TagCLIP: A Local-to-Global Framework to Enhance Open-Vocabulary Multi-Label Classification of CLIP without Training

Yuqi Lin[1,3], Minghao Chen[2*], Kaipeng Zhang[3*], Hengjia Li[1], Mingming Li[1], Zheng Yang[4], Dongqin Lv[6], Binbin Lin[5], Haifeng Liu[1], Deng Cai[1,4]

[1]State Key Lab of CAD&CG, College of Computer Science, Zhejiang University
[2]Hangzhou Dianzi University
[3]Shanghai Artificial Intelligence Laboratory
[4]FABU Inc.
[5]School of Software Technology, Zhejiang University
[6]Nantong Port Group
{linyq5566, minghaochen01}@gmail.com, kp_zhang@foxmail.com

Abstract

Contrastive Language-Image Pre-training (CLIP) has demonstrated impressive capabilities in open-vocabulary classification. The class token in the image encoder is trained to capture the global features to distinguish different text descriptions supervised by contrastive loss, making it highly effective for single-label classification. However, it shows poor performance on multi-label datasets because the global feature tends to be dominated by the most prominent class and the contrastive nature of softmax operation aggravates it. In this study, we observe that the multi-label classification results heavily rely on discriminative local features but are overlooked by CLIP. As a result, we dissect the preservation of patch-wise spatial information in CLIP and proposed a local-to-global framework to obtain image tags. It comprises three steps: (1) patch-level classification to obtain coarse scores; (2) dual-masking attention refinement (DMAR) module to refine the coarse scores; (3) class-wise reidentification (CWR) module to remedy predictions from a global perspective. This framework is solely based on frozen CLIP and significantly enhances its multi-label classification performance on various benchmarks without dataset-specific training. Besides, to comprehensively assess the quality and practicality of generated tags, we extend their application to the downstream task, i.e., weakly supervised semantic segmentation (WSSS) with generated tags as image-level pseudo labels. Experiments demonstrate that this *classify-then-segment* paradigm dramatically outperforms other annotation-free segmentation methods and validates the effectiveness of generated tags. Our code is available at https://github.com/linyq2117/TagCLIP.

Introduction

Contrastive Language-Image Pre-training (CLIP) (Radford et al. 2021) has recently emerged as a powerful vision-language model. It is pre-trained on a large-scale dataset of image-text pairs and has shown impressive performance in image-text matching tasks (Zhou et al. 2022; Crowson et al. 2022; Gu et al. 2021). By transferring this matching ability to the classification task, we can recognize arbitrary text labels and achieve open-vocabulary classification. However, most existing open-vocabulary works focus on the single-label classification task while multi-label classification, which aims to recognize all the relevant categories or concepts in an image, is a more practical and challenging task. In Figure 1, we find that the performance is unsatisfactory on multi-label classification datasets. Specifically, the classification logits predicted by the class token tend to be dominated by the most prominent class, while some inconspicuous objects, e.g., with small size, are usually underrated. It stems from two main reasons: (1) CLIP is trained to align image-text pairs with contrastive loss, which aims to match an image with its corresponding text descriptions and distinguish it from others. The softmax operation introduced by this loss creates competition among different classes, which is detrimental to the multi-label setting. (2) CLIP is trained to represent an entire image through a unique

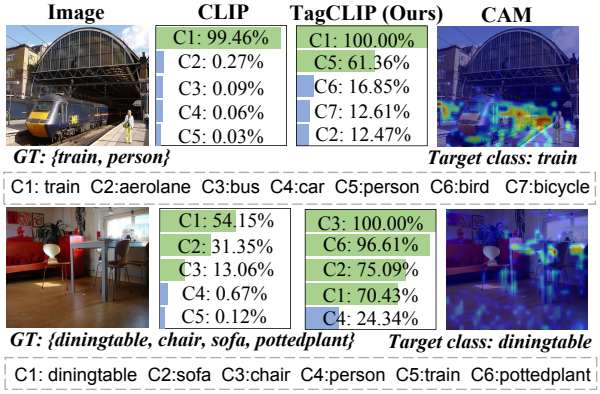

Figure 1: Visualizations of multi-label classification results and CAMs of some target classes. The middle two columns demonstrate that original CLIP (Radford et al. 2021) usually fails to recognize inconspicuous categories while our TagCLIP can identify them well. The last column presents some CAMs of specific classes and indicates that classification mainly depends on some discriminative local features. All results are based on ViT-B/16 and we leverage Grad-CAM (Selvaraju et al. 2017) to obtain CAMs for CLIP.

*Corresponding authors

global embedding using the class token, without explicitly capturing the local features of specific regions. However, in the multi-label setting, discriminative local features are more helpful. This preference for local features can be observed in the Class Activation Map (CAM) (Zhou et al. 2016) shown in Figure1, where the highly responsive regions for the target class mainly correspond to specific local cues. Therefore, it is necessary to explore the spatial information preserved in CLIP-ViT to take advantage of discriminative local cues.

In general, the final output feature map of a model is commonly utilized for localization tasks, e.g., object detection (Ren et al. 2015) or segmentation (Chen et al. 2017). However, we observe that the localization quality of CLIP-ViT is not effective for the last feature map (seeing Figure 3). We delve into the underlying factors and find that the attention operation in the last layers is irrational for dense tokens, leading to the lack of spatial information in the final output feature map. Alternatively, by forwarding the feature map of the penultimate layer without the self-attention operation at the last layer (denoted as penultimate layer for short), the spatial information is effectively preserved. This enables us to extract local features from CLIP, enhancing its capability for capturing fine-grained details.

Building on the observation above, we further propose a novel framework called TagCLIP to enhance the multi-label classification capability of the original CLIP without training. This framework follows a local-to-global paradigm and consists of three steps. First, we ignore the attention operation at the last layer of CLIP-ViT and perform patch-level classification based on the penultimate layer to obtain corresponding classification score maps for each class. Second, to refine the initial scores and mitigate potential noise, we introduce a dual-masking attention refinement strategy based on the Multi-Head Self-Attention (MHSA) inherent in ViT. Finally, we propose a class-wise reidentification module to further improve the primary predictions from the global view. This double-check approach can filter out some falsely detected classes and improve the scores of missed cases. The whole framework remarkably improves the multi-label classification performance of CLIP. It is based solely on frozen CLIP and enables open-vocabulary multi-label classification without the need for dataset-specific training.

To further validate the quality and practicality of generated tags, we integrate TagCLIP with downstream tasks, where it serves as a generalizable annotator that provides high-quality pseudo labels. It can benefit many downstream tasks, e.g., self-training (Zoph et al. 2020; Wang et al. 2022; Xie et al. 2020), and weakly supervised learning (Lin et al. 2023; Xie et al. 2022; Xu et al. 2022b). In this paper, we explore the application of TagCLIP by integrating the generated labels with weakly supervised semantic segmentation (WSSS). The combination of open-vocabulary multi-label classification and WSSS enables annotation-free segmentation. Unlike previous works (Zhou, Loy, and Dai 2022; Van Gansbeke et al. 2021) following the bottom-up paradigm, we surprisingly find this novel *classify-then-segment* paradigm leads to significant performance gains, which indicates the importance of image-level supervision to the segmentation task.

The main contributions can be summarized as follows:

- We explore the spatial information in CLIP at the patch level and find that the attention operation in the last layer breaks spatial information. On this basis, we propose a local-to-global framework TagCLIP to enhance the multi-label classification performance of the original CLIP without any extra training.
- Experiment results demonstrate the effectiveness of our TagCLIP. It unlocks the potential of original CLIP and can generate high-quality image tags. Our method achieves significant performance gains compared to original CLIP and other works across different benchmarks.
- We integrate the proposed TagCLIP with the downstream WSSS task and find this *classify-then-segment* paradigm achieves remarkable improvement over other methods.

Related Works

Contrastive Language-Image Pre-training

Contrastive Language-Image Pre-training(CLIP) (Radford et al. 2021) connects visual concepts with textual descriptions and has empowered many computer vision tasks with language ability. It consists of an image and text encoder, and is jointly trained to align the two modalities with over 400 million image-text pairs. The image-text matching ability can be transferred to the downstream zero-shot tasks. However, the pre-training task is image-level, and only *class* token is trained to capture the global feature. For multi-label classification task, the region-level feature is preferred. Some works (Raghu et al. 2021; Ghiasi et al. 2022) explore the spatial information in the patches of deep ViT layers but the results are unsatisfactory. This paper makes it possible by ignoring the last attention operation, and leverages obtained local features to benefit multi-label classification.

Open-Vocabulary Multi-Label Classification

Multi-Label Classification aims to predict a set of labels for an image. Conventionally, a multi-label classification task is transformed into a set of binary classification tasks, which are solved by optimizing a binary cross-entropy loss function. The proposed methods can be categorized into three main directions: 1) Improving loss functions (Ridnik et al. 2021; Wu et al. 2020). 2) Modeling label correlations (Chen et al. 2019b,a; Ye et al. 2020). 3) Locating regions of interest (Wang et al. 2017; You et al. 2020). To deal with unseen labels, multi-label zero-shot learning (ML-ZSL) is developed to transfer knowledge from seen classes to unseen classes. The keys to this task are the alignment of the image with its relevant label embeddings and the relation between seen and unseen label embeddings. Existing works realize it from the perspective of finding principle directions (Ben-Cohen et al. 2021) or adopting attention module (Narayan et al. 2021; Huynh and Elhamifar 2020).

Different from ML-ZSL, visual-related language data like image captions can be used as auxiliary supervision in the open vocabulary setting. The open-vocabulary multi-label recognition can classify multi-label images via arbitrary textual names or descriptions. According to the complexity, existing methods can be divided into two groups. 1) The first

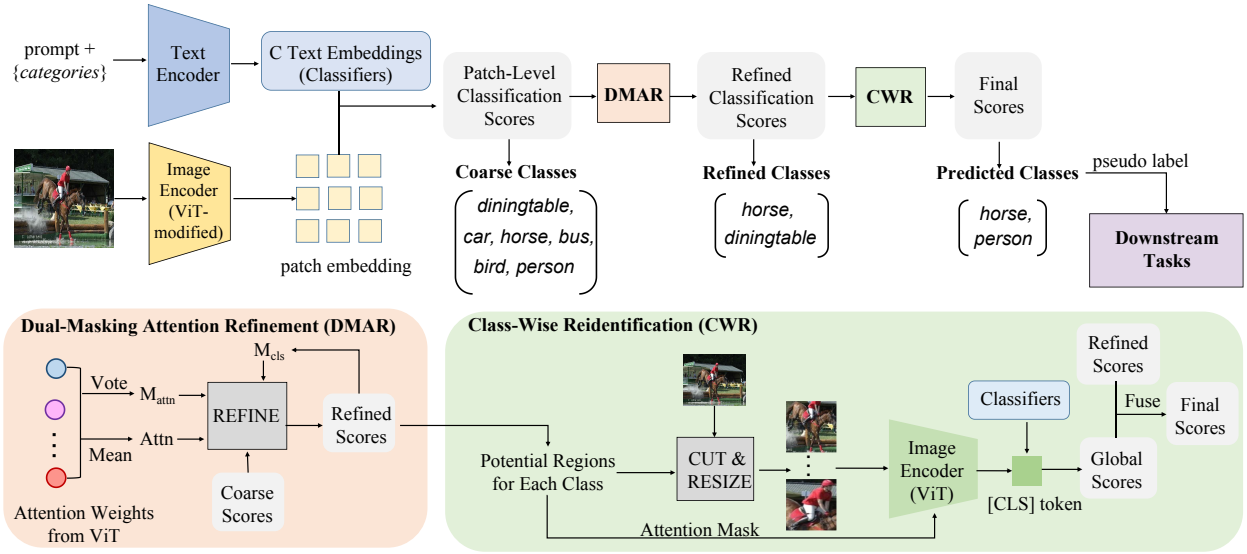

Figure 2: An overview of our proposed framework. The framework consists of three steps, i.e., patch-level classification, dual-masking attention refinement (DMAR), and class-wise reidentification (CWR). C is the total number of classes. "ViT-modified" means ignoring the last self-attention operation to maintain spatial information. We threshold the predicted probability scores with 0.5 to obtain predicted classes. The predicted image tags can be treated as pseudo labels for downstream tasks, e.g., WSSS.

group requires additional training processes on seen classes or specific curated data. These methods require fine-tuning on target datasets (He et al. 2023; Sun, Hu, and Saenko 2022) or training from scratch using massive data (Guo et al. 2023), both of which have complex training processes. 2) The second group is merely based on pre-trained models without further training or extra information (Li et al. 2023). Our work falls into the second group, which is more challenging but convenient to use. Similar to CLIP-Surgery (Li et al. 2023), we improve the classification ability of CLIP from the perspective of model explainability. The difference is that we leverage a local-to-global framework, while CLIP-Surgery only relies on global embedding.

Annoation-Free Semantic Segmentation

In the annotation-free segmentation setting, no annotation is provided during training, which is corresponding to unsupervised semantic segmentation (USS). Primary USS methods leverage self-supervised learning to learn pixel-level representation (Ji, Henriques, and Vedaldi 2019; Cho et al. 2021; Ziegler and Asano 2022; Ke et al. 2022; Hwang et al. 2019; Van Gansbeke et al. 2021) and the learned representations can then be employed to cluster image segments via K-means or linear classifiers. These bottom-up approaches are difficult to distinguish different classes with similar appearances or identify classes with varied appearances.

Another similar setting is open-vocabulary segmentation. Its target is to segment an image with arbitrary categories described by texts instead of fixed labeling vocabularies. It typically addresses the closed-set limitation via training on weak supervision signals, e.g., image-text pairs (Xu et al. 2022a; Luo et al. 2023). Differently, recent works (Zhou, Loy, and Dai 2022; Shin, Xie, and Albanie 2022b,a) are merely based on pre-trained CLIP and require no extra annotations. The performance gain of these methods is still limited for the lack of high-level semantic guidance. We denote all the above methods leveraging image-text pairs or the pre-trained model as CLIP-based methods.

Method

In this section, we introduce our CLIP-based multi-label classification framework, TagCLIP, which is depicted in Figure 2. We first review the architecture of CLIP-ViT and investigate the spatial information preserved in the patches. Then, we introduce the proposed local-to-global framework for multi-label classification without annotations and fine-tuning. Finally, we present the application of generated image tags on the downstream WSSS task.

Analysis of CLIP

CLIP (Radford et al. 2021) consists of an image encoder and a text encoder and is jointly trained to align the two modalities with large-scale image-text pairs. For the image encoder with transformer architecture, a [*cls*] token is pre-trained to capture the global feature. Given the ViT with L layers, the forward propagation of the last transformer layer is expressed as follows:

$$\hat{X}^L = X^{L-1} + \mathbf{a}^L, \tag{1}$$

$$= X^{L-1} + A^L \left(X^{L-1} W_V^L\right), \tag{2}$$

$$A^L = \sigma\left(\frac{(X^{L-1}W_Q^L)(X^{L-1}W_V^L)^T}{\sqrt{d}} + M^L\right), \tag{3}$$

$$X^L = \hat{X}^L + \mathrm{MLP}(\hat{X}^L), \tag{4}$$

where X^{L-1} represents the output tokens of the *L-1* layer, $\mathbf{a}^L$ and MLP represent the self-attention and the MLP modules in the transformer block. A^L encodes the attention weights at layer L. σ represents the softmax normalization, d is the dimension of X^{L-1}, M^L is attention mask for A^L. W_Q, W_K, W_V are linear projection weights to generate *query, key, value* in MHSA. X^L consists of the [*cls*] token and remaining tokens (denoted as dense tokens):

$$X^L = [x^L_{cls}, x^L]. \tag{5}$$

As mentioned in Introduction, the contrastive loss and global embedding in the original CLIP will harm the multi-label classification. Alternatively, region-level features are better suited to recognize multiple categories in an image. As only the [*cls*] token is used during contrastive pre-training, the localization ability of the original CLIP is weak (Zhong et al. 2022). There is a major performance degradation when applying the pre-trained CLIP model for localization tasks (e.g., only 16.2% mIoU for segmentation by leveraging the final output feature map in Table 1).

We hypothesize that the spatial information is remained in feature maps of CLIP in the previous layer but lacks in the last layer for the following reasons: (1) The query and key in the last attention layer are merely involved in the optimization of [*cls*] token to perform weighted sum operation and globalizes information during pre-training. It is a special design for [*cls*] token but is meaningless and redundant for remaining dense tokens. (2) The [*cls*] token plays a relatively minor role throughout the vision transformer and is not used for globalization until the last layer (Ghiasi et al. 2022). Therefore, it scarcely affects local features in previous layers. To verify it, we use ViT-B/16 with 12 layers and treat the encoded text features as classifiers to classify each dense token outputted by the last two layers. To make the feature embedded into the same feature space, we let the dense token outputted by the penultimate layer pass the rest layer without self-attention:

$$\hat{x}_{dense} = x^{L-1} + \mathbf{c}^L, \tag{6}$$

$$= x^{L-1} + x^{L-1}W_V^L, \tag{7}$$

$$x_{dense} = \hat{x}_{dense} + \mathrm{MLP}(\hat{x}_{dense}). \tag{8}$$

We provide qualitative and quantitative results in Figure 3 and Table 1. The results demonstrate spatial information is preserved in the penultimate layer but lacks in the last layer. Therefore, it is feasible to omit the last self-attention operation and perform classification based on the projected output of the penultimate layer to discover the discriminative features for target classes.

CLIP-Based Multi-Label Classification

This section introduces our proposed local-to-global framework for multi-label classification, including patch-level classification to obtain coarse scores, dual-masking attention refinement (DMAR) to refine coarse scores, and the class-wise reidentification (CWR) module to double-check the potential predictions.

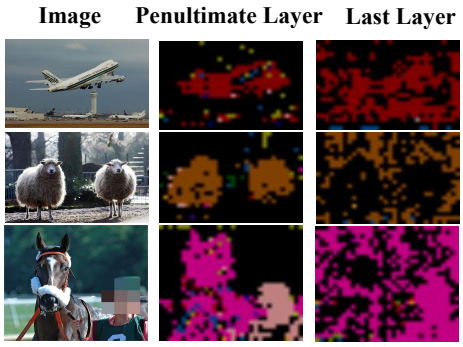

Figure 3: Qualitative results of the patch-level classification upon x_{dense} and x^L outputted by the last two layers of CLIP-ViT respectively. The last self-attention operation breaks spatial information in ViT.

Last Self-Attention	mAP	mIoU
✓	82.7	16.2
✗	85.4	41.6

Table 1: Quantitative results for the effect of last self-attention operation in terms of classification (mAP) and segmentation (mIoU) on PASCAL VOC 2012 validation set.

Coarse Classification To perform patch-level classification, the output feature map based on the penultimate layer $x_{dense} \in R^{N \times D}$ is leveraged. The output of the text encoder is denoted as $T \in R^{D \times C}$, which acts as the classifier based on the text inputs. N, D, C represent token length, token dimension and class number, respectively. The classification score for each patch in x_{dense} is calculated as:

$$s_i = Linear(x_{dense,i}) * T, \tag{9}$$

where i represents the spatial index of each patch. $Linear$ is the last layer of CLIP to map the encoded image features and text features into the unified space in CLIP. s_i reflects the similarity of image token and C text descriptions, and the similarity scores will be forwarded to the softmax function to normalize these scores over all classes (Note that the softmax operation is optional, but we find it significant for CLIP and validate it in the experiment section). The probability classification score of class c for each dense token i can be obtained as follows:

$$P_{coarse}(i, c) = \frac{\exp(s_i^c)}{\sum_{k=1}^{C} \exp(s_i^k)}. \tag{10}$$

Dual-Masking Attention Refinement (DMAR) The initial patch-level classification scores obtained from Equation 10 often suffer from noise, hindering them from serving as a reliable criterion for class identification (e.g., leading to false positives for classification in Figure 5). Prior approaches typically utilize pairwise affinity to refine dense classification maps, but require training extra layers (Ahn and Kwak 2018; Ahn, Cho, and Kwak 2019). In contrast, the

vision transformer's inherent self-attention mechanism captures the pairwise affinity between patches, allowing us to refine the patch-wise classification scores without incurring additional computational costs. A common way is directly using attention weights from the last few layers (Xu et al. 2022b) or all layers (Gao et al. 2021) of ViT and performing the refinement as follows:

$$P_{refined} = \frac{1}{|\psi|} \sum_{l \in \psi} A_l * P_{coarse}, \tag{11}$$

where $P_{coarse} \in R^{N \times C}$ denotes the coarse score map, $A_l \in R^{N \times N}$ represents the attention weight in the *l-th* layer of ViT, ψ represents the index set of used attention layer and $|\psi|$ is its number of elements.

However, the affinity captured in the original ViT's MHSA is inaccurate (Ru et al. 2022), potentially misleading the refinement process. To address it, we propose a dual-masking strategy, the key idea of which is to neglect unconfident elements in both attention weights $A \in R^{N \times N \times L}$ and coarse score maps $P_{coarse} \in R^{N \times C}$. For attention weights, we generate an attention mask $M_{attn} \in R^{N \times N}$ to select confident elements by leveraging a voting-style approach across all L attention layers. Each confident position should have prominent attention value (exceeding layer-wise mean value) in at least K layers, which can be represented as:

$$M_{attn}(i,j) = 1, \, if \sum_{l=1}^{L} \mathbb{I}_{(A(i,j,l) > \bar{A}_l)}(A) > K, \tag{12}$$

where $\mathbb{I}$ is the indicator function, $\bar{A}_l$ is the mean value of $l\text{-}th$ layer. The refinement procedure is then illustrated as:

$$\hat{P}_{refined} = \frac{1}{|\psi|} \sum_{l \in \psi} M_{attn} \odot A_l * P_{coarse}, \tag{13}$$

where $\odot$ denotes the Hadamard product. For coarse score maps, we calculate the average score for each class based on $\hat{P}_{refined}$ and produce a expanded class-wise mask $M_{cls} \in R^{N \times N \times C}$ by ignoring unconfident positions (below the average score). The final refined scores for each class c can be obtained as follows:

$$P_{refined}(c) = \frac{1}{|\psi|} \sum_{l \in \psi} M_{attn} \odot A_l \odot M_{cls}(c) * P_{coarse}(c). \tag{14}$$

Class-Wise Reidentification (CWR) Although patch-level classification can discover target classes by discriminative local features, it may result in misclassification for the lack of a comprehensive view. Therefore, we propose a class-wise reidentification module to further remedy the primary predicted scores for each class from a global view. Specifically, given refined classification scores $P_{refined} \in R^{N \times C}$, we can obtain the confidence of each class P_{local} by corresponding most outstanding patches:

$$P_{local}(c) = \max_i (P_{coarse}(i,c)), \tag{15}$$

For each class, we pick out the highly responsive patches from $P_{refined}$ and form the class-related region (class-wise mask). We crop the image by the bounding box of the region and resize it to a specific size, e.g., 224×224. The class-wise mask serves as the attention mask in ViT to exclude patches that do not belong to the class. We input the class-wise image into original CLIP and use [*cls*] token for classification. The obtained global results P_{global} are merged with local scores P_{local} to take advantage of both local and global views.

$$P_{final} = \lambda P_{local} + (1 - \lambda) P_{global}, \tag{16}$$

where λ is a coefficient to balance the local and global effect and is simply set to 0.5 in our experiments. Through this fusion process, we can effectively incorporate the valuable insights provided by both local and global views, thereby enhancing the overall classification performance.

Appplication on the Downstream Task

Multi-label classification is a practical task with wide-ranging applications in downstream tasks that rely on image-level labels. In this paper, we explore the use of TagCLIP in conjunction with existing Weakly Supervised Semantic Segmentation (WSSS) methods to tackle annotation-free semantic segmentation. Given image-level labels, most WSSS works (Wang et al. 2020; Xie et al. 2022) leverage Class Activation Mapping (CAM) to find the target class's related regions in the image and generate segmentation masks based on it. The use of category information provides valuable high-level guidance, enabling WSSS to perform remarkably well, even approaching the performance of fully-supervised settings. We select CLIP-ES (Lin et al. 2023) for its outstanding accuracy and efficiency. It is also a training-free framework based on frozen CLIP and more details can be found in (Lin et al. 2023). By leveraging this efficient WSSS method, the whole *classify-then-segment* paradigm requires no dataset-specific training and can realize annotation-free segmentation. We denote this framework as CLS-SEG.

Experiment

Experimental Setup

Dataset and Evaluation Metrics. To verify the performance of multi-label classification, for fair comparisons, we evaluate our method on PASCAL VOC 2007 (Everingham et al. 2010) and MS COCO 2014 (Lin et al. 2014). following (Guo et al. 2023). The PASCAL VOC 2007 contains 20 categories and we evaluate on the test set with 4952 images. MS COCO 2014 includes 80 categories, and we take the 40137 images as validation set following the official split. For downstream semantic segmentation, we conduct experiments on three commonly used datasets, including PASCAL VOC 2012 (Everingham et al. 2010), MS COCO 2017 (Lin et al. 2014) and COCO-Stuff (Caesar, Uijlings, and Ferrari 2018). For Pascal VOC 2012, there are 20 foreground classes, and the remaining pixels are background. The validation set with 1449 images is used for validation. COCO 2017 has 5000 validation images with 80 categories and a background class. COCO-stuff has 4172 validation images of 171 low-level categories. We employ 27 mid-level categories setting following (Shin, Xie, and Albanie 2022b).

Method	Extra Training Data	VOC	COCO
Supervised specialist:			
SARB	10% Data	83.5	75.5
DualCoOp	10% Data	90.3	78.7
TAI-DPT	10% Data	93.3	81.5
Open-vocabulary generalist:			
TAI-DPT	COCO captions	88.3	65.1
CLIP†	None	79.5	54.2
CLIP	None	85.8	63.3
DPT†	None	83.4	59.6
DPT	None	86.2	64.3
CLIPSurgery	None	85.4	61.2
TagCLIP(Ours)	None	**92.8**	**68.8**

Table 2: Experimental results of multi-label classification. † represents not using softmax on classification scores.

Method	VOC	COCO	COCO-Stuff
Vanilla USS methods			
IIC	9.8	-	6.7
MaskContrast	35.0	3.73	-
TransFGU	37.2	12.7	17.5
MaskDistill	45.8	-	-
PiCIE	-	-	13.8
PiCIE+H	-	-	14.4
CLIP-based methods			
MaskCLIP‡	42.1	20.2	23.9
CLIPSurgery‡	41.5	25.2	29.7
GroupViT	52.3	24.3	-
SegCLIP	52.6	26.5	-
ReCo	34.2	17.1	26.3
NamedMask	59.2	27.7	-
CLS-SEG (Ours)	**64.8**	**34.0**	**30.1**
CLS-SEG*(Ours)	**68.7**	**35.3**	**31.0**

Table 3: Results of annotation-free semantic segmentation. The vanilla USS results are based on K-means clustering. ‡ represents we re-implement it with the same experimental setting as ours. * means using denseCRF to postprocess.

Note that our classification framework TagCLIP and segmentation framework CLS-SEG are both training-free and can directly evaluate on the validation set. We employ mean average precision (mAP) as the evaluation metric for multi-label classification and the mean Intersection over Union (mIoU) for semantic segmentation.

Implementation Details. Our experiments are based on ViT-B/16 pre-trained by CLIP. For multi-label classification, images remain at their original resolution. In every operation where a confidence threshold is required, threshold 0.5 is substituted if not otherwise specified, such as thresholds for selecting highly responsive patches in CWR. We adopt the 80 prompts used in CLIP (Radford et al. 2021) and background set in (Lin et al. 2023). To determine the potential classes in an image according to classification logits, we first perform min-max normalization to scale the logits to $[0, 1]$ and then set 0.5 to determine positive categories.

Experimental Results

Multi-label classification. To demonstrate the effectiveness of our proposed TagCLIP, we compare it with other CLIP-based approaches. Some supervised specialist methods leverage partial data on downstream datasets to train customized models, including SARB (Pu et al. 2022), DualCoOp (Sun, Hu, and Saenko 2022), TAI-DPT (Guo et al. 2023). The use of downstream data limits their generalization. Another training-based manner has no access to downstream data but trains on curated caption data, which enables arbitrary category recognition. The others are merely based on frozen CLIP and thus inherit its outstanding generalization capability, including CLIP (Radford et al. 2021), DPT (Guo et al. 2023), CLIPSurgery (Li et al. 2023).

In Table 2, † represents directly treating logits before softmax as classification scores (Guo et al. 2023) because these logits can reflect the similarity between image and text features. We find that there is a major performance degradation without softmax activation, which may stem from the use of contrastive loss during pre-training of CLIP. Results in Table 2 demonstrate that our proposed framework performs surprisingly well. It enhances the multi-label classification performance of original CLIP by a large margin, i.e., 7.0% and 5.5% on VOC and COCO, respectively. Our method surpasses all works that require no extra training data on both VOC and COCO. It also compares favorably with the works requiring extra data and training. More experiment results are available in Appendix.

Segmentation performance. We provide our annotation-free segmentation result with tags generated by TagCLIP as pseudo labels and compare them with both **vanilla USS methods** (including IIC (Ji, Henriques, and Vedaldi 2019), MaskContrast (Van Gansbeke et al. 2021), TransFGU (Yin et al. 2022), MaskDistill (Van Gansbeke, Vandenhende, and Van Gool 2022), PiCIE(+H) (Cho et al. 2021)) and recent **CLIP-based works** (including MaskCLIP (Zhou, Loy, and Dai 2022), CLIPSurgery (Li et al. 2023), GroupViT (Xu et al. 2022a), SegCLIP (Luo et al. 2023), ReCo (Shin, Xie, and Albanie 2022b), NamedMask (Shin, Xie, and Albanie 2022a)) in Table 3.

We observe that CLS-SEG outperforms vanilla USS and other CLIP-based methods dramatically on all three datasets, demonstrating the high quality of generated tags and the effectiveness of this *classify-then-segment* paradigm. From Figure 4, we find that high-level conceptual guidance provided by category information in an image is essential to obtain high-quality segmentation masks because: 1) it prevents false predictions caused by confusing textures among semantically similar classes, e.g., the skin of the cow and sheep; 2) it can comprehensively identify some categories with large intra-class variance, e.g., different parts of a person can be identified as a whole with superior semantic concepts. Results indicate that classification helps segmentation and may provide inspiration for future research.

Ablation Study

Effect of DMAR and CWR. In Table 4, we evaluate the effect of DMAR and CWR in terms of classification and

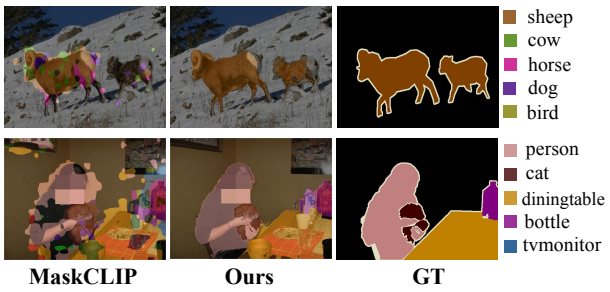

Figure 4: Visualizations of segmentation results for MaskCLIP (Zhou, Loy, and Dai 2022) and ours. MaskCLIP has more false positives for the lack of category information.

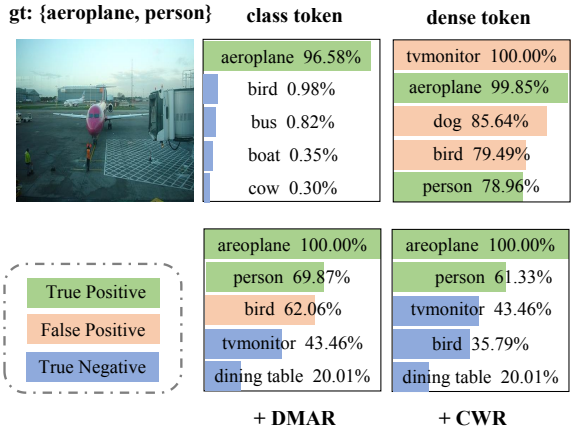

Figure 5: The classification results of different strategies. We use 0.5 as the threshold by default.

segmentation. The DMAR can refine the coarse scores remarkably and CWR can further boost the performance. We provide a qualitative case in Figure 5. After DMAR, most irrelevant categories can be suppressed. The CWR can coordinate with DMAR to double-check the refined scores from a global view. Therefore, the scores of false positives and false negatives can be suppressed and improved, respectively.

Effect of attention layers used in DMAR module. To determine the appropriate attention layers in CLIP-ViT for classification score refinement, we first compare single-layer attention weight with multi-layer in terms of classification (*precision, recall and f1-score*) and segmentation (*mIoU*) performance. From Figure 6, we can draw the following conclusions: 1) Fusing multi-layer attention weights usually performs better and more robustly than single-layer. 2) The performance of the first few attention layers is unsatisfactory, which mainly stems from the weak attention and features these layers learned. 3) The last attention layer is inaccurate among the last few layers, which corresponds to our analysis above. We also present the performance of our proposed dual-masking strategy, which effectively mitigates the impact of noise and improves original attention refinement in most cases. This strategy demonstrates significant precision gains with only a slight recall drop, leading to better classification and segmentation performance in general.

Coarse Score	DMAR	CWR	mAP	mIoU
✓			85.4	30.9
✓		✓	88.0	55.2
✓	✓		93.9	63.7
✓	✓	✓	94.1	64.8

Table 4: Results for the effectiveness of DMAR and CWR module in terms of classification and semantic segmentation. The results are evaluated on the VOC 2012 val set.

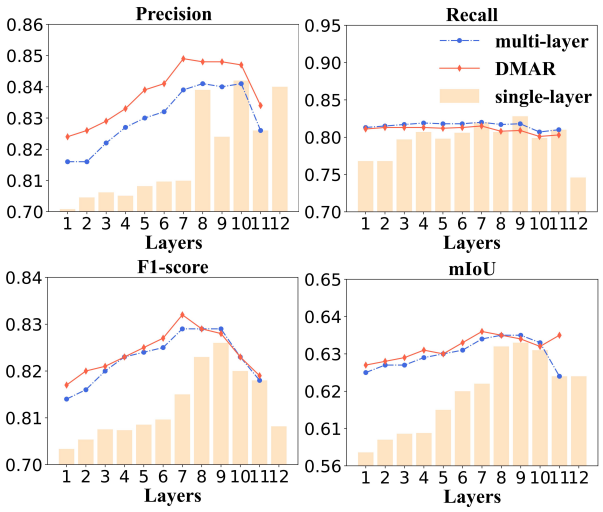

Figure 6: Comparison of single-layer and multi-layer attention refinement in terms of classification and segmentation tasks. For the single-layer setting, each tick *i* on the x-axis represents merely adopting attention weight in *i-th* layer. For the multi-layer setting, the *i-th* x-tick means fusing *i-th to 11-th* layers attention weights to refine coarse classification scores. We rule out the last attention layer during fusing.

Based on these observations, we fuse the last four attention weights except the last one in our experiments.

Conclusion

This paper proposes TagCLIP, a simple and effective framework designed to enhance the multi-label classification capability of the original CLIP. It follows a local-to-global paradigm and consists of three key steps: patch-level classification, dual-masking attention refinement (DMAR), and class-wise reidentification (CWR). Benefiting from these steps, TagCLIP unlocks the potential of CLIP and can serve as a generalizable annotator that provides high-quality image tags without dataset-specific training. Additionally, we validate the practicality of treating generated tags as pseudo labels for the downstream weakly supervised semantic segmentation (WSSS) task and find this *classify-then-segment* paradigm surpasses previous bottom-up style annotation-free segmentation methods remarkably. This demonstrates the effectiveness and versatility of TagCLIP and highlights its potential in various downstream applications.

Acknowledgements

This work was supported in part by The National Nature Science Foundation of China (Grant Nos: 62273303, 62273301, 62273302, 62036009, 61936006), in part by Ningbo Key R&D Program (No.2023Z231, 2023Z229), in part by Yongjiang Talent Introduction Programme (Grant No: 2022A-240-G), in part by the Key R&D Program of Zhejiang Province, China (2023C01135), in part by the National Key R&D Program of China (NO.2022ZD0160100).

References

Ahn, J.; Cho, S.; and Kwak, S. 2019. Weakly supervised learning of instance segmentation with inter-pixel relations. In *CVPR*.

Ahn, J.; and Kwak, S. 2018. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *CVPR*.

Ben-Cohen, A.; Zamir, N.; Ben-Baruch, E.; Friedman, I.; and Zelnik-Manor, L. 2021. Semantic diversity learning for zero-shot multi-label classification. In *ICCV*, 640–650.

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

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, 40(4): 834–848.

Chen, T.; Xu, M.; Hui, X.; Wu, H.; and Lin, L. 2019a. Learning semantic-specific graph representation for multi-label image recognition. In *ICCV*, 522–531.

Chen, Z.-M.; Wei, X.-S.; Wang, P.; and Guo, Y. 2019b. Multi-label image recognition with graph convolutional networks. In *CVPR*, 5177–5186.

Cho, J. H.; Mall, U.; Bala, K.; and Hariharan, B. 2021. Picie: Unsupervised semantic segmentation using invariance and equivariance in clustering. In *CVPR*, 16794–16804.

Crowson, K.; Biderman, S.; Kornis, D.; Stander, D.; Hallahan, E.; Castricato, L.; and Raff, E. 2022. Vqgan-clip: Open domain image generation and editing with natural language guidance. In *ECCV*, 88–105. Springer.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *IJCV*, 88(2): 303–338.

Gao, W.; Wan, F.; Pan, X.; Peng, Z.; Tian, Q.; Han, Z.; Zhou, B.; and Ye, Q. 2021. Ts-cam: Token semantic coupled attention map for weakly supervised object localization. In *ICCV*, 2886–2895.

Ghiasi, A.; Kazemi, H.; Borgnia, E.; Reich, S.; Shu, M.; Goldblum, M.; Wilson, A. G.; and Goldstein, T. 2022. What do vision transformers learn? a visual exploration. *arXiv preprint arXiv:2212.06727*.

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

Guo, Z.; Dong, B.; Ji, Z.; Bai, J.; Guo, Y.; and Zuo, W. 2023. Texts as images in prompt tuning for multi-label image recognition. In *CVPR*, 2808–2817.

He, S.; Guo, T.; Dai, T.; Qiao, R.; Shu, X.; Ren, B.; and Xia, S.-T. 2023. Open-vocabulary multi-label classification via multi-modal knowledge transfer. In *AAAI*, volume 37, 808–816.

Huynh, D.; and Elhamifar, E. 2020. A shared multi-attention framework for multi-label zero-shot learning. In *CVPR*, 8776–8786.

Hwang, J.-J.; Yu, S. X.; Shi, J.; Collins, M. D.; Yang, T.-J.; Zhang, X.; and Chen, L.-C. 2019. Segsort: Segmentation by discriminative sorting of segments. In *ICCV*, 7334–7344.

Ji, X.; Henriques, J. F.; and Vedaldi, A. 2019. Invariant information clustering for unsupervised image classification and segmentation. In *ICCV*, 9865–9874.

Ke, T.-W.; Hwang, J.-J.; Guo, Y.; Wang, X.; and Yu, S. X. 2022. Unsupervised Hierarchical Semantic Segmentation with Multiview Cosegmentation and Clustering Transformers. In *CVPR*, 2571–2581.

Li, Y.; Wang, H.; Duan, Y.; and Li, X. 2023. Clip surgery for better explainability with enhancement in open-vocabulary tasks. *arXiv preprint arXiv:2304.05653*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*.

Lin, Y.; Chen, M.; Wang, W.; Wu, B.; Li, K.; Lin, B.; Liu, H.; and He, X. 2023. Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. In *CVPR*, 15305–15314.

Luo, H.; Bao, J.; Wu, Y.; He, X.; and Li, T. 2023. Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. In *ICML*, 23033–23044. PMLR.

Narayan, S.; Gupta, A.; Khan, S.; Khan, F. S.; Shao, L.; and Shah, M. 2021. Discriminative region-based multi-label zero-shot learning. In *ICCV*, 8731–8740.

Pu, T.; Chen, T.; Wu, H.; and Lin, L. 2022. Semantic-aware representation blending for multi-label image recognition with partial labels. In *AAAI*, volume 36, 2091–2098.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*.

Raghu, M.; Unterthiner, T.; Kornblith, S.; Zhang, C.; and Dosovitskiy, A. 2021. Do vision transformers see like convolutional neural networks? *NeurIPS*, 34: 12116–12128.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, 28.

Ridnik, T.; Ben-Baruch, E.; Zamir, N.; Noy, A.; Friedman, I.; Protter, M.; and Zelnik-Manor, L. 2021. Asymmetric loss for multi-label classification. In *ICCV*, 82–91.

Ru, L.; Zhan, Y.; Yu, B.; and Du, B. 2022. Learning affinity from attention: End-to-end weakly-supervised semantic segmentation with transformers. In *CVPR*, 16846–16855.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*.

Shin, G.; Xie, W.; and Albanie, S. 2022a. Namedmask: Distilling segmenters from complementary foundation models. *arXiv preprint arXiv:2209.11228*.

Shin, G.; Xie, W.; and Albanie, S. 2022b. ReCo: Retrieve and Co-segment for Zero-shot Transfer. In *NeurIPS*.

Sun, X.; Hu, P.; and Saenko, K. 2022. Dualcoop: Fast adaptation to multi-label recognition with limited annotations. *NeurIPS*, 35: 30569–30582.

Van Gansbeke, W.; Vandenhende, S.; Georgoulis, S.; and Van Gool, L. 2021. Unsupervised semantic segmentation by contrasting object mask proposals. In *ICCV*, 10052–10062.

Van Gansbeke, W.; Vandenhende, S.; and Van Gool, L. 2022. Discovering object masks with transformers for unsupervised semantic segmentation. *arXiv preprint arXiv:2206.06363*.

Wang, X.; Wu, Z.; Lian, L.; and Yu, S. X. 2022. Debiased learning from naturally imbalanced pseudo-labels. In *CVPR*, 14647–14657.

Wang, Y.; Zhang, J.; Kan, M.; Shan, S.; and Chen, X. 2020. Self-Supervised Equivariant Attention Mechanism for Weakly Supervised Semantic Segmentation. In *CVPR*.

Wang, Z.; Chen, T.; Li, G.; Xu, R.; and Lin, L. 2017. Multi-label image recognition by recurrently discovering attentional regions. In *ICCV*, 464–472.

Wu, T.; Huang, Q.; Liu, Z.; Wang, Y.; and Lin, D. 2020. Distribution-balanced loss for multi-label classification in long-tailed datasets. In *ECCV*, 162–178. Springer.

Xie, J.; Hou, X.; Ye, K.; and Shen, L. 2022. CLIMS: Cross Language Image Matching for Weakly Supervised Semantic Segmentation. In *CVPR*.

Xie, Q.; Luong, M.-T.; Hovy, E.; and Le, Q. V. 2020. Self-training with noisy student improves imagenet classification. In *CVPR*, 10687–10698.

Xu, J.; De Mello, S.; Liu, S.; Byeon, W.; Breuel, T.; Kautz, J.; and Wang, X. 2022a. GroupViT: Semantic Segmentation Emerges from Text Supervision. In *CVPR*, 18134–18144.

Xu, L.; Ouyang, W.; Bennamoun, M.; Boussaid, F.; and Xu, D. 2022b. Multi-class Token Transformer for Weakly Supervised Semantic Segmentation. In *CVPR*.

Ye, J.; He, J.; Peng, X.; Wu, W.; and Qiao, Y. 2020. Attention-driven dynamic graph convolutional network for multi-label image recognition. In *ECCV*, 649–665. Springer.

Yin, Z.; Wang, P.; Wang, F.; Xu, X.; Zhang, H.; Li, H.; and Jin, R. 2022. TransFGU: a top-down approach to fine-grained unsupervised semantic segmentation. In *ECCV*, 73–89. Springer.

You, R.; Guo, Z.; Cui, L.; Long, X.; Bao, Y.; and Wen, S. 2020. Cross-modality attention with semantic graph embedding for multi-label classification. In *AAAI*, volume 34, 12709–12716.

Zhong, Y.; Yang, J.; Zhang, P.; Li, C.; Codella, N.; Li, L. H.; Zhou, L.; Dai, X.; Yuan, L.; Li, Y.; et al. 2022. Regionclip: Region-based language-image pretraining. In *CVPR*, 16793–16803.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *CVPR*.

Zhou, C.; Loy, C. C.; and Dai, B. 2022. Extract free dense labels from clip. In *ECCV*.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.

Ziegler, A.; and Asano, Y. M. 2022. Self-Supervised Learning of Object Parts for Semantic Segmentation. In *CVPR*, 14502–14511.

Zoph, B.; Ghiasi, G.; Lin, T.-Y.; Cui, Y.; Liu, H.; Cubuk, E. D.; and Le, Q. 2020. Rethinking pre-training and self-training. *NeurIPS*, 33: 3833–3845.