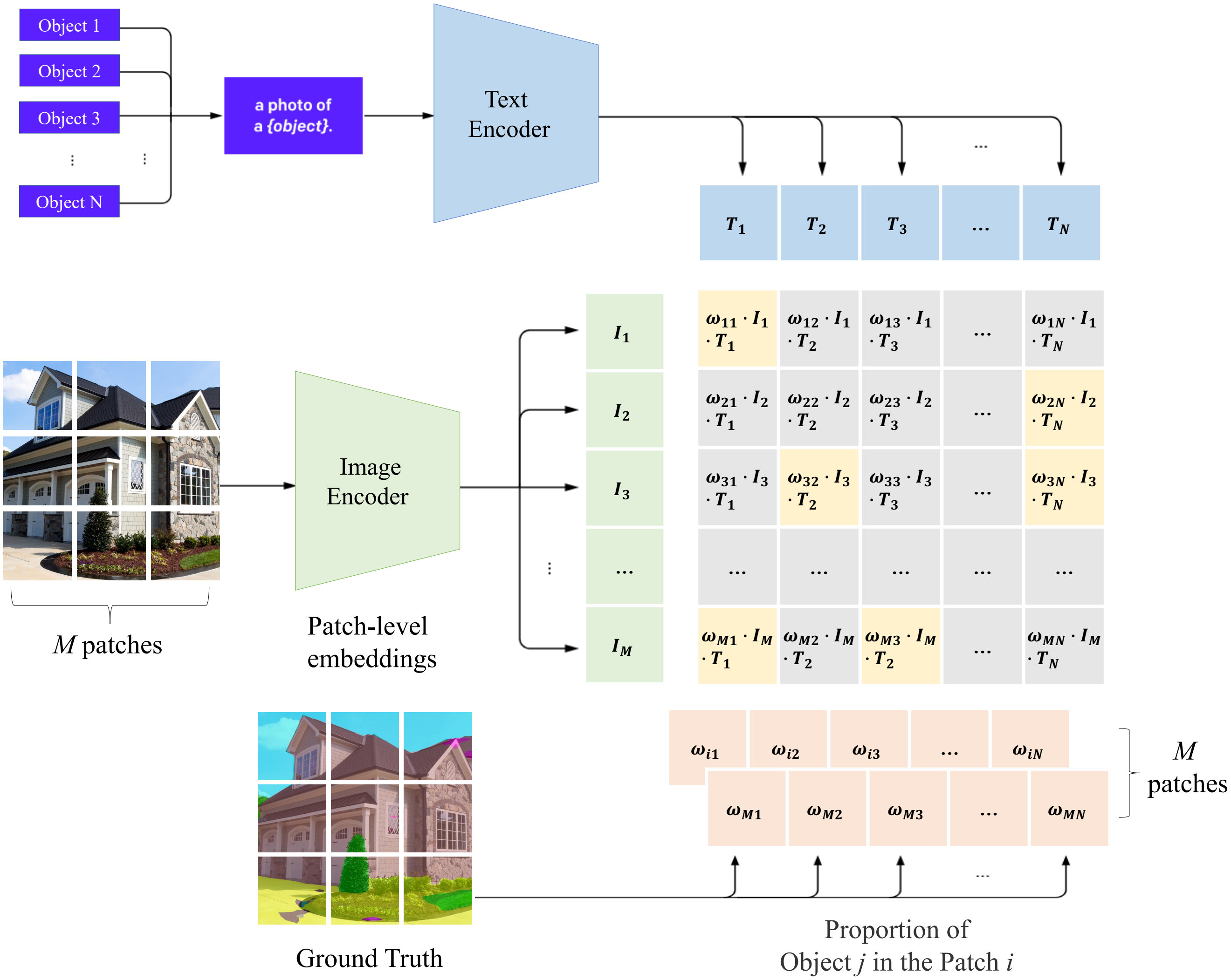
Object 1
Object 2
Object 3
Object N
a photo of a {object}.
Text Encoder
T_1
T_2
T_3
...
T_N
M patches
Image Encoder
Patch-level embeddings
I_1
I_2
I_3
...
I_M
$\omega_{11} \cdot I_1 \cdot T_1$
$\omega_{12} \cdot I_1 \cdot T_2$
$\omega_{13} \cdot I_1 \cdot T_3$
$\omega_{1N} \cdot I_1 \cdot T_N$
$\omega_{21} \cdot I_2 \cdot T_1$
$\omega_{22} \cdot I_2 \cdot T_2$
$\omega_{23} \cdot I_2 \cdot T_3$
$\omega_{2N} \cdot I_2 \cdot T_N$
$\omega_{31} \cdot I_3 \cdot T_1$
$\omega_{32} \cdot I_3 \cdot T_2$
$\omega_{33} \cdot I_3 \cdot T_3$
$\omega_{3N} \cdot I_3 \cdot T_N$
$\omega_{M1} \cdot I_M \cdot T_1$
$\omega_{M2} \cdot I_M \cdot T_2$
$\omega_{M3} \cdot I_M \cdot T_3$
$\omega_{MN} \cdot I_M \cdot T_N$
ω_{i1}
ω_{i2}
ω_{i3}
ω_{iN}
ω_{M1}
ω_{M2}
ω_{M3}
ω_{MN}
M patches
Ground Truth
Proportion of Object j in the Patch i

Figure 4: The overall structure of the proposed weighted patch-level contrastive loss. Our approach utilizes patch representations derived from the transformer image encoder, focusing on achieving alignment at the most granular level. Recognizing that a single image patch may contain multiple semantic concepts, we advocate for a novel technique of contrasting image patches with corresponding texts. This is facilitated by employing soft labels, which are calculated based on the count of pixels in each patch.

Table 1
The Category Similarity between various validate datasets and the COCO Stuff training dataset. Measured by Hausdorff distance and cosine similarity based on CLIP text encoder.

Dataset	Category Similarity
Pascal VOC	0.91
Pascal Context-59	0.86
Pascal Context-459	0.70
ADE20K-150	0.73
ADE20K-847	0.57

(118K images), a validation set (5K images), and a test set (41K images). In our experiments, we use the entire 118K training set as the default training data.

ADE20K-150 Zhou, Zhao, Puig, Fidler, Barriuso and Torralba (2017) is a large-scale scene understanding dataset with 20K training images and 2K validation images which contains a total of 150 classes.

ADE20K-847 Zhou et al. (2017) has the same set of images as ADE20K-150 but includes a greater number of annotated classes (847 classes). This dataset presents a challenge for open-vocabulary semantic segmentation.

Pascal VOC Everingham and Winn (2011) consists of 20 classes of semantic segmentation annotations. The training set contains 1464 images, and the validation set contains 1449 images. Following previous methods Cho et al. (2023), we report PAS-20 using the standard 20 object classes and also report the score for PAS-20^b, which defines the “background” as additional classes.

Pascal Context-59 Mottaghi, Chen, Liu, Cho, Lee, Fidler, Urtasun and Yuille (2014) is a dataset for semantic understanding, comprising 5K training images and 5K validation images. It includes a total of 59 most frequent annotated classes.

Table 2
Comparison with previous SOTA. The best-performing results are presented in bold. For a fair comparison, we utilize the mIoU as the evaluation metric. All our results are trained with full COCO-Stuff Lin, Maire, Belongie, Hays, Perona, Ramanan, Dollár and Zitnick (2014) dataset and tested on the validation datasets directly. Our method achieves the new state-of-the-art.

Methods	VLM	Training dataset	Additional dataset	A-847	PC-459	A-150	PC-59	PAS-20	PAS-20^b
SPNet Xian et al. (2019)	-	PASCAL VOC	-	-	-	-	24.3	18.3	-
ZS3Net Bucher et al. (2019)	-	PASCAL VOC	-	-	-	-	19.4	38.3	-
LSeg Li et al. (2022)	ViT-B/32	PASCAL VOC-15	-	-	-	-	-	47.4	-
ZegFormer Ding, Xue, Xia and Dai (2022a)	ViT-B/16	COCO-Stuff-156	-	4.9	9.1	16.9	42.8	86.2	62.7
ZegFormer† Ding et al. (2022a)	ViT-B/16	COCO-Stuff	-	5.6	10.4	18.0	45.5	89.5	65.5
ZSseg Xu et al. (2022)	ViT-B/16	COCO-Stuff	-	7.0	-	20.5	47.7	88.4	-
OpenSeg Ghiasi et al. (2022)	ALIGN	COCO Panoptic	Localized Narrative	4.4	7.9	17.5	40.1	-	63.8
OVSeg Liang et al. (2022)	ViT-B/16	COCO-Stuff	COCO Caption	7.1	11.0	24.8	53.3	92.6	-
CAT Cho, Shin, Hong, An, Lee, Arnab, Seo and Kim (2023)	ViT-B/16	COCO-Stuff	-	8.4	16.6	27.2	57.5	93.7	78.3
LSeg Li et al. (2022)	ViT-B/32	PASCAL VOC-15	-	-	-	-	-	52.3	-
OVSeg Liang et al. (2022)	ViT-L/14	COCO-Stuff	COCO Caption	9.0	12.4	29.6	55.7	94.5	-
SAN Xu, Zhang, Wei, Hu and Bai (2023)	ViT-L/14	COCO-Stuff	-	12.4	15.7	32.1	57.7	94.6	-
CAT Cho et al. (2023)	ViT-L/14	COCO-Stuff	-	10.8	20.4	31.5	62.0	96.6	81.8
CAT Cho et al. (2023)	ViT-H/14	COCO-Stuff	-	12.4	20.1	34.4	61.2	96.7	80.2
CAT Cho et al. (2023)	ViT-G/14	COCO-Stuff	-	13.3	21.4	36.2	61.5	97.1	81.4
Ours	ViT-L/14	COCO-Stuff	-	11.6	20.6	31.8	**62.1**	96.6	**82.0**
Ours	ViT-H/14	COCO-Stuff	-	13.1	20.1	34.5	61.3	96.7	80.1
Ours	ViT-G/14	COCO-Stuff	-	**14.4**	**21.6**	**36.3**	61.7	**97.2**	81.2

Table 3
Ablation studies on each proposed component. The L^p denotes the proposed physically regularized loss, and the L^{sac} denotes the weighted patched alignment loss. Our baseline is CAT Cho et al. (2023) For a fair comparison, we utilize the mIoU as the evaluation metric. All results are trained with full COCO-Stuff Lin et al. (2014) dataset and tested on the validation datasets directly.

Methods	A-847	PC-459	A-150	PC-59	PAS-20	PAS-20^b
Baseline	10.8	20.4	31.5	62.0	96.6	81.8
Baseline + L^p	11.5	20.5	31.6	62.2	96.6	82.0
Baseline + L^p + L^{sac}	11.6	20.6	31.8	62.1	96.6	82.0

Pascal Context-459 Mottaghi et al. (2014) shares the same images as Pascal Context-59 but includes a larger number of annotated classes (459 classes). This dataset is also widely used for open-vocabulary semantic segmentation.

Dataset Analysis: To analyze the capability of the trained model on the open set task, following the SAN Xu et al. (2023), we evaluate the relationship between the evaluated datasets to the trained dataset by calculating their similarity using the Hausdorff Distance. To compute the pairwise similarity, we extract the text embeddings of each concept using the pre-trained CLIP text encoder (ViT-L/14) and compute the cosine similarity. The results are presented in Table 1. Among the five validation datasets, Pascal VOC exhibits a high similarity score of up to 0.9, indicating their suitability for assessing the in-domain open-vocabulary ability in terms of visual categories. Conversely, Pascal Context-459, ADE20K-150, and ADE20K-847 yield lower similarity scores, suggesting their effectiveness in evaluating the cross-domain open-vocabulary ability, which is more challenging.

Table 4
Ablation studies on different kinds of physical information as regularization. the PL denotes the proposed physically regularized loss. For a fair comparison, we utilize the mIoU as the evaluation metric. All results are trained with full COCO-Stuff Lin et al. (2014) dataset and test on the validation datasets directly.

L^{sac}	SIFT	Superpoint	Canny	EdgeGT	A-847	PC-459	A-150	PC-59	PAS-20	PAS-20^b
✓					10.8	20.4	31.5	62.0	96.6	81.8
✓	✓				11.6	20.6	31.8	62.1	96.6	82.0
✓		✓			11.7	20.6	31.6	62.1	96.6	82.1
✓			✓		11.5	20.7	31.9	62.2	96.6	81.9
✓				✓	11.6	20.5	31.7	62.3	96.6	81.9
✓	✓	✓			11.7	20.8	31.8	62.0	96.8	82.0
✓	✓		✓		11.5	20.8	31.6	62.1	96.6	82.0
✓	✓	✓	✓	✓	11.6	20.7	31.7	62.2	96.8	82.1

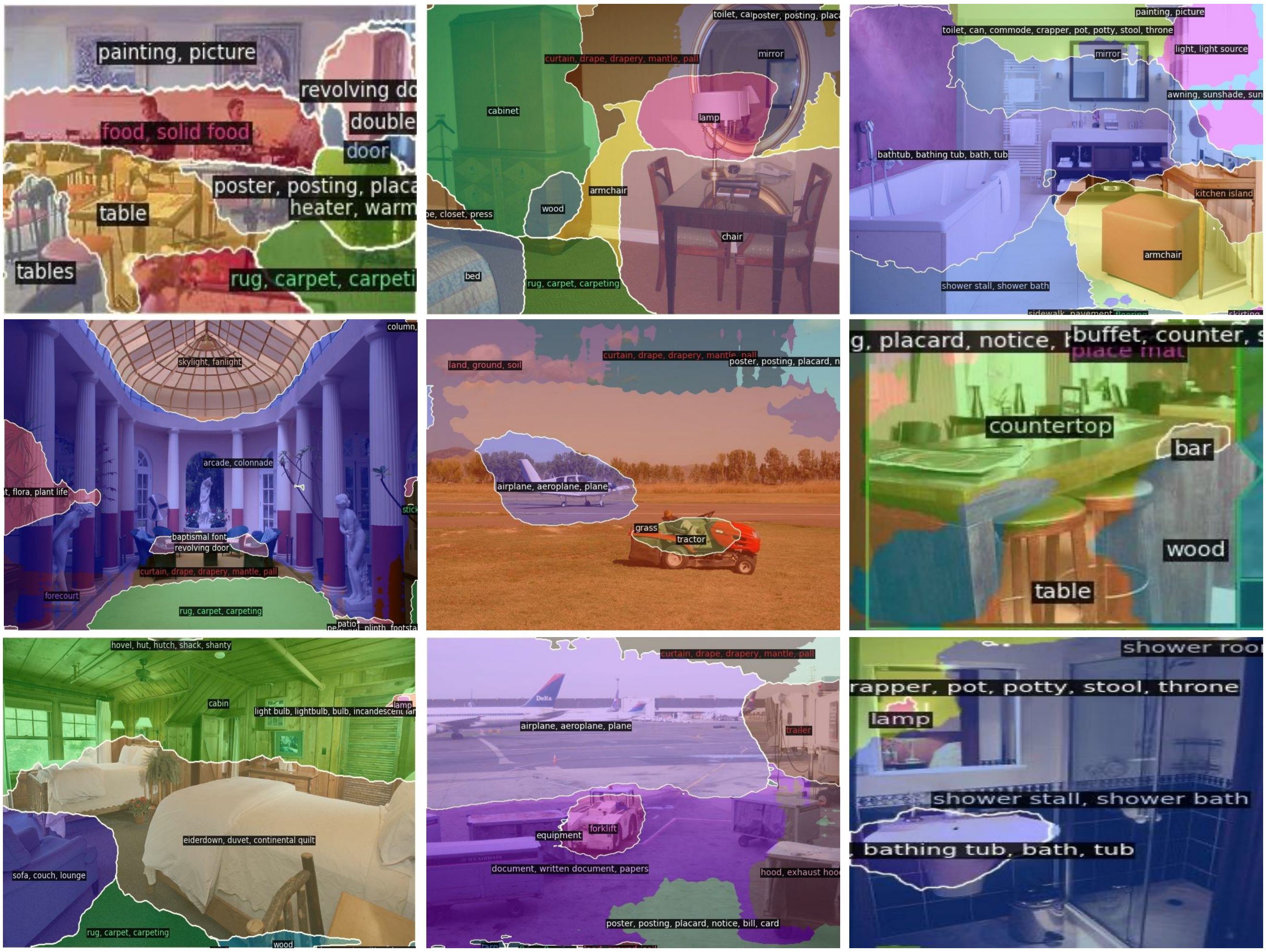

Figure 5: Segmentation Results on the ADE20K-847 Validation Dataset: The figures display segmentation masks for only the annotated categories within each image. Notably, these results demonstrate our model's capability to identify and locate objects across categories and datasets that were neither included in training nor previously encountered. This underscores the robustness and adaptability of our segmentation approach, effectively handling unseen categories and enhancing generalization.

4.1.2. Evaluation Metric

Following previous methods, we utilize the mean of class-wise intersection over union (mIoU) to validate the performance of our models.

4.2. Comparison with Previous Methods

4.2.1. Training Setting

All models are trained using the training set from the COCO Stuff dataset. Experimental settings such as the initial learning rate, weight decay, batch size, training iterations, and optimizers align with baselines. Following training, model evaluation is conducted directly on the validation datasets. Specifically, our implementation is based on PyTorch and utilizes Detectron2. We employ the AdamW optimizer with a learning rate of 2×10^{-4} for our model and 2×10^{-6} for CLIP, setting the weight decay to 10^{-4}. The batch size is configured to 4, utilizing 4 NVIDIA RTX A5000 GPUs for efficient training.

To clarify, our training process is end-to-end without employing pre-training strategies like contrastive learning. Instead, we integrate the novel loss function as a regularization term for open-vocabulary segmentation, a choice that aligns with our objective to directly optimize for the task at hand. This decision was made based on our intent to explore the effectiveness of the proposed loss within a straightforward and focused training framework.

4.2.2. Overall Comparison

Table 2 presents a comprehensive comparison of our method with the state-of-the-art approaches. Notably, all the methods utilize the CLIP ViT pre-trained models and are trained on the COCO Stuff dataset or larger datasets. Our method outperforms these approaches, demonstrating a substantial improvement for different visual-language models (VLM), including CLIP ViT-L/14, and ViT-H/14. Remarkably, the performance gains are most pronounced on the ADE-847 dataset. As indicated in Table 1, ADE-847 exhibits fewer classes similar to those in the training dataset COCO Stuff, which is the most challenging validate dataset.

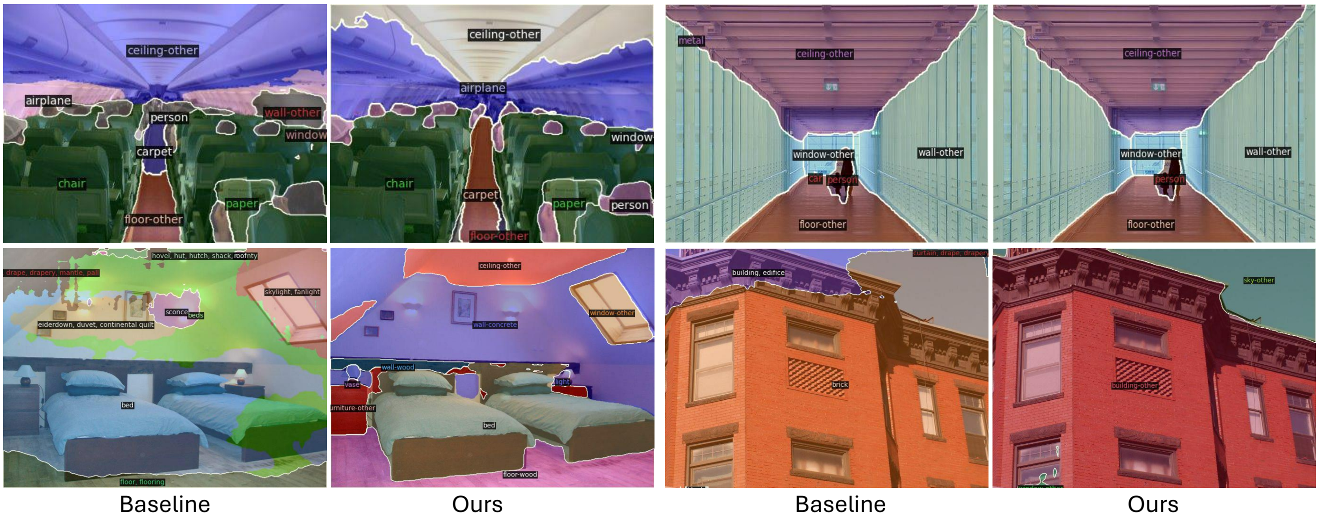

Figure 6: Comparison of Segmentation Results on the ADE20K Validation Dataset: Our method demonstrates improved segmentation results compared to the baseline.

This shows the superior open-vocabulary recognition capability of our approach.

4.3. Ablation Studies

The effectiveness of each component As demonstrated in Table 3, we conduct ablation experiments to illustrate the effectiveness of our proposed losses. Each component contributes to the improvement over the baseline. Combining these losses (L^p and L^{sac}), our method further enhances the baseline.

Different kinds of physical regularization To comprehensively validate the effectiveness of various low-level structural constraints in regularizing open vocabulary segmentation tasks, we conducted an ablation analysis involving the integration of edges and keypoints as distinct forms of physical structure in the regularization process.

For edge extraction, we explored two methodologies. First, we employed the Canny edge detection algorithm Canny (1986), a widely recognized technique in computer vision. Acknowledging potential limitations as Canny Edge may not be accurate, we adopted an alternative strategy for edge computation in the second method, generating ground truth edges using labels from semantic masks (EdgeGT). Moving beyond edge-based regularization, we investigated the utility of keypoints as a regularization mechanism. We employed two methodologies for computing keypoints as pseudo ground truths: the scale-invariant feature transform (SIFT) Lindeberg (2012) and Superpoint DeTone, Malisiewicz and Rabinovich (2018). Table 4 shows the results, demonstrating consistent and comparable performance across all types of physical structural regularization.

4.4. Result Visualization

As illustrated in Figure 5, we show our segmentation results on the ADE20K validation dataset. Notably, our method achieves precise object localization even without prior training on this dataset or access to specific category information.

Furthermore, we present a visual comparison of segmentation results between our method and the baseline in Figure 6. Our approach not only demonstrates enhanced segmentation accuracy but also reflects the robustness of our proposed method.

5. Conclusion

In this paper, we address the challenge of applying the CLIP model to open-vocabulary semantic segmentation. The CLIP model, trained using image-level contrastive learning, lacks the pixel-level recognition capability required for segmentation tasks. We propose a novel approach by dividing each image into patches and aligning them with text classification. This weighted patch-aligned contrastive loss enables more accurate segmentation results by effectively capturing distinctions between different object classes. Additionally, to mitigate bias issues in the base training datasets, we introduce a physical structure regularized loss. Our method shows promise in improving segmentation performance and achieves new state-of-the-art on the open vocabulary segmentation tasks.

Acknowledgement

This research work is supported by the Agency for Science, Technology and Research (A*STAR) under its MTC Programmatic Funds (Grant No. M23L7b0021).

References

Alayrac, J., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., Simonyan, K., 2022.

Flamingo: a visual language model for few-shot learning. arXiv preprint arxiv:2204.14198 .

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D., 2015. Vqa: Visual question answering, in: Proceedings of the IEEE international conference on computer vision, pp. 2425–2433.

Bucher, M., Vu, T.H., Cord, M., Pérez, P., 2019. Zero-shot semantic segmentation. NeuralIPS 32, 468–479.

Canny, J., 1986. A computational approach to edge detection. IEEE TPAMI PAMI-8, 679–698. doi:10.1109/TPAMI.1986.4767851.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020a. A simple framework for contrastive learning of visual representations, in: ICML, PMLR. pp. 1597–1607.

Chen, X., Fan, H., Girshick, R., He, K., 2020b. Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 .

Chen, X., He, K., 2021. Exploring simple siamese representation learning, in: CVPR, pp. 15750–15758.

Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J., 2020c. Uniter: Learning universal image-text representations, in: ECCV, Springer.

Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R., 2022. Masked-attention mask transformer for universal image segmentation.

Cho, S., Shin, H., Hong, S., An, S., Lee, S., Arnab, A., Seo, P.H., Kim, S., 2023. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. arXiv preprint arXiv:2303.11797 .

DeTone, D., Malisiewicz, T., Rabinovich, A., 2018. Superpoint: Self-supervised interest point detection and description, in: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp. 224–236.

Ding, J., Xue, N., Xia, G.S., Dai, D., 2022a. Decoupling zero-shot semantic segmentation, in: CVPR, pp. 11583–11592.

Ding, Z., Wang, J., Tu, Z., 2022b. Open-vocabulary panoptic segmentation with maskclip. arXiv preprint arXiv:2208.08984 .

Everingham, M., Winn, J., 2011. The pascal visual object classes challenge 2012 (voc2012) development kit. Pattern Analysis, Statistical Modelling and Computational Learning, Tech. Rep 8, 5.

Ghiasi, G., Gu, X., Cui, Y., Lin, T.Y., 2022. Scaling open-vocabulary image segmentation with image-level labels.

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., 2020a. Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems 33, 21271–21284.

Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.A., Guo, Z.D., Azar, M.G., et al., 2020b. Bootstrap your own latent: A new approach to self-supervised learning. arXiv preprint arXiv:2006.07733 .

Gu, X., Lin, T.Y., Kuo, W., Cui, Y., 2021. Open-vocabulary object detection via vision and language knowledge distillation. arXiv preprint arXiv:2104.13921 .

Han, C., Wang, Q., Cui, Y., Cao, Z., Wang, W., Qi, S., Liu, D., 2023. E^ 2vpt: An effective and efficient approach for visual prompt tuning. arXiv preprint arXiv:2307.13770 .

He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum contrast for unsupervised visual representation learning, in: CVPR, pp. 9729–9738.

Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y., 2021. Clipscore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718 .

Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q.V., Sung, Y., Li, Z., Duerig, T., 2021. Scaling up visual and vision-language representation learning with noisy text supervision. arXiv preprint arXiv:2102.05918 .

Li, B., Weinberger, K.Q., Belongie, S., Koltun, V., Ranftl, R., 2022. Language-driven semantic segmentation, in: ICLR.

Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J., 2024a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems 36.

Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al., 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks, in: ECCV, Springer. pp. 121–137.

Li, X., Yuan, H., Li, W., Ding, H., Wu, S., Zhang, W., Li, Y., Chen, K., Loy, C.C., 2024b. Omg-seg: Is one model good enough for all segmentation?, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 27948–27959.

Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., Zhang, H., Zhang, P., Vajda, P., Marculescu, D., 2022. Open-vocabulary semantic segmentation with mask-adapted clip. arXiv preprint arXiv:2210.04150 .

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

Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection, in: ICCV.

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context, in: ECCV, pp. 740–755.

Lindeberg, T., 2012. Scale invariant feature transform .

Liu, D., Cui, Y., Tan, W., Chen, Y., 2021. Sg-net: Spatial granularity network for one-stage video instance segmentation, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 9816–9825.

Liu, W., Wu, Z., Zhao, Y., Fang, Y., Foo, C.S., Cheng, J., Lin, G., 2024. Harmonizing base and novel classes: A class-contrastive approach for generalized few-shot segmentation. International Journal of Computer Vision 132, 1277–1291.

Liu, W., Zhang, C., Lin, G., Liu, F., 2020. Crnet: Cross-reference networks for few-shot segmentation, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 4165–4173.

Lu, J., Batra, D., Parikh, D., Lee, S., 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. arXiv preprint arXiv:1908.02265 .

Lu, Y., Zhang, J., Sun, S., Guo, Q., Cao, Z., Fei, S., Yang, B., Chen, Y., 2023. Label-efficient video object segmentation with motion clues. IEEE Transactions on Circuits and Systems for Video Technology .

Mitrovic, J., McWilliams, B., Walker, J., Buesing, L., Blundell, C., 2020. Representation learning via invariant causal mechanisms. arXiv preprint arXiv:2010.07922 .

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

Mukhoti, J., Lin, T.Y., Poursaeed, O., Wang, R., Shah, A., Torr, P.H., Lim, S.N., 2023. 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.

Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., Lischinski, D., 2021. Styleclip: Text-driven manipulation of stylegan imagery, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2085–2094.

Qin, Z., Han, C., Wang, Q., Nie, X., Yin, Y., Xiankai, L., 2023. Unified 3d segmenter as prototypical classifiers. Advances in Neural Information Processing Systems 36, 46419–46432.

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. arXiv preprint arXiv:2103.00020 .

Roh, B., Shin, W., Kim, I., Kim, S., 2021. Spatially consistent representation learning, in: CVPR, pp. 1144–1153.

Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., Dai, J., 2019. Vl-bert: Pre-training of generic visual-linguistic representations. arXiv preprint arXiv:1908.08530 .

Wang, F., Liu, H., 2021. Understanding the behaviour of contrastive loss, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 2495–2504.

Wang, J., Sun, G., Wang, P., Liu, D., Dianat, S., Rabbani, M., Rao, R., Tao, Z., 2024. Text is mass: Modeling as stochastic embedding for text-video retrieval, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16551–16560.

Wang, W., Han, C., Zhou, T., Liu, D., 2022a. Visual recognition with deep nearest centroids. arXiv preprint arXiv:2209.07383 .

Wang, W., Liang, J., Liu, D., 2022b. Learning equivariant segmentation with instance-unique querying. Advances in Neural Information Processing Systems 35, 12826–12840.

Wang, Z., Lu, Y., Li, Q., Tao, X., Guo, Y., Gong, M., Liu, T., 2022c. Cris: Clip-driven referring image segmentation, in: CVPR, pp. 11686–11695.

Wei, C., Wang, H., Shen, W., Yuille, A., 2020. Co2: Consistent contrast for unsupervised visual representation learning. arXiv preprint arXiv:2010.02217 .

Xian, Y., Choudhury, S., He, Y., Schiele, B., Akata, Z., 2019. Semantic projection network for zero-and few-label semantic segmentation, in: CVPR, pp. 8256–8265.

Xu, M., Zhang, Z., Wei, F., Hu, H., Bai, X., 2023. Side adapter network for open-vocabulary semantic segmentation, in: CVPR, pp. 2945–2954.

Xu, M., Zhang, Z., Wei, F., Lin, Y., Cao, Y., Hu, H., Bai, X., 2022. A simple baseline for open-vocabulary semantic segmentation with pretrained vision-language model, in: ECCV, Springer. pp. 736–753.

Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y., 2022. Coca: Contrastive captioners are image-text foundation models. arXiv preprint arxiv:2205.01917 .

Yuan, L., Chen, D., Chen, Y., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., Liu, C., Liu, M., Liu, Z., Lu, Y., Shi, Y., Wang, L., Wang, J., Xiao, B., Xiao, Z., Yang, J., Zeng, M., Zhou, L., Zhang, P., 2021. Florence: A new foundation model for computer vision. arXiv preprint arxiv:2111.11432 .

Zhao, H., Puig, X., Zhou, B., Fidler, S., Torralba, A., 2017. Open vocabulary scene parsing, in: ICCV, pp. 2002–2010.

Zhao, X., Vemulapalli, R., Mansfield, P.A., Gong, B., Green, B., Shapira, L., Wu, Y., 2021. Contrastive learning for label efficient semantic segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10623–10633.

Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A., 2017. Scene parsing through ade20k dataset, in: CVPR, pp. 633–641.

Zhou, C., Loy, C.C., Dai, B., 2021. Denseclip: Extract free dense labels from clip. arXiv preprint arXiv:2112.01071 .