SCLIP: Rethinking Self-Attention for Dense Vision-Language Inference

Feng Wang, Jieru Mei, and Alan Yuille

Johns Hopkins University

Abstract. Recent advances in contrastive language-image pretraining (CLIP) have demonstrated strong capabilities in zero-shot classification by aligning visual and textual features at an image level. However, in dense prediction tasks, CLIP often struggles to localize visual features within an image and fails to attain favorable pixel-level segmentation results. In this work, we investigate in CLIP's spatial reasoning mechanism and identify that its failure of dense prediction is caused by a *location misalignment* issue in the self-attention process. Based on this observation, we propose a training-free adaptation approach for CLIP's semantic segmentation, which only introduces a very simple modification to CLIP but can effectively address the issue of *location misalignment*. Specifically, we reform the self-attention mechanism with leveraging query-to-query and key-to-key similarity to determine attention scores. Remarkably, this minimal modification to CLIP significantly enhances its capability in dense prediction, improving the original CLIP's 14.1% average zero-shot mIoU over eight semantic segmentation benchmarks to 38.2%, and outperforming the existing SoTA's 33.9% by a large margin. Code is available at `https://github.com/wangf3014/SCLIP`.

Keywords: CLIP · Self-attention · Semantic segmentation

1 Introduction

In the era of large foundation models, intensive pretraining followed by minimal adaptations to various downstream tasks is becoming a new paradigm for transfer learning. Nonetheless, in contrast to the significant success of foundation models in natural language processing [4, 14, 45], most visual models have yet to exhibit a comparable level of zero-shot transfer learning capabilities in various downstream tasks [3, 34]. By introducing language supervision and learning on web-scale datasets, Contrastive Language-Image pretraining (CLIP) models [28, 44] are able to generalize visual representations into open-vocabulary inference and demonstrate remarkable zero-shot classification results, yet this capability remains very limited when it comes to more complex tasks such as semantic segmentation.

Specifically, CLIP performs zero-shot classification by matching image-level representations with a range of target text embeddings, by which it achieves over 70% test accuracy on ImageNet [13] when paired with proper prompting

Fig. 1: Open-vocabulary semantic segmentation examples. We evaluate on two images from COCO [5] (the 3rd and the 5th examples) and three high-resolution images in the wild, where our SCLIP consistently generates high quality segmentation masks yet the original CLIP fails to correctly localize objects. We display the corresponding text query of each segmentation mask, where *"g. retriever"* and *"b. collie"* in the first example denote golden retriever and border collie, respectively.

strategies [44]. However, directly transferring this inference protocol to semantic segmentation fails to achieve favorable results. For example, when equipped with a ViT-Base/16 [17] encoder and fed with a 224×224 resolution input image, CLIP can obtain a 14×14 dense feature map; and by simply associating such patch-level representations with text embeddings, CLIP yields a mere 3.1% mIoU on ADE20k [70] and 5.7% mIoU on COCO-Stuff [5]. Considering the supervised counterparts that often produce around 40% mIoU on this two benchmarks, this result is not really comparable. As a result, CLIP still relies on careful finetuning and in-domain adaptations for downstream dense prediction tasks [39, 64, 73].

In this work, we investigate in CLIP's potential for dense prediction and find out whether the weak supervisions of CLIP can benefit various vision tasks with minimal downstream adaptations. We start with a qualitative analysis. As shown in Figure 1, we conduct simple open-vocabulary semantic segmentation experiments on five sample images from COCO [5] or in the wild, where the vanilla CLIP model often presents incorrect dense predictions and noisy segmentation masks. However, despite its poor semantic segmentation performance, we find that CLIP is actually able to roughly recognize what objects appear in the image yet wrongly localizes them. For instance, in the second example, we set 10 target categories including *flamingo*, *water*, *land*, with distractors such as *sky*, *building*, and *person*, but although CLIP accurately obtains the correct categories such

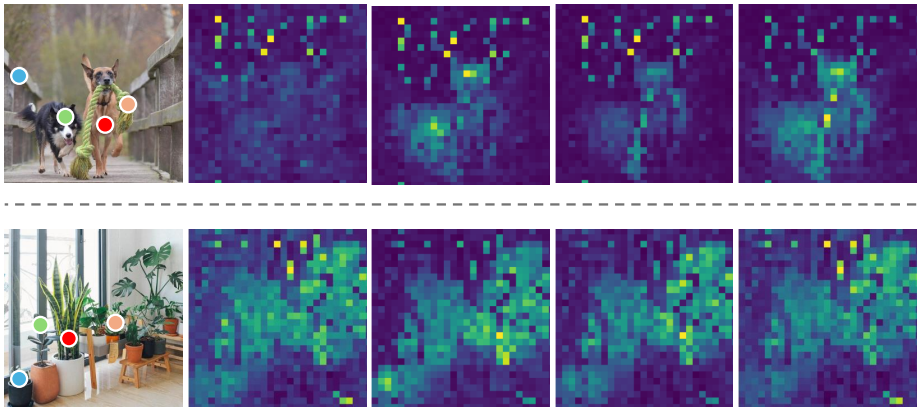

Fig. 2: Final layer attention maps of vanilla CLIP with a ViT-Base/16 image encoder. We display the attention maps of four points (marked in different colors) for each example. It shows that each local visual token attends to a wide range of positions and the attention maps often share similar patterns, indicating that CLIP learns spatial-invariant visual features.

as *water* and *flamingo*, it predicts the opposite localizations (*i.e.*, predicts *water* for flamingos and *flamingo* for water and land).

This qualitative study suggests that the poor segmentation performance of CLIP is caused by a spatial misalignment of the patch representations, instead of a failure in extracting dense visual features. This observation makes us suspect that the problem lies in CLIP's self-attention modules because they are responsible for arranging spatial information. In Figure 2, we illustrate several examples of CLIP self-attention patterns, where each map represents the attention scores for a specific point in the image (marked in different colors). As is shown, CLIP attention maps can reflect the shape of major objects, but appears to be very similar across many different source points in the image. This suggests that CLIP learns ***spatial-invariant*** visual features, implying that the local features tend to be invariant to their spatial positions in the image, and the model focuses on a holistic visual representation.

However, in dense prediction tasks like semantic segmentation, we actually desire ***spatial-covariant*** features, which implies the local representations should change accordingly to their spatial positions in an image. To this end, we rethink the purpose of self-attention and introduce **C**orrelative **S**elf-**A**ttention (CSA), a novel self-attention mechanism that facilitates covariant visual features. Specifically, in contrast to the original self-attention that employs two projection matrices (*i.e.*, the query and key) to determine attention scores, our CSA module only projects the input once to find pairwise correlations of visual tokens, which encourages each local token to attend to itself and to the positions sharing similar information with it.

Surprisingly, we find that after making this change, our CSA mechanism is very effective to adapt CLIP into dense prediction tasks. In detail, we develop our new approach **SCLIP** (**S**egmentation-adapted **CLIP** model) by employing a CSA module in place of the original self-attention block in CLIP vision encoder[1]. It is noteworthy that the CSA module is not sensitive to its projection weights so we can simply reuse the pretrained parameters of original self-attention in CLIP, which makes SCLIP a tuning-free approach for semantic segmentation using a stand-alone CLIP model.

Empirical study on our SCLIP model showcases its notable effectiveness, with yielding both impressive qualitative and quantitative outcomes: we obtain an average mIoU of 38.2% over eight semantic segmentation benchmarks such as PASCAL Context [40] and COCO-Stuff [5], substantially outperforming the existing state-of-the-art methods such as MaskCLIP [71] (30.3%), GroupViT [61] (30.7%), and TCL [8] (33.9%) that support zero-shot and open-vocabulary semantic segmentation. In Figure 1, we also show the qualitative results obtained by SCLIP for images in the COCO [5] dataset and in the wild, where our model yields very clear and accurate segmentation masks, especially for the high-resolution inputs (*e.g.*, the case of two dogs sitting on the boat). The primary contributions of this work can be summarized as follows:

- First, We identify the reasons of CLIP's failure in semantic segmentation, and address them by introducing a novel Correlative Self-Attention (CSA) mechanism, while extensive experiments demonstrate significant results.
- Next, our SCLIP approach outperforms the existing methods [8, 49, 61, 71] with neither fine-tuning nor any additional parameters given a pretrained CLIP model, which validates the good transferability of vision-language models in dense prediction tasks.
- Further, in this work, a minimal modification to CLIP yields very significant improvements in semantic segmentation, which provides with an important data point that the weakly-supervised pretraining paradigm with language guidance has very good potentials to function as a visual foundation model that supports a wide range of downstream tasks.

2 Related Work

Transferable Visual Foundation Models. Self-supervised pretraining has recently demonstrated good potentials in learning transferable visual representations. The models pretrained with re-constructive objectives such as masked image modeling [2, 23, 58] or discriminative objectives such as contrastive learning [7, 9, 21, 24, 26, 54] exhibit strong capabilities in adapting various visual tasks when sufficient downstream training data is available. Similarly, denoising diffusion models [15, 27, 50] that allow high-resolution conditional image generation

[1] Here we focus on transformer-based image encoders for CLIP. Compared with ResNet [25] encoders, the vision transformers [17] are more suitable for zero-shot transfer into semantic segmentation, since they have 1) a global receptive field, and 2) lower down-sampling ratios (*e.g.*, 16× for ViT-Base/16 *vs.* 32× for ResNet-50)

and Segment Anything models [30, 76] that facilitate semantic-agnostic image segmentation can also serve as foundation models with transferable visual features.

When incorporated with language guidance, such foundation models can be really powerful to allow open-vocabulary and zero-shot transfer learning for downstream visual tasks [35, 42, 44, 47]. A representative model is CLIP [44], which pioneers to align visual and textual features by contrastive pretraining. Based on this, a series of follow-up works extend its scale [28, 43, 66, 68], applications [22, 48, 61, 71], and downstream inference protocols [19, 53, 72].

Open-Vocabulary Segmentation. To fully utilize the advancements of vision-language models in zero-shot and open-vocabulary visual inference, extensive follow-up work has been initiated to investigate their applications in dense prediction tasks. For example, GroupViT [61] introduces group tokens into its vision encoder and pretrains with language guidance, leading to an open-vocabulary model that well applies to semantic segmentation tasks. Also, MaskCLIP [71] and CLIP Surgery [33] make simple modification to vision transformers and enables CLIP's coarse feature localization. The study of language-guided segmentation is continuously explored [8, 20, 32, 36, 39, 41, 46, 49, 59, 62–65, 69, 73].

Self-Attention for Dense Visual Features. A series of related research has demonstrated that the potential of vision transformers in extracting dense visual features can be augmented by employing varied self-attention mechanisms. For example, in contrast to the vanilla self attention used in CLIP and conventional vision transformers [17, 44], the Local Attention mechanism constrains the spatial feature aggregation within a local window so that to encourage fine-grained features [37, 38, 52, 57, 74]. Localized visual features can also be encouraged with modified self-attention mechanisms such as MaskCLIP [71] which discards the processing of query and key vectors in its last transformer layer (equivalent to local attention with window being one) and MSSA [67] that reduces the attention projections into a single matrix. In addition, some segmentation or detection-oriented transformer models leverage cross attention to map local visual features into semantic tokens [6, 10, 11, 51, 61]. Also, the models equipped with Axial Attention [16, 55, 56] or Deformable Attention [60, 75] demonstrate strong capabilities in dense prediction.

3 Method

The central concept of our method is transforming the spatial-invariant visual features learned from the CLIP paradigm into covariant representations by architectural modifications, so that the CLIP models can generalize to dense prediction tasks. As we discussed in Section 1, the ***spatial-invariant*** features indicate that the model produces similar representations for different locations within an image and they tend to share holistic information (see Figure 2), which is favorable in image-level tasks such as classification. In contrast, the ***spatial-covariant*** features encourage each local token to effectively represent the visual information of its corresponding position, which is conductive to pixel-level dense

prediction tasks such as semantic segmentation. We develop our approach SCLIP (Segmentation-adapted CLIP model) by introducing a new self-attention mechanism as it can re-organize the spatial information. The details can be found below.

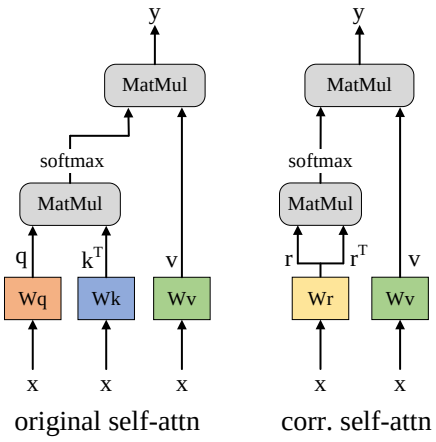

Fig. 3: An architectural comparison between the original self-attention and our correlative self-attention mechanism. Our method determines attention scores by pairwise correlations between the local tokens.

3.1 Re-Visiting the Original Self-Attention

In conventional vision transformers [17], each input image of size $3 \times w \times h$ is initially divided into a number of non-overlapping patches, with each patch subsequently being projected into a vectorized feature $\boldsymbol{x}_i \in \mathbb{R}^d$, where d denotes the dimension of the model's feature space. Each layer of the vision transformer receives a collection of visual tokens $\boldsymbol{X} = \{\boldsymbol{x}_{\text{cls}}, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_l\} \in \mathbb{R}^{(l+1)\times d}$ as input, with $\boldsymbol{x}_{\text{cls}} \in \mathbb{R}^d$ denoting the class token, $l = wh/p^2$ denoting the total number of image patches ($p \times p$ size for each), and each local visual token $\boldsymbol{x}_i \in \mathbb{R}^d$ ($i = 1, 2, \ldots, n$) associated with a distinct position within the input image.

We illustrate the pipeline of the traditional self-attention block in Figure 3 (left). Formally, the attention map $\boldsymbol{Attn} \in \mathbb{R}^{(l+1)\times(l+1)}$ is computed by

$$\boldsymbol{Attn} = \text{Softmax}\left(\boldsymbol{X}\boldsymbol{W}_q\boldsymbol{W}_k^T\boldsymbol{X}^T/\sqrt{d}\right), \tag{1}$$

where $\boldsymbol{W}_q, \boldsymbol{W}_k \in \mathbb{R}^{d\times d}$ are projection parameters learned from pretraining. Note that here we only consider the single-head self-attention for easy description. In

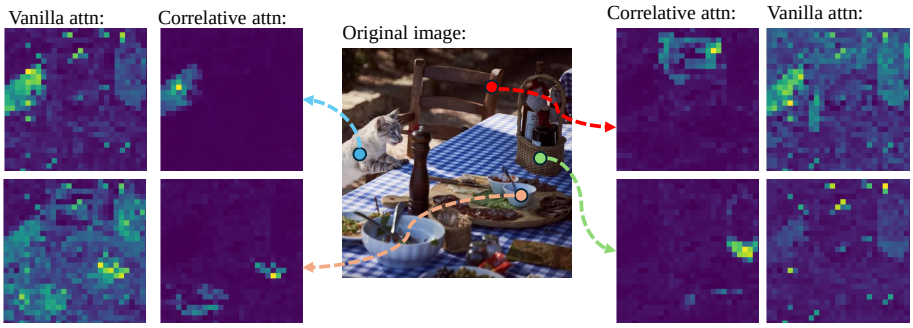

Fig. 4: Comparison of attention maps. We show the attention maps of the last transformer layer in CLIP vision encoder equipped with the original self-attention (right) and our correlative self-attention (left). Our correlative self-attention exhibits spatially covariant patterns as the attention maps are distinct to different source points and show clear boundaries of semantic objects (*e.g.*, the chair and the cat).

CLIP, the vision encoders are pretrained to represent each input image by a single feature vector, which encourages the self-attention blocks to extract holistic visual representations and consequently facilitates spatial-invariant features. As mentioned above, these invariant features prevent CLIP from performing dense prediction tasks, so necessary modifications should be made for its self-attention modules to allow semantic segmentation.

A very straightforward way to this end is forcing each visual token $\boldsymbol{x}_i$ only attending to itself, *i.e.*, setting the attention map $\boldsymbol{Attn}$ to an identical matrix $\boldsymbol{I}_{(l+1)\times(l+1)}$ regardless of the input. In this way, each local visual token only receives information from its corresponding position so that visual features are well localized. In practice, MaskCLIP [71] uses this attention map in CLIP vision encoder's last layer and obtains a non-trivial improvement in semantic segmentation. For example, it increases CLIP's mIoU on COCO-Stuff [5] from 5.7% to 16.7%. However, as this approach strictly constrains the receptive field of local tokens, the model may easily over-focus on low-level features and thus produces noisy dense predictions [8, 71].

3.2 Correlative Self-Attention

To facilitate spatial-covariant features, we introduce **C**orrelative **S**elf-**A**ttention (CSA) mechanism which computes attention scores by pairwise correlations across local tokens, with an overall pipeline illustrated in Figure 3. Formally, we have

$$\boldsymbol{Attn} = \text{Softmax}\left(\boldsymbol{X}\boldsymbol{W}_r\boldsymbol{W}_r^T\boldsymbol{X}^T/\tau\right), \tag{2}$$

where $\boldsymbol{X} \in \mathbb{R}^{(l+1)\times d}$ denotes the input and $\boldsymbol{W}_r \in \mathbb{R}^{d\times d}$ is the newly introduced projection matrix. The temperature coefficient τ is by default set to $\sqrt{d}$ following traditional self-attention. This change makes self-attention depend on the

distance between the feature vectors at different positions, with an underlying idea that the tokens $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ assign high attention scores to each other if they have high cosine similarity after a projection. Compared with the conventional mechanism, this correlative self-attention is more suitable for dense prediction tasks for the following reasons.

First, in vision transformers, the feature localization can be intuitively reflected by the magnitude of the diagonal elements of the matrix $\boldsymbol{Attn} \in \mathbb{R}^{(l+1)\times(l+1)}$. Specifically, each element $a_{ij} \in [0, 1]$ of $\boldsymbol{Attn}$ measures the attention score of $\boldsymbol{x}_i$ to $\boldsymbol{x}_j$, so high diagonal values indicate that each local token mainly attends to its own position and the visual information of each position is consequently well localized. This explains why MaskCLIP [71] works, where it forces $a_{ij} = 0, i \neq j$ and $a_{ij} = 1, i = j$. In the CSA module, the diagonal attention scores are also enhanced since the correlation between $\boldsymbol{x}_i\boldsymbol{W}_r$ and $\boldsymbol{x}_j\boldsymbol{W}_r$ always reaches its maximum when $i = j$ (supposing both vectors are normalized).

In addition to its notable feature localization abilities, the CSA module also thoroughly accounts for the semantic correlations across local tokens, so that it produces robust and smooth dense prediction results. Intuitively, for each local token $\boldsymbol{x}_i$, CSA imparts high attention scores not only to $\boldsymbol{x}_i$ itself but also to tokens that share similar semantic content. We visualize this effect in Figure 4, where for each source point, only the positions with high semantic similarity to it are assigned with noticeable attentions, and therefore the corresponding object (*e.g.*, the chair and the cat) of each source point can be clearly recognized in the attention maps.

Further, the matrix $\boldsymbol{W}_r$ in CSA functions as a distance measure between the features in different positions, so our model is not sensitive to the parameters of this projection layer since changing $\boldsymbol{W}_r$ only alters the form of the distance measure. In the experiments, we find that it is unnecessary to specifically train this projection matrix, but instead manually assigning it or even employing an ensemble of randomly initialized matrices can consistently obtain very competitive results (see Section 4.3 and Table 2 for details). Notably, CSA's insensitivity to model parameters provides with good potentials of zero-shot adaptation into dense prediction tasks when given a pretrained CLIP model. With this merit, we can develop our segmentation model using CSA without introducing any additional parameters nor any downstream fine-tuning.

3.3 Segmentation-Adapted CLIP Model

To develop our SCLIP approach, we employ a pretrained CLIP model with a ViT-Base/16 [17] image encoder as backbone. Generally, when it comes to adapting CLIP into a downstream task without introducing additional parameters, we actually regard its last or last several layers as a task-specific decoder head. Based on our observation of the self-attention patterns in different layers, we regard the last transformer block of CLIP's image encoder as the decoding layer to implement the adaptations while leaving the remaining components unchanged.

In this decoding layer, we replace the original self-attention block by our CSA module and reuse its parameters of $\boldsymbol{W}_q$ and $\boldsymbol{W}_k$ as our projection matrices.

Formally, we have

$$\begin{aligned} \boldsymbol{Attn} = &\operatorname{Softmax}\left(\boldsymbol{X}\boldsymbol{W}_q\boldsymbol{W}_q^T\boldsymbol{X}^T/\tau\right) \\ &+ \operatorname{Softmax}\left(\boldsymbol{X}\boldsymbol{W}_k\boldsymbol{W}_k^T\boldsymbol{X}^T/\tau\right), \end{aligned} \tag{3}$$

which makes a training-free adaptation since the matrices $\boldsymbol{W}_q$ and $\boldsymbol{W}_k$ can be directly loaded from CLIP.

Post-processing of dense visual features. In dense prediction tasks, we generally have a simple but very essential pre-hypothesis of spatial continuity, which suggests that in an image, the adjacent pixels or patches tend to share similar visual features. This prior knowledge can be easily introduced in fully supervised training as the labels of segmentation masks actually satisfy this hypothesis. However, in CLIP-like weakly supervised pretraining, there is no such explicit constraint to limit the spatial continuity of dense visual features, with only positional embeddings added in the input layer. Therefore, the existing zero-shot segmentation models often rely on specific post-processing strategies to refine or smooth their segmentation masks (*e.g.*, PAMR [1] for TCL [8] and DenseCRF [31] for ReCo [49]).

However, we argue that such post-processing approaches should not be employed by default since ensuring the spatial continuity of output is also an integral part of the inference capability of semantic segmentation models. In our experiments, we find SCLIP to be very robust in this aspect, which does not rely on any refinement or smoothing strategies to produce good segmentation results.

4 Experiments

4.1 Experiment Settings

Datasets. We evaluate our method on six commonly used semantic segmentation benchmarks, including PASCAL VOC 2012 [18], PASCAL Context [40], Cityscapes [12], ADE20k [70], COCO-Stuff and COCO-Object [5]. Considering the background category, we additionally evaluate on two variant datasets for PASCAL VOC and PASCAL Context. For clear reference, we denote VOC21, Context60 as the original datasets with a background class, and VOC20, Context59 as the variant without this category. In prior works such as GroupViT [61] and TCL [8], they evaluate with input images resized to have a shorter side of 448 and then performing slide inference with a 448×448 window and 224 stride. However, in our experiments, we find that using a smaller input size with denser sliding stride can lead to slightly higher results (*e.g.*, 0.2% mIoU on PASCAL Context). Specifically, we resize input images with a short side of 336 and perform slide inference with a 224×224 window and 112 stride. This protocol introduces a similar level of computation as that of GroupViT, yet better fits the original input size of CLIP (*e.g.*, 224 for ViT-Base) and is also friendly to parallel computing. For Cityscapes [12], we resize with a 560 shorter side due to the particular high resolution of its original images. A detailed comparison of image pre-processing protocols can be found in Table 4.

Table 1: Evaluation results (mIoU, %) of our method and the baseline models on eight semantic segmentation benchmarks. The methods with an asterisk symbol * denote using a PAMR [1] post-processing strategy which introduces heavy computation cost so we de-emphasize these results. Our results are marked in gray. The best results on each dataset are **bolded**.

Method	*With a background category*			*Without background category*					Avg.
	VOC21	Context60	Object	VOC20	City.	Ctx59	ADE20k	Stuff.	
CLIP [44]	18.8	9.9	8.1	49.4	6.5	11.1	3.1	5.7	14.1
MaskCLIP [71]	43.4	23.2	20.6	74.9	24.9	26.4	11.9	16.7	30.3
GroupViT [61]	52.3	18.7	27.5	79.7	18.5	23.4	10.4	15.3	30.7
ReCo [49]	25.1	19.9	15.7	57.7	21.6	22.3	11.2	14.8	23.5
TCL [8]	51.2	24.3	30.4	77.5	23.5	30.3	14.9	19.6	33.9
CLIP-Surg [33]	-	-	-	-	31.4	29.3	-	21.9	-
OVSeg. [63]	53.8	20.4	25.1	-	-	-	5.6	-	-
SegCLIP [39]	52.6	24.7	26.5	-	-	-	-	-	-
SCLIP (ours)	**59.1**	**30.4**	**30.5**	**80.4**	**32.2**	**34.2**	**16.1**	**22.4**	**38.2**
Approaches with pamr post-processing:									
CLIP*	19.8	8.7	10.4	54.2	7.0	11.7	3.6	5.9	15.2
MaskCLIP*	52.0	28.2	22.6	72.1	30.1	31.5	14.0	20.0	33.8
GroupViT*	52.7	19.5	27.9	81.5	21.7	24.4	11.8	16.9	32.1
ReCo*	27.2	21.9	17.3	62.4	23.2	24.7	12.4	16.3	25.7
TCL*	55.0	30.4	31.6	83.2	24.3	33.9	17.1	22.4	37.2
SCLIP* (ours)	**61.7**	**31.5**	**32.1**	**83.5**	**34.1**	**36.1**	**17.8**	**23.9**	**40.1**

Baselines. CLIP [44] is a direct baseline for our method to compare the difference of dense prediction performance between the original self-attention and our CSA mechanisms. In detail, we first extract textual embeddings of the target class names from CLIP's language encoder and then directly align them with CLIP vision encoder's dense features. We also consider the open-vocabulary semantic segmentation models derived from CLIP or similar vision language models as stronger baselines, which includes MaskCLIP [71], ReCo [49], and TCL [8]. For These methods, we report the higher numbers between our re-implementation based on their official code bases and results from the existed work [8]. We additionally compare them with recent baselines such as SegCLIP [39] and OVSegmentor [63], for which we directly take the results in their original papers.

Following TCL [8], we do not permit the post-processing strategies with very heavy computation cost such as Dense CRF [31], and do not consider the baselines that borrow well-pretrained models other than CLIP [29,62]. We by default discard the Pixel-Adaptive Mask Refinement (PAMR) [1] technique for post-processing of segmentation masks as it also introduces intensive computation and may easily obscure the inherent inference capabilities of the segmentation models.

4.2 Main Results

Table 1 summarizes the comparison of various zero-shot semantic segmentation models, where our SCLIP consistently achieves the best performance across eight evaluated benchmarks, with notable leads in PASCAL Context (34.2%), Cityscapes (32.2%), and ADE20k (16.1%). Overall, the average performance of SCLIP stands at 38.0%, which is significantly higher than the second-best average performance by TCL at 33.9%. This suggests that SCLIP provides a robust improvement over existing methods and testifies the significant effectiveness of the newly introduced correlative self-attention. Aside from the competitive baseline methods, we also report the evaluation results of the vanilla CLIP model with its original self-attention in the image encoder. As a result, this straightforward protocol fails to obtain a comparable performance as other baseline methods, indicating the incompatibility of directly transferring the original self-attention to dense prediction tasks.

In Table 1 there are also results of additionally employing a PAMR post-processing layer, where almost all approaches can benefit from it with a similar level of improvements. For example, our SCLIP attains a 1.9% average mIoU increase over the eight datasets while the baselines of GroupViT and TCL get 1.4% and 3.3%, respectively. Contrary to what is reported in the TCL paper [8] where MaskCLIP experiences a degradation in predictive performance after using the PAMR module, we find that by simply searching for suitable PAMR hyper-parameters, it can achieve a 3.5% increase in mIoU compared with its original version. We suggest to disable this refinement strategy in the default settings of open-vocabulary segmentation since it is computationally intensive, but instead turn to some lightweight smoothing methods for the predictions.

4.3 Ablation Study

Projection matrices in correlative self-attention. We want to find out the effect of choosing different types of projection matrices in our correlative self-attention block. As previously discussed, the CSA module theoretically accepts any non-zero projections as its $\boldsymbol{W}_r$, and we by default ensemble the $\boldsymbol{W}_q$ and $\boldsymbol{W}_k$ in CLIP's original self-attention for that (shown in Equation 3). Here we compare four more variants to testify its robustness.

1. *Identity Projection*: we directly measure the pairwise correlations by inputs $\boldsymbol{X}$, leading to a very simple protocol of $\boldsymbol{Attn} = \text{Softmax}(\boldsymbol{X}\boldsymbol{X}^T/\tau)$. Note that this is not equivalent to MaskCLIP which directly forces the $\boldsymbol{Attn}$ to be an identity matrix.
2. *Ensemble of Random Initializations*: we randomly initialize several projection matrices as $\boldsymbol{W}_r$ and then average their corresponding attention scores. Formally, we have $\boldsymbol{Attn} = \frac{1}{n}\sum_{i=1}^{n}\text{softmax}(\boldsymbol{X}\boldsymbol{W}_i\boldsymbol{W}_i^T\boldsymbol{X}^T/\tau)$.
3. *Projection with Single $\boldsymbol{W}_q$ or $\boldsymbol{W}_k$*: We load single $\boldsymbol{W}_q$ or $\boldsymbol{W}_k$ as our $\boldsymbol{W}_r$ to ablate the effect of combining both.

4. *Learned Projection*: To fully exploit the potential of CSA, we specifically learn a projection matrix from the training split of each dataset. The model is able to converge well with few training samples (we use 64 for each dataset) due to the few learnable parameters.

Table 2: Ablation results (mIoU, %) of projection matrices in correlative self-attention. n denotes the number of random projection matrices used in this experiment. Our default setting is marked in gray. The best result on each dataset is **bolded**.

Mode	PASCAL VOC	PASCAL Context	COCO-Stuff
Single projection matrix for CSA:			
Identity projection	57.5	33.0	21.5
W_q projection	58.2	33.5	21.7
W_k projection	58.4	33.1	21.8
Learned projection	**60.4**	**34.7**	**22.6**
Random projection matrices (average of 5 trials):			
$n = 1$	57.1	32.4	20.6
$n = 4$	58.0	32.7	20.9
$n = 16$	58.1	32.7	21.2
Default	59.1	34.2	22.4

The results are summarized in Table 2. Overall, the performance differences among the various modes on the three datasets are minimal, showcasing the robustness of the proposed CSA mechanism. This is particularly notable in scenarios where only a single matrix is randomly initialized, which still yields respectable results, such as an mIoU of 57.1% on the PASCAL VOC dataset. Also, while the learned projection mode achieves the highest performance, the margin of improvement over the default training-free structure is not substantial. Given this modest gain, investing significant effort into in-domain training for learned projections may not be recommended. Excluding the learned projection, our default method consistently attains the best results. This suggests that the proposed CSA is highly compatible with the pretrained projection parameters from CLIP. This compatibility is a testament to the efficacy of CSA when paired with the robust features provided by CLIP's pretrained projections.

Alternative approaches for feature localization. There also exist some potential approaches to enable CLIP localizing visual features. For example, we can sharpen the attention map by simply adjusting the temperature parameters of the CLIP vision encoder, which prevents it from overly attending to global information and instead concentrates the features on a few specific positions. We denote this approach as Attention Sharpening and compare its effect to our method. Similarly, by employing local attention techniques, *i.e.*, calculating attention scores only within a given window, we can facilitate the CLIP model in anchoring visual features to their corresponding locations. However, this method

Table 3: Ablation results of potential approaches for feature localization. Our default setting is marked in gray. The best results are **bolded**.

Approach	VOC21	Ctx59	Stuff
Attention sharpening			
$\tau = 8$ (CLIP default)	18.8	**13.3**	**5.7**
$\tau = 2$	**21.7**	9.5	4.1
$\tau = 0.5$	15.6	6.0	4.1
$\tau \to 0$ (hard max)	14.8	5.7	4.2
Local attention			
window size = 3	**42.9**	**25.5**	**16.0**
window size = 5	30.5	18.3	8.2
window size = 7	28.1	17.9	8.0
Attention map from early stages			
from layer #1	41.5	26.2	**16.8**
from layer #3	**43.0**	26.4	16.2
from layer #5	41.7	**26.8**	15.4
from layer #7	21.9	17.3	10.1
SCLIP (ours)	**59.1**	**34.2**	**22.4**

comes at the cost of losing the global receptive field inherent in the vision transformer models, preventing the model from reasoning with the assistance of tokens outside the local domain. It's noteworthy that the MaskCLIP [71] algorithm can be considered as a special case of local attention when the window size is set to one. We also observe that actually the early stages of the vision transformer attend to relatively small local regions. Therefore, a possible way for CLIP feature localization is to directly borrow the attention maps in early stages in place of those in the decoding layer.

As summarized in Table 3, the three alternative strategies may offer considerable enhancements over the baseline CLIP model when specific parameters are adjusted, yet they fall notably short when compared with our method. Specifically, the attention sharpening approach fails to obtain performance improvements in most cases, and only achieves a 2.9% mIoU gain on PASCAL VOC with $\tau = 2$. When we apply local attention with a window size of three, the evaluation performance is promising and almost parallels that of MaskCLIP across three different datasets. In addition, the heuristic approach of directly borrowing attention maps from early stages shows relatively better results, with a 26.8% mIoU on PASCAL Context59 and a 16.8% mIoU on COCO-Stuff, which even outperforms MaskCLIP.

This ablation study suggests that the mere focus on local visual features does not effectively convert a weakly supervised pre-trained model such as CLIP for semantic segmentation challenges. In contrast, our method, which incorporates a correlative self-attention mechanism that considers the relationship between local features and overarching semantic contexts, proves to be more adept for visual reasoning tasks across diverse scales.

Image pre-processing. As discussed in Section 4.1, we adopts a new pre-processing protocol that resizes each input image with the shorter side fixed to 336 instead of 448, and perform slide inference with a smaller window of 224 and stride of 112 than previous methods [8, 61]. To ablate the effect of this protocol, we present a detailed comparison of different pre-processing strategies in Table 4.

Table 4: Ablation results of image pre-processing on PASCAL VOC. "Img size" in this table denotes the length of the shorter side of the resized image. Our default setting is marked in gray. The best result is **bolded**.

Mode	Img size	Window	Stride	Flops	VOC21	Context59	COCO-Stuff
#1	224	224	112	1×	56.5	32.0	20.5
#2	336	224	112	∼ 4×	59.1	34.2	22.4
#3	336	336	112	∼ 3×	58.6	34.1	21.9
#4	448	224	112	∼ 9×	**60.4**	**35.3**	**23.4**
#5	448	448	224	∼ 4×	58.9	34.3	22.1

As is shown, in general, larger image sizes combined with smaller windows and strides lead to better performance, although they come with an increased computational cost as indicated by the higher number of Flops. Specifically, utilizing too small an image size results in substantial information loss and a marked decrease in performance, as seen with mode #1, which achieves only a 56.5% mIoU. Larger image sizes can enhance prediction accuracy (as demonstrated by mode #4), yet the improvement is not significant compared to the default setting (mode #2).

Compared to the established default settings of existing work (mode #5), the proposed protocol achieves better results with an equivalent amount of computation. This is possibly attributed to two factors: first, CLIP inherently performs better with its original input size of 224×224 pixels without fine-tuning; and second, our setting reduces the window stride, leading to smoother outputs. Furthermore, even when using the same pre-processing approach (mode #5), our SCLIP outperforms the existing (SoTA) model, with a 58.9% mIoU compared to TCL's 51.2% on PASCAL VOC.

5 Conclusion

In this work, we propose to enhance CLIP's potentials for dense prediction tasks by introducing a novel correlative self-attention mechanism, which functions as a task-specific decoder head for semantic segmentation in our approach. The adaptation significantly improves its performance in dense vision-language inference, achieving a 38.2% average zero-shot mIoU across eight benchmarks evaluated in this paper, outperforming the existing state-of-the-art models by a large margin. We demonstrate that minimal modifications to the existing CLIP model can yield substantial improvements in its functionality. The significant increase in zero-shot mIoU scores across various benchmarks testifies to the effectiveness of our approach. Notably, our model outperforms the existing baseline methods without any fine-tuning or additional parameters involved, which underscores the robust potential of the CLIP-like weakly-supervised pretraining paradigm in creating versatile visual foundation models.

Acknowledgements

This work was supported by ONR N00014-23-1-2641.

References

1. Araslanov, N., Roth, S.: Single-stage semantic segmentation from image labels. In: CVPR (2020)
2. Bao, H., Dong, L., Wei, F.: BEiT: BERT pre-training of image transformers. arXiv preprint arXiv:2106.08254 (2021)
3. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., et al.: On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 (2021)
4. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. NeurIPS (2020)
5. Caesar, H., Uijlings, J., Ferrari, V.: Coco-stuff: Thing and stuff classes in context. In: CVPR (2018)
6. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV (2020)
7. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: ICCV (2021)
8. Cha, J., Mun, J., Roh, B.: Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs. In: CVPR (2023)
9. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML (2020)
10. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: CVPR (2022)
11. Cheng, B., Schwing, A., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation. NeurIPS (2021)
12. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR (2016)
13. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: CVPR (2009)
14. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
15. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. NeurIPS (2021)
16. Dong, X., Bao, J., Chen, D., Zhang, W., Yu, N., Yuan, L., Chen, D., Guo, B.: Cswin transformer: A general vision transformer backbone with cross-shaped windows. In: CVPR (2022)
17. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
18. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. IJCV (2015)

19. Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.: Clip-adapter: Better vision-language models with feature adapters. arXiv preprint arXiv:2110.04544 (2021)
20. Ghiasi, G., Gu, X., Cui, Y., Lin, T.Y.: Scaling open-vocabulary image segmentation with image-level labels. In: ECCV (2022)
21. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., Piot, B., kavukcuoglu, k., Munos, R., Valko, M.: Bootstrap your own latent a new approach to self-supervised learning. In: NeurIPS (2020)
22. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. arXiv preprint arXiv:2104.13921 (2021)
23. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: CVPR (2022)
24. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR (2020)
25. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
26. Hénaff, O.J., Koppula, S., Alayrac, J.B., Van den Oord, A., Vinyals, O., Carreira, J.: Efficient visual pretraining with contrastive detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10086–10096 (2021)
27. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
28. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: ICML (2021)
29. Karazija, L., Laina, I., Vedaldi, A., Rupprecht, C.: Diffusion models for zero-shot open-vocabulary segmentation. arXiv preprint arXiv:2306.09316 (2023)
30. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
31. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials. NeurIPS (2011)
32. Li, K., Wang, Z., Cheng, Z., Yu, R., Zhao, Y., Song, G., Liu, C., Yuan, L., Chen, J.: Acseg: Adaptive conceptualization for unsupervised semantic segmentation. In: CVPR (2023)
33. Li, Y., Wang, H., Duan, Y., Li, X.: Clip surgery for better explainability with enhancement in open-vocabulary tasks. arXiv preprint arXiv:2304.05653 (2023)
34. Liang, P.P., Zadeh, A., Morency, L.P.: Foundations and recent trends in multimodal machine learning: Principles, challenges, and open questions. arXiv preprint arXiv:2209.03430 (2022)
35. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. arXiv preprint arXiv:2304.08485 (2023)
36. Liu, Q., Wen, Y., Han, J., Xu, C., Xu, H., Liang, X.: Open-world semantic segmentation via contrasting and clustering vision-language embedding. In: ECCV (2022)
37. Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., et al.: Swin transformer v2: Scaling up capacity and resolution. In: CVPR (2022)
38. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV (2021)

39. Luo, H., Bao, J., Wu, Y., He, X., Li, T.: Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. In: ICML (2023)
40. Mottaghi, R., Chen, X., Liu, X., Cho, N.G., Lee, S.W., Fidler, S., Urtasun, R., Yuille, A.: The role of context for object detection and semantic segmentation in the wild. In: CVPR (2014)
41. Mukhoti, J., Lin, T.Y., Poursaeed, O., Wang, R., Shah, A., Torr, P.H., Lim, S.N.: Open vocabulary semantic segmentation with patch aligned contrastive learning. In: CVPR (2023)
42. Nichol, A.Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., Mcgrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In: ICML (2022)
43. Pham, H., Dai, Z., Ghiasi, G., Kawaguchi, K., Liu, H., Yu, A.W., Yu, J., Chen, Y.T., Luong, M.T., Wu, Y., et al.: Combined scaling for open-vocabulary image classification. arXiv preprint arXiv:2111.10050 (2021)
44. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
45. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research (2020)
46. Ranasinghe, K., McKinzie, B., Ravi, S., Yang, Y., Toshev, A., Shlens, J.: Perceptual grouping in contrastive vision-language models. In: ICCV (2023)
47. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
48. Shen, S., Li, L.H., Tan, H., Bansal, M., Rohrbach, A., Chang, K.W., Yao, Z., Keutzer, K.: How much can clip benefit vision-and-language tasks? arXiv preprint arXiv:2107.06383 (2021)
49. Shin, G., Xie, W., Albanie, S.: Reco: Retrieve and co-segment for zero-shot transfer. NeurIPS (2022)
50. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: ICML (2015)
51. Strudel, R., Garcia, R., Laptev, I., Schmid, C.: Segmenter: Transformer for semantic segmentation. In: ICCV (2021)
52. Tu, Z., Talebi, H., Zhang, H., Yang, F., Milanfar, P., Bovik, A., Li, Y.: Maxvit: Multi-axis vision transformer. In: ECCV (2022)
53. Wang, F., Li, M., Lin, X., Lv, H., Schwing, A.G., Ji, H.: Learning to decompose visual features with latent textual prompts. arXiv preprint arXiv:2210.04287 (2022)
54. Wang, F., Wang, H., Wei, C., Yuille, A., Shen, W.: Cp2: Copy-paste contrastive pretraining for semantic segmentation. ECCV (2022)
55. Wang, H., Zhu, Y., Adam, H., Yuille, A., Chen, L.C.: Max-deeplab: End-to-end panoptic segmentation with mask transformers. In: CVPR (2021)
56. Wang, H., Zhu, Y., Green, B., Adam, H., Yuille, A., Chen, L.C.: Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In: ECCV (2020)
57. Wang, W., Chen, W., Qiu, Q., Chen, L., Wu, B., Lin, B., He, X., Liu, W.: Crossformer++: A versatile vision transformer hinging on cross-scale attention. arXiv preprint arXiv:2303.06908 (2023)
58. Wei, C., Fan, H., Xie, S., Wu, C.Y., Yuille, A., Feichtenhofer, C.: Masked feature prediction for self-supervised visual pre-training. In: CVPR (2022)
59. Wu, S., Zhang, W., Xu, L., Jin, S., Li, X., Liu, W., Loy, C.C.: Clipself: Vision transformer distills itself for open-vocabulary dense prediction. arXiv preprint arXiv:2310.01403 (2023)

60. Xia, Z., Pan, X., Song, S., Li, L.E., Huang, G.: Vision transformer with deformable attention. In: CVPR (2022)
61. Xu, J., De Mello, S., Liu, S., Byeon, W., Breuel, T., Kautz, J., Wang, X.: Groupvit: Semantic segmentation emerges from text supervision. In: CVPR (2022)
62. Xu, J., Liu, S., Vahdat, A., Byeon, W., Wang, X., De Mello, S.: Open-vocabulary panoptic segmentation with text-to-image diffusion models. In: CVPR (2023)
63. Xu, J., Hou, J., Zhang, Y., Feng, R., Wang, Y., Qiao, Y., Xie, W.: Learning open-vocabulary semantic segmentation models from natural language supervision. In: CVPR (2023)
64. Xu, M., Zhang, Z., Wei, F., Hu, H., Bai, X.: Side adapter network for open-vocabulary semantic segmentation. In: CVPR (2023)
65. Yi, M., Cui, Q., Wu, H., Yang, C., Yoshie, O., Lu, H.: A simple framework for text-supervised semantic segmentation. In: CVPR (2023)
66. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: Coca: Contrastive captioners are image-text foundation models. arXiv preprint arXiv:2205.01917 (2022)
67. Yu, Y., Chu, T., Tong, S., Wu, Z., Pai, D., Buchanan, S., Ma, Y.: Emergence of segmentation with minimalistic white-box transformers. arXiv preprint arXiv:2308.16271 (2023)
68. Yuan, L., Chen, D., Chen, Y.L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., et al.: Florence: A new foundation model for computer vision. arXiv preprint arXiv:2111.11432 (2021)
69. Yun, S., Park, S.H., Seo, P.H., Shin, J.: Ifseg: Image-free semantic segmentation via vision-language model. In: CVPR (2023)
70. Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ade20k dataset. IJCV (2019)
71. Zhou, C., Loy, C.C., Dai, B.: Extract free dense labels from clip. In: ECCV (2022)
72. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. arXiv preprint arXiv:2109.01134 (2021)
73. Zhou, Z., Lei, Y., Zhang, B., Liu, L., Liu, Y.: Zegclip: Towards adapting clip for zero-shot semantic segmentation. In: CVPR (2023)
74. Zhu, L., Wang, X., Ke, Z., Zhang, W., Lau, R.W.: Biformer: Vision transformer with bi-level routing attention. In: CVPR (2023)
75. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)
76. Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Gao, J., Lee, Y.J.: Segment everything everywhere all at once. arXiv preprint arXiv:2304.06718 (2023)