MAPS: Memory Augmented Panoptic Segmentation

Vatsal Agarwal[1], Saksham Suri[1], Max Ehrlich[1], and Abhinav Shrivastava[1]

University of Maryland, College Park

Abstract. Recently, there has been increased interest in developing models that can use external knowledge for both vision and natural language processing tasks. When faced with complex or unfamiliar scenes, the network can then retrieve instances that share similar visual concepts from its external memory and use them to enrich its visual recognition capabilities. In this work, we leverage the representational power of vision-language models and the abundance of large-scale data to build a memory bank for explicit retrieval. In contrast to other external memory designs, we exploit available labels in the form of object/part-segmentation maps to ensure that the memory stores pertinent granular visual concepts. Moreover, our memory is highly modular and can be easily substituted depending on the task. We introduce a paradigm to equip networks with our memory bank for standard image-segmentation tasks including panoptic, instance, and semantic segmentation. Our results highlight a remarkable improvement in performance over state-of-the-art architectures on ADE20K, COCO, and Cityscapes.

Keywords: Memory Augmented Recognition · Panoptic Segmentation

1 Introduction

Humans have a remarkable capacity to learn and perceive the visual world around them. Imagine looking for your keys in a cluttered office space. As you glance across the room, your eyes might filter for smaller objects that appear shiny and metallic. But how did your brain know to focus on these characteristics? Simply, it had seen enough keys to know that they had these attributes and effortlessly stored that knowledge in memory [26, 46]. This life-long learning enables humans to build a vast library of visual patterns that can be used flexibly across various environments and tasks.

Studies in neuroscience have highlighted the role of memory in providing top-down signals relevant to the given task [11, 13, 18]. More specifically, there has been a plethora of research indicating that humans rely heavily on contextual cues for object recognition and localization. Say that we perceive our current environment to appear similar to a kitchen. Based on our previous experiences, we would anticipate seeing objects such as utensils or appliances. These expectations prime our eyes to pay attention to features corresponding to those objects.

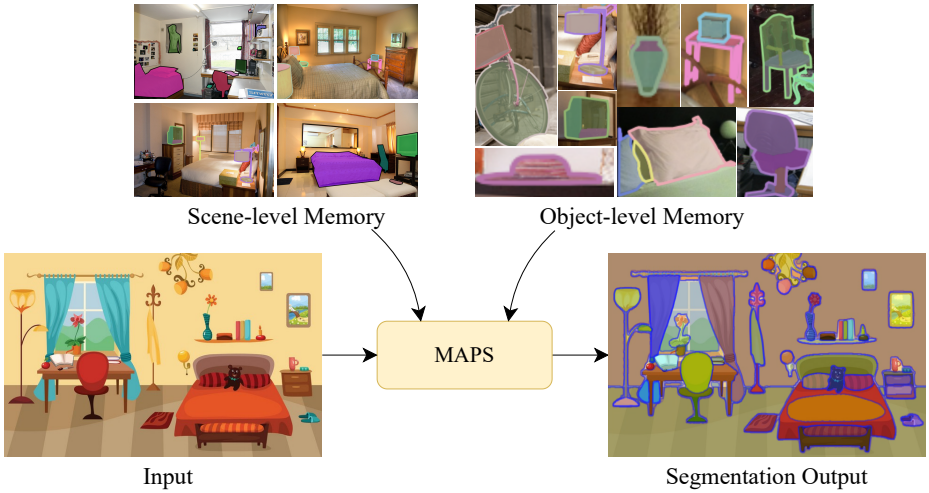

Fig. 1: We present **MAPS**, our framework for **M**emory **A**ugmented **P**anoptic **S**egmentation. We incorporate scene level and object level memory to improve performance on downstream tasks. The memory helps introduce more context as well as diverse instances of related objects coming from similar scenes.

While such feedback is crucial for human visual processing, contemporary deep networks are limited in harnessing both contextual and object-level information to guide information processing. Instead, current models learn to encode such knowledge directly in their weights via large-scale training [55,58]. Although this has shown great promise, these models struggle to integrate such information to learn more fine-grained relations. Furthermore, updating the model's knowledge requires further training or fine-tuning of the model which may not be computationally feasible for large networks [20, 32].

As a result, there has been an increased interest in building retrieval-augmented methods for various vision tasks [19, 20, 30, 32, 53]. These works propose building an external database consisting of images and their corresponding captions. Then during training, the model can find similar images and learn to fuse the retrieved image and text features generated from a pre-trained vision-language model. However, such a design is restrictive and cumbersome. Constructing a large database is challenging and storing it in memory may also be prohibitively expensive. Additionally, these approaches primarily rely on high-level information for retrieving similar examples and may fail to adapt to tasks requiring more fine-grained knowledge.

In this paper, we present **MAPS** a framework that enables **M**emory **A**ugmented **P**anoptic **S**egmentation. Specifically, we propose an alternative paradigm to build a general-purpose memory bank that is both scalable and modular. Most notably, we formulate our memory as a two-stage pipeline where our memory bank stores scene and object-level information. Our motivation behind this design is that the scene-level information captures the associations between context

and objects, while the object-level stores structural knowledge about the object. The combination of both enables our memory to store valuable information across multiple granularities. Furthermore, we propose using labeled datasets to generate our memory. With such a strategy, we aim to improve the precision and robustness of the memory features. It is important to note however that our framework can easily be extended to weakly-labeled and unlabeled datasets as well.

Following previous works and to best capitalize on the increasing availability of multi-modal image-text data, we propose using CLIP [37] to encode image features for our memory. To demonstrate the effectiveness of our proposed memory design, we apply it to the task of image segmentation, which is challenging as it requires the model to both accurately localize and recognize a vast number of objects. In particular, we describe how to take an existing state-of-the-art segmentation network, namely Mask2Former [6], and augment it with our proposed memory pipeline. Our results indicate significant and consistent improvement over challenging baselines on the COCO [29], ADE20K [62] and Cityscapes [9] datasets for panoptic, instance, and semantic segmentation.

An overview of the full pipeline is shown in Fig. 1 and consists of two steps. Given a new scene, we first find semantically similar images and retrieve their global features which encode what objects are in the scene and the relationships between them. For instance, if the network perceives the scene to be similar to a bedroom, when it queries the memory, it will retrieve features corresponding to other bedrooms. In this case, the global representations capture important contextual information such as the objects in the room (e.g. beds and shelves) as well as some of the compositional information (e.g. a bed is generally next to the shelf). We fuse this retrieved knowledge with the network's current features using a learnable cross-attention module.

From these same images, we then retrieve more granular information about the objects in each scene. In the case of the bedroom, our memory can return feature embeddings corresponding to beds or lamps. On the other hand, if the scene is similar to a kitchen, the network could obtain features related to appliances such as a stove or utensils such as forks and knives. These granular features implicitly encode structural knowledge about each object and can be used to help the network reinforce or correct its understanding of different objects in the scene. We again utilize cross-attention to enhance our current features with these retrieved features. Finally, the enriched features are used to generate the final segmentation masks.

In summary, our contributions are as follows:

- We propose a new paradigm for constructing and integrating external knowledge via scene-level and object-level memory to incorporate high-level and fine-grained information respectively. This memory is highly modular and scales easily with more data.
- We introduce a lightweight pipeline to seamlessly equip state-of-the-art networks with our scene-level and object-level memory and demonstrate its

effectiveness for the Mask2Former architecture, which is an advanced state-of-the-art segmentation network.
– We find that our proposed memory pipeline significantly improves model performance on challenging benchmark datasets.

2 Related Works

Context-Based Methods. The role of context plays in understanding the world around us has been a focus of study in both cognitive neuroscience and computer vision [14, 18, 39, 40, 59]. Numerous studies in psychology highlight how the human visual system relies on contextual relationships for detecting and recognizing objects in various settings [13]. It has been theorized that the human eye can quickly capture a "scene gist" and use this information in a top-down manner to retrieve prior knowledge that may be useful for parsing the scene. This knowledge comes in the form of contextual associations as the presence of some objects can trigger predictions about other objects that may be in the scene and thereby impact guesses about what the object is [11, 13].

In computer vision, there has been considerable research in trying to harness contextual information to address various vision problems such as image classification [35], object detection [43], crowd counting [41], and scene parsing [12, 61]. Early seminal works modeled the contextual associations for object detection via conditional random fields [36] or modeling feature statistics [45]. [33] proposes modeling object relationships for scene parsing via an exemplar-based model where each relationship can be categorized as either a contextual relation or structural similarity. [35] aims to reconcile compositionality and contextual relationships to build classifiers that can generalize to unseen compositions of seen concepts. More recently, [31] aims to understand contextual relationships as a self-supervised learning task and encode this knowledge via a learnable external memory module. Our work builds upon this literature as we aim to encode contextual relationships in our memory for the universal segmentation task.

Retrieval-Augmented Networks. Recently, there has been increased interest in leveraging external knowledge to aid deep networks for a variety of tasks in both computer vision and natural language processing [30, 32, 42, 47, 49, 52, 53]. The primary goal of these works is to harness the potential of large-scale image-text or text data to improve the transfer ability of general-purpose models such as CLIP [37] for specific vision tasks. Our work differs from these approaches in several key ways. Most notably, these works focus on using external knowledge to improve contrastive-learning models, while we primarily focus on extending an existing segmentation network for the universal segmentation task. [49] is most similar to our work in that they also propose using training data to form their retrieval knowledge base and demonstrate improved performance when using task-specific labels. Our goal in this work is to extend this idea to computer vision and use available labeled data as a way for storing contextual and object-specific relationships.

Image Segmentation. Image segmentation requires densely labeling every image pixel. One of the seminal works around this was the Mask R-CNN [16]

which proposed adding a mask head to the Faster R-CNN [38] style detection networks for instance segmentation. There have been works built over this [1, 4, 25, 44] with one of the more recent ones being ViTDet [27] which replaced the backbone architecture to a state-of-art MAE transformer [10, 15]. For a related and more challenging task of Panoptic Segmentation [24] there have been works like Panoptic-DeepLab [5], Panoptic-FPN [24] which have proposed specialized architectures for this task. DETR [2] showcased a new way to do panoptic segmentation in an end-to-end manner inherently learning proposals and labels through learnable queries. Deformable-DETR [64] built over a similar architecture and made it much more efficient to train while improving the performance. Both of them relied on the transformer architecture which made use of self-attention operation. Along similar lines, MaskFormer [7] also does both semantic as well as panoptic segmentation. Panoptic SegFormer [28] builds upon Deformable-DETR and introduces a deep supervision strategy to speed up training and decouples the query between things and stuff for improved performance. K-Net [60] furthermore introduces a unified segmentation approach that excels in the panoptic segmentation task. We build over Mask2Former [6] which is a popularly used architecture and can perform "universal image segmentation" which includes panoptic, instance or semantic segmentation.

3 Method

In this section, we introduce some preliminaries and then detail the specifics of our memory design and construction. We then present a seamless pipeline to equip the current state-of-the-art Mask2Former network with our proposed memory retrieval.

3.1 Revisiting Mask2Former

Recently, there has been significant progress made in developing models for unified image segmentation. Namely, these architectures take inspiration from DETR and formulate segmentation as a mask classification problem. To this end, they learn a set of query tokens to capture the things and stuff in the image. These queries are then used to generate the final set of segmentation masks. This approach is both simple and effective and many following works have obtained state-of-the-art results across multiple segmentation and scene-parsing tasks. As such, we build our framework around this DETR-style architecture and adopt Mask2Former as our base model.

Here, we spend some time reviewing the key aspects of the Mask2Former architecture. Rather than performing a per-pixel classification, Mask2Former and other similar works learn to predict a set of binary masks and their corresponding class label. To do this, they represent each potential segment as a learnable feature vector, also known as an object query.

Formally, given an image, a pre-trained backbone network is used to extract an intermediate feature representation containing high-level semantics. These

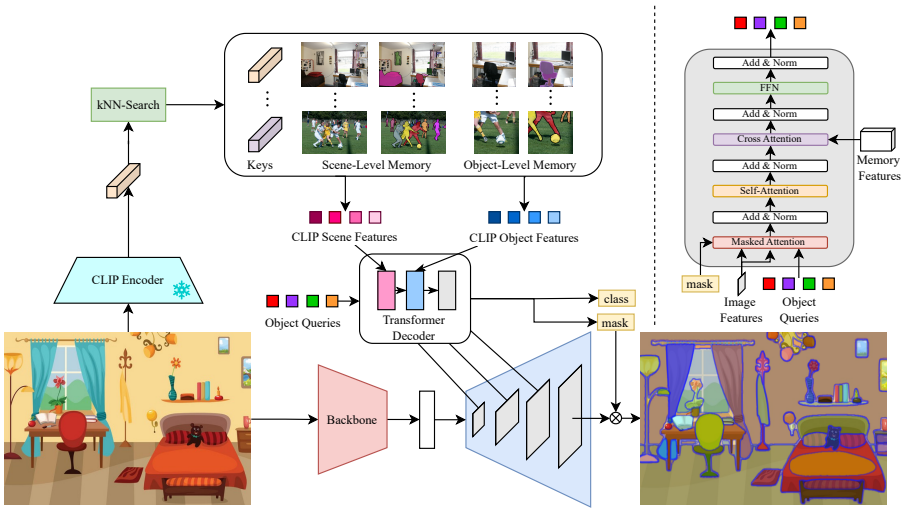

Fig. 2: Overview of MAPS. For a given image during training we encode it using a frozen CLIP encoder. The extracted embedding is used to query our scene-level and part-level memory. These memories are also constructed using CLIP embeddings and encode scene-level as well object-level information. Once the relevant things are retreived from both memories, a cross attention layer modifies the learnable queries in the model by interacting with retrieved features. This interaction with memory features through cross attention helps introduce more context of the scene as well as instance features coming from diverse instances. Finally, the model predicts segmentations for the input image using the modified queries.

features are then progressively upsampled via a convolutional pixel decoder network. Simultaneously, a transformer decoder is fed these pixel-level features to refine the initial set of object query features. This occurs over a series of nine attention layers. After multiple rounds of self-attention through the transformer decoder, the final object queries are then used to generate per-object classifications. These queries are also combined with the final set of pixel features to generate each of the binary masks for the full panoptic segmentation. To improve the performance, deep supervision is also applied and intermediate masks are generated at each stage of the pixel-decoder.

In this work, we propose modifying the transformer decoder to use the retrieved memory features to refine the object query features. In the following sections, we detail the design and intuition behind our proposed memory construction as well as the integration strategy used to fuse memory features with the segmentation model.

3.2 Building a Robust External Memory

Our memory pipeline is composed of two separate memory banks, namely a scene-level memory and an object-level memory. We formulate our memory as

Table 1: Panoptic Segmentation results on COCO `panoptic val` set with 133 categories.

Method	Backbone	Params	PQ	PQ^{th}	PQ^{st}	AP^{Th}_{pan}	$mIOU_{pan}$
Panoptic-FPN [23]	R101 [17]	-	39.0	45.9	28.7	33.3	41.0
Panoptic-DeepLab [5]	X71 [8]	-	39.7	43.9	33.2	-	-
SOLOv2 [50]	R50	-	42.1	49.6	30.7	-	-
DETR [2]	R50	43M	43.4	48.2	36.3	31.1	-
Panoptic FCN [54]	R50	37M	43.6	49.3	35.0	-	-
K-Net [60]	R50	-	47.1	51.7	40.3	-	-
CMT-DeepLab [56]	R50	-	48.5	-	-	-	-
MaskFormer [7]	R50	45M	46.5	51.0	39.8	33.0	57.8
Panoptic-Segformer [28]	R50	51M	49.6	54.4	42.4	-	-
Mask2Former† [6]	R50	44M	51.7	57.6	43.0	**41.7**	61.0
OneFormer [21]	R50	47M	51.5	-	-	42.5	61.2
MAPS	R50	45.9M	**52.0**	**57.7**	**43.4**	**41.7**	**62.1**

† indicates our own training results

a set of key-value pairs and choose CLIP as our feature extractor due to its ability to extract semantically rich feature representations for both image and text data. Specifically, we use a pre-trained CLIP-ViT-B model with a patch size of 16 as our image encoder and freeze it before generating the memory features.

Scene Memory for Context-Object Associations The scene-level memory bank stores information about what objects make up a particular scene. When retrieved, this knowledge can be used by the network to refine its current object-level features. We choose the COCO 2017 [29] training set for our memory due to its wide variety of scenes and objects as well as its extensive number of annotations including panoptic segmentation masks.

For a given image, we use the CLS token from our image encoder as the key to compactly store global "scene gist" information for efficient querying. These features make up the values for the first stage of our memory as well. Preliminary experiments demonstrate that the CLS tokens encode relevant information to retrieve semantically similar scenes and we show some example retrievals in Figure 3. Given that ViT architectures rely on the CLS token to encode global representations, we argue that they also effectively store scene-level contextual information that can then be exploited for the segmentation task.

Incorporating Object-Level Priors The goal of our object-level memory is to provide the network with more structural knowledge about different objects such as parts and attributes. We hypothesize that such granular knowledge can aid the network in discerning more fine-grained nuances between similar objects and thereby help in the downstream task. Broadly, we encode objects in our memory as tokens by pooling extracted spatial image features using the provided segmentation masks. This entails that our extracted features have good spatial granularity. We discuss how we achieve this below.

Table 2: Panoptic Segmentation results on ADE20K `val` set with 150 categories.

Method	Backbone	PQ	PQ^{th}	PQ^{st}	AP^{Th}_{pan}	$mIOU_{pan}$
MaskFormer [7]	R50	34.7	32.2	39.7	-	-
Mask2Former† [6]	R50	39.1	39.8	38.7	26.2	45.3
kMaX-DeepLab [57]	R50	41.5	-	-	-	45.0
OneFormer [21]	R50	**41.9**	-	-	27.3	47.3
MAPS	R50	40.7	40.3	41.5	27.0	47.0
MAPS-Aug	R50	41.0	40.3	42.2	**27.3**	**48.6**
MAPS-Aug-CLS	R50	41.1	**40.3**	**42.8**	**27.3**	48.3

† indicates our own training results

First, similar to the scene-level memory, we extract CLS tokens for each image and use them as the keys for indexing. However, in this memory, we wish to store object-level information and thus require extracting dense features. Thus, directly using the CLIP ViT-B backbone is problematic as its input resolution is fixed to 224×224 and thus the intermediate grid features only have 14×14 resolution and therefore lack the necessary spatial granularity to extract robust object embeddings. The fixed input-size requirement of ViTs is an additional limitation as the input images must be downsized or cropped before being fed to the network, resulting in a loss of information.

To overcome this challenge, we follow the protocol proposed by [48] to produce dense CLIP spatial features. Namely, for any image with resolution $H \times W$, we resize the image such that the smaller image dimension is set to 448 and scale the larger dimension such that it is divisible by the model patch size p. Then a tiling strategy is applied where the image is split into 224×224 crops and each crop is fed to the CLIP image encoder. Overlapping crops are taken as well to handle the remaining areas and the non-overlapping features are mapped to the final output feature with a resolution that roughly is $H/p \times W/p$. Please refer to [48] for more details about the dense extraction strategy. Given this output feature map and a set of binary masks for each object in the scene, we resize the features to match the mask resolution and then pool them for each mask giving us a set of object-level representations.

3.3 Augmenting Mask2Former with Memory

Here, we describe the full procedure for augmenting the Mask2Former model with our memory pipeline as shown in Fig. 2. In designing this protocol, we aim to make as few changes as possible to the original architecture to ensure ease of adoption. Moreover, we take great care to introduce as few parameters as possible to the architecture and also try to reduce additional memory usage.

We make two modifications to the Mask2Former training pipeline architecture. First, we use a pre-trained CLIP image encoder to generate index features during the forward pass for each image in the training set. It is important to note that while the rest of the Mask2Former architecture is learnable, the CLIP encoder is frozen during training. This helps us leverage the semantics already captured by CLIP. Our next change is the addition of a lightweight cross-attention

layer to transform the learnable object query features using the retrieved CLIP memory features. This is required to incorporate the knowledge from the memory into the training. Next, we will delve deeper into the design choices of how scene-level and object-level memory is retrieved and used to augment the object query features.

Memory Integration We apply the same protocol for both the scene and object-level memory. While generating the index features via CLIP is fairly straightforward, it is not immediately clear how the retrieved features should be fused with the learnable object queries and how many samples should be retrieved. Directly using the retrieved features or simply adding them to the queries may cause the performance to degrade as the object queries need to learn not only how to recognize an object, but also how to localize it. To this end, we propose an adaptive fusion mechanism via cross-attention layers to fuse the query tokens with the retrieved scene features. We place this cross-attention module within the Transformer decoder block.

Formally, let l represent the current layer index and $\mathbf{X}_l \in \mathbb{R}^{N \times C}$ represent the object queries at the l^{th} layer where N is the number of queries and $M_{\text{scene}} \in \mathbb{R}^{N_{\text{retr}} \times C}$ represent the retrieved memory features where N_{retr} is the number of retrieved features. We can then compute $Q_l = f_Q(\mathbf{X}_{l-1})$ and $K_l, V_l \in \mathbb{R}^{N_{\text{retr}} \times C}$ by applying the functions $f_K(\cdot)$ and $f_V(\cdot)$ upon M_{scene}. Here, Q_l represents the query features while K_l and V_l denote the key and value features respectively. The functions f_Q, f_K, and f_V are implemented as standard linear projection layers. Then we perform standard cross-attention as shown in Eq. 1.

$$\mathbf{X}_l = \text{softmax}(\mathbf{Q}_l \mathbf{K}_l^T)\mathbf{V}_l + \mathbf{X}_{l-1} \tag{1}$$

Such a design enables each query to attend to all retrieved features from the memory in an efficient manner. Compared to self-attention, this approach only requires linear runtime complexity and we find that it performs well in multiple settings.

Finally, when retrieving features from both memory banks, we partition the Transformer decoder layers into three splits with three layers in each split. The first split takes advantage of scene-level memory to enhance object-level representations using the context features. Then, the second split uses the object-level memory to focus on fine-grained details and improve the granularity of the object query representations. Finally, the last split uses standard Mask2Former layers without any memory. We do this to limit the computational cost of querying memory and also allow the network to focus on the current input and process the retrieved knowledge.

4 Experiments

We evaluate our approach on several challenging benchmark datasets to showcase the efficacy of our proposed memory pipeline and integration strategy. Further-

Table 3: Panoptic Segmentation results on Cityscapes `val` set with 19 categories.

Method	Backbone	Iters	PQ	PQ^{th}	PQ^{st}	AP^{Th}_{pan}	$mIOU_{pan}$
Panoptic-FPN [23]	R101	65K	58.1	52.0	62.5	33.0	75.7
Panoptic-DeepLab [5]	R50	60K	60.3	-	-	32.1	78.7
Mask2Former [6]	R50	90K	62.1	-	-	37.3	77.5
kMaX-DeepLab [57]	R50	60K	**64.3**	-	-	**38.5**	**79.7**
Mask2Former† [6]	R50	45K	60.5	52.4	66.3	34.0	75.3
MAPS	R50	45K	60.7	52.5	66.6	34.6	77.9
MAPS-Aug	R50	45K	61.0	52.3	**67.3**	34.2	78.1
MAPS-Aug-CLS	R50	45K	61.0	**52.7**	67.0	35.6	79.1

† indicates our own training results

more, we perform detailed ablations across multiple datasets and segmentation tasks.

4.1 Experimental Details

Datasets. We examine the performance of our proposed pipeline using three popular image-segmentation datasets that are suitable for panoptic, semantic, and instance segmentation, namely COCO [29], ADE20K [62], and Cityscapes [9]. The COCO dataset consists of 118,000 images for training and 5,000 images for validation and covers approximately 80 "things" and 53 "stuff" categories. The ADE20K dataset contains images depicting approximately 100 "things" and 50 "stuff" categories over 20,210 training and 2,000 validation images. Finally, the Cityscapes dataset covers 19 total categories (11 stuff and 8 things) and has 2,975 images for training with 500 images for validation and 1,525 images for testing.

Evaluation Metrics. We report the standard evaluation metrics for each of the image segmentation tasks. Specifically, for panoptic segmentation, we use the standard Panoptic Quality (PQ) metric [24]. For instance segmentation, we report the Average precision (AP) metric, and for semantic segmentation, we report mean Intersection-over-Union (mIOU).

4.2 Implementation Details

Here we detail the specific implementation choices we make for building and retrieving from our memory as well as other design choices related to the Mask2Former model. For the Mask2Former model, all experiments are conducted with the ResNet-50 [17] backbone.

We construct each memory bank using the FAISS library [22] and use the Hierarchical Small Navigable Worlds index (HSNW) [34] for fast and efficient querying of the memory. For both memory banks, we retrieve five samples for each instance to balance between retrieving diverse samples and computational cost. For the scene-memory, we apply our cross-attention module for the first three blocks of the transformer decoder, while for the part-memory, we use the cross-attention module for the third to sixth blocks of the decoder. To generate

Table 4: Ablation of extra layers on ADE20K. We demonstrate that our approach improves upon adding additional layers to Mask2Former.

Method	Params	PQ	AP	mIOU
Mask2Former	44M	39.1	26.2	45.3
Mask2Former (6)	45.6	39.8	26.4	46.1
Mask2Former (9)	46.4M	40.6	27.3	46.5
MAPS (6)	45.9M	**40.7**	**27.0**	**47.0**

Table 5: Ablation of augmentation strategies on ADE20K. We demonstrate that adding augmentations to the cross-attention improves model performance.

Augmentation	PQ	AP	mIOU
None	40.7	27.0	47.0
Swapped	39.6	26.9	46.9
Random-Sampling	40.5	27.2	47.5
Input-Masking	41.3	27.4	47.5
Layer-Masking	40.4	27.4	46.6

the features for querying the memory, we use CLIP-ViT-B/16 as our image encoder. The rest of the pipeline follows that of Mask2Former and we refer the reader to [6] for more details about the segmentation architecture and loss hyperparameters and post-processing.

4.3 Training Settings

We train our model using the Detectron2 [51] library and follow Mask2Former's settings for training. Specifically, we use the AdamW optimizer and the same step-learning scheduler as Mask2Former. Further details about the training pipeline can be found in [6]. For all experiments, we train our model using panoptic segmentation labels and follow the respective training recipes for COCO and Cityscapes. For ADE20K, we observe that a simple change to the learning rate schedule enables better training convergence. Specifically, we still our network for 160K iterations, but keep the learning rate the same across training with a 10% drop at the 100K iteration and another 10% drop at the 140K iteration.

4.4 Quantitative Evaluation

We start by benchmarking our approach on the COCO `panoptic val` set and show these results in Table 1. It can be seen that our memory augmented approach outperforms all baselines on all metrics with reduced number of parameters and no special tricks used for the segmentation task. Specifically, we perform 0.3 points better than Mask2Former and 0.5 points more than OneFormer which has been trained jointly on panoptic, instance, and segmentation tasks. Most notably, we achieve the highest mIOU with a 1 point increase compared to baselines. This highlights that our memory is useful in localizing objects and generating more precise segmentation masks.

Table 6: Ablation of Memory Design. Here we delineate all of our memory design choices including two different backbones to generate the CLS token used for indexing the memory as well as forming the scene-level memory as well as four dense extraction strategies to form our object-level memory.

Row	CLIP-CLS	DINO-CLS	CLIP-Image	DINO-Image	MaskCLIP-Image	TiledCLIP-Image	PQ	AP	mIOU
1	✓	✗	✗	✗	✗	✗	**40.3**	**27.2**	**47.0**
2	✗	✓	✗	✗	✗	✗	39.7	26.7	46.0
3	✓	✗	✓	✗	✗	✗	39.6	26.6	45.9
4	✓	✗	✗	✓	✗	✗	40.0	27.0	46.3
5	✓	✗	✗	✗	✓	✗	40.1	26.9	46.2
6	✓	✗	✗	✗	✗	✓	**40.7**	27.0	**47.0**

Table 7: Ablation of each component during inference. Specifically, we turn each component off in the forward pass.

Row	Scene Level	Obj. Level	PQ	AP	mIOU
1	✗	✗	33.3	24.4	40.3
2	✓	✗	40.2	26.7	46.7
3	✗	✓	34.1	25.6	41.9
4	✓	✓	40.7	27.0	47.0

Table 8: Ablation on the number of retrieved samples from memory.

# Retrievals	1	5	10
PQ	39.7	40.7	40.5
AP	26.7	27.0	27.1
mIOU	46.4	47.0	46.9

We next show results for the ADE20K benchmark where we see similar improvements Table 2. Along with our base model using the TiledCLIP strategy, we also experiment with augmentation strategies to the cross-attention to improve model robustness and also further utilize CLIP knowledge by appending the current image's CLIP CLS token to our scene memory. We see that while our plain MAPS framework does improve upon Mask2Former, adding these extra components improves our panoptic quality by 0.4 points and improves mIOU by roughly 1.6 points and outperforming existing state-of-the-art on similar backbones.

Finally we observe improved performance on the Cityscapes dataset compared to existing methods Table 3. While only training for 45K iterations, we observe competitive performance on mIOU where we almost reach state-of-the-art performance with 79.2%. We also closely match Mask2Former and beat its results at that iteration number by 0.5 points. Overall, we observe that the segmentation network equipped with memory is better initialized throughout the training process and consistently has improved performance.

4.5 Ablation Experiments

Next we ablate over the different components of our approach. In Table 4, we showcase that our model outperforms adding extra self-attention layers to the Mask2Former backbone by about 1 point in PQ, 0.6 pts in AP and 1 pt in mIOU. Therefore, we avoid incurring expensive runtime costs as well as additional mem-

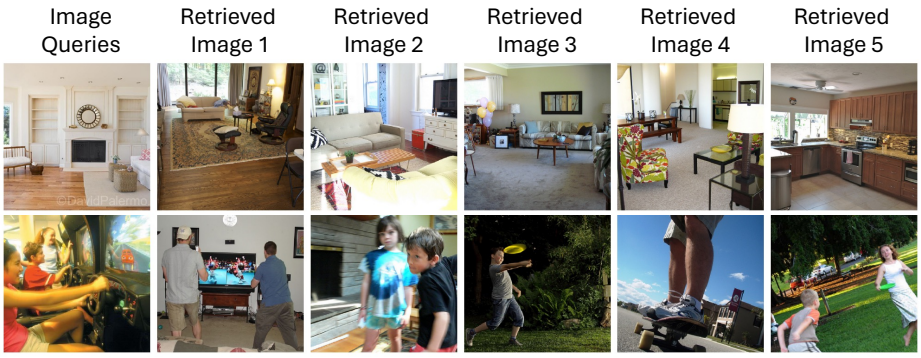

Fig. 3: Examples of image level retrievals from the memory. The first column corresponds to the training image being used to query the memory. The following five rows corresponds to the retrievals from the memory for the corresponding sample.

ory for computing the attention matrices. This demonstrates that our approach benefits from the memory and not the negligible increase in parameters.

Next in Table 6, we examine the effect of different backbones for generating our memory features. Namely, we experiment with CLIP and DINO [3] for generating the CLS tokens that represent the scene memory and are used for indexing as well as MaskCLIP [63] and the base CLIP model for obtaining dense features. We observe that CLIP is best for obtaining global features and that TiledCLIP with CLIP obtains the best PQ and mIOU with 40.7 and 47.0 respectively.

Next, we consider how augmentation strategies affect the cross-attention utilization. Specifically, we experiment with swapping the scene and object memory, adding random sampling to sample new retrievals at each iteration, and random masking of the memory features across the input and layers. The results are shown in Table 5. It can be seen that adding augmentations can add up to 0.6 points in performance boosts for PQ and 0.5 points in mIOU, namely for input-masking. As such, our augmented version relies on random-sampling and input-masking where the cross-attention is randomly masked once at the beginning of the forward pass.

Lastly, we consider the number of samples retrieved Table 8 as well as the individual impact of each memory Table 7. We observe that after 5 samples, there is minimal improvements and that the scene-level memory enables the most benefits in performance, while the object-level memory adds about 1 pt in performance across the board.

4.6 Retrieval Analysis

Here we examine the quality of our retrievals for scene and object-level memory. Some example scene-level retrievals are shown in Fig. 3. For each row, the leftmost image corresponds to the current instance that the model is processing and the right five images are the retrieved samples. We see that the retrieved

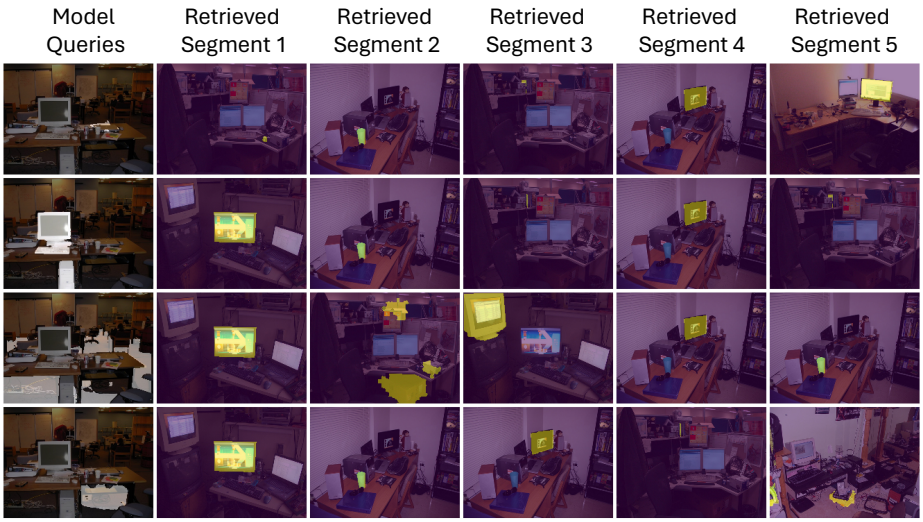

Fig. 4: Examples of retrieved segment queries for different queries in the transformer. The first column corresponds to the region corresponding to where the query looks at in the given image. The following five columns are the regions corresponding to the retrieved images for highest similarity region retrievals.

samples very closely match the queried image with respect to foreground and background. We show object-level retrievals in Fig. 4 shows retrieved segments for a given image query. It can be seen that for query images corresponding to certain objects like computer it looks at other computer screens. At the same time to predict a certain segmentation, it looks at context which might correspond to other objects.

5 Conclusion

In this work, we propose MAPS a novel framework for Memory Augmented Panoptic Segmentation. We present all aspects of our approach and justify different design decisions regarding memory construction, memory incorporation and training. Our memory design allows for the incorporation of associations between context as well as diverse instances coming from the retrieved samples. We equip the state-of-the-art Mask2Former architecture with our memory and delineate an effective yet efficient way to transfer information from the memory features to the segmentation query features. Through extensive experiments and ablations we show improvements over existing baselines across multiple datasets. Our results demonstrate strong improvements as a result of our method, especially in the low-data regime, and show promise in generalizing to more fine-grained tasks as well. While applied to panoptic segmentation we feel our framework is general and can be extended to other tasks.

References

1. Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6154–6162 (2018)
2. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
3. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
4. Chen, K., Pang, J., Wang, J., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Shi, J., Ouyang, W., et al.: Hybrid task cascade for instance segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4974–4983 (2019)
5. Cheng, B., Collins, M.D., Zhu, Y., Liu, T., Huang, T.S., Adam, H., Chen, L.C.: Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12475–12485 (2020)
6. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1290–1299 (2022)
7. Cheng, B., Schwing, A., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation. Advances in Neural Information Processing Systems **34**, 17864–17875 (2021)
8. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1251–1258 (2017)
9. 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: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
11. Fenske, M.J., Aminoff, E., Gronau, N., Bar, M.: Top-down facilitation of visual object recognition: object-based and context-based contributions. Progress in brain research **155**, 3–21 (2006)
12. Fu, J., Liu, J., Wang, Y., Li, Y., Bao, Y., Tang, J., Lu, H.: Adaptive context network for scene parsing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6748–6757 (2019)
13. Gazzaley, A., Nobre, A.C.: Top-down modulation: bridging selective attention and working memory. Trends in cognitive sciences **16**(2), 129–135 (2012)
14. Greene, M.R.: Statistics of high-level scene context. Frontiers in psychology **4**, 777 (2013)
15. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)

16. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
18. Hollingworth, A.: Does consistent scene context facilitate object perception? Journal of experimental psychology: General **127**(4), 398 (1998)
19. Iscen, A., Caron, M., Fathi, A., Schmid, C.: Retrieval-enhanced contrastive vision-text models. arXiv preprint arXiv:2306.07196 (2023)
20. Iscen, A., Fathi, A., Schmid, C.: Improving image recognition by retrieving from web-scale image-text data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19295–19304 (2023)
21. Jain, J., Li, J., Chiu, M.T., Hassani, A., Orlov, N., Shi, H.: Oneformer: One transformer to rule universal image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2989–2998 (2023)
22. Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with GPUs. IEEE Transactions on Big Data **7**(3), 535–547 (2019)
23. Kirillov, A., Girshick, R., He, K., Dollar, P.: Panoptic feature pyramid networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
24. Kirillov, A., He, K., Girshick, R., Rother, C., Dollár, P.: Panoptic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9404–9413 (2019)
25. Kirillov, A., Levinkov, E., Andres, B., Savchynskyy, B., Rother, C.: Instancecut: from edges to instances with multicut. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5008–5017 (2017)
26. Le-Hoa Võ, M., Wolfe, J.M.: The role of memory for visual search in scenes. Annals of the New York Academy of Sciences **1339**(1), 72–81 (2015)
27. Li, Y., Mao, H., Girshick, R., He, K.: Exploring plain vision transformer backbones for object detection. In: European Conference on Computer Vision. pp. 280–296. Springer (2022)
28. Li, Z., Wang, W., Xie, E., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P., Lu, T.: Panoptic segformer: Delving deeper into panoptic segmentation with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1280–1289 (June 2022)
29. Lin, T.Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., Dollár, P.: Microsoft coco: Common objects in context (2015)
30. Liu, H., Son, K., Yang, J., Liu, C., Gao, J., Lee, Y.J., Li, C.: Learning customized visual models with retrieval-augmented knowledge. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15148–15158 (2023)
31. Liu, X., Sikarwar, A., Kreiman, G., Shi, Z., Zhang, M.: Reason from context with self-supervised learning (2023)
32. Long, A., Yin, W., Ajanthan, T., Nguyen, V., Purkait, P., Garg, R., Blair, A., Shen, C., van den Hengel, A.: Retrieval augmented classification for long-tail visual recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6959–6969 (2022)
33. Malisiewicz, T., Efros, A.: Beyond categories: The visual memex model for reasoning about object relationships. Advances in neural information processing systems **22** (2009)

34. Malkov, Y.A., Yashunin, D.A.: Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**(4), 824–836 (2020). https://doi.org/10.1109/TPAMI.2018.2889473
35. Misra, I., Gupta, A., Hebert, M.: From red wine to red tomato: Composition with context. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1792–1801 (2017)
36. Murphy, K.P., Torralba, A., Freeman, W.: Using the forest to see the trees: A graphical model relating features, objects, and scenes. Advances in neural information processing systems **16** (2003)
37. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2021)
38. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)
39. Riesenhuber, M., Poggio, T.: Hierarchical models of object recognition in cortex. Nature neuroscience **2**(11), 1019–1025 (1999)
40. Rosenfeld, A., Biparva, M., Tsotsos, J.K.: Priming neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 2011–2020 (2018)
41. Sam, D.B., Babu, R.V.: Top-down feedback for crowd counting convolutional neural network. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
42. Shen, S., Li, C., Hu, X., Xie, Y., Yang, J., Zhang, P., Gan, Z., Wang, L., Yuan, L., Liu, C., et al.: K-lite: Learning transferable visual models with external knowledge. Advances in Neural Information Processing Systems **35**, 15558–15573 (2022)
43. Shrivastava, A., Gupta, A.: Contextual priming and feedback for faster r-cnn. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. pp. 330–348. Springer (2016)
44. Tian, Z., Shen, C., Chen, H.: Conditional convolutions for instance segmentation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16. pp. 282–298. Springer (2020)
45. Torralba, A.: Contextual priming for object detection. International journal of computer vision **53**, 169–191 (2003)
46. Vo, M.L.H., Boettcher, S.E., Draschkow, D.: Reading scenes: How scene grammar guides attention and aids perception in real-world environments. Current opinion in psychology **29**, 205–210 (2019)
47. Wallingford, M., Ramanujan, V., Fang, A., Kusupati, A., Mottaghi, R., Kembhavi, A., Schmidt, L., Farhadi, A.: Neural priming for sample-efficient adaptation. arXiv preprint arXiv:2306.10191 (2023)
48. Walmer, M., Suri, S., Gupta, K., Shrivastava, A.: Teaching matters: Investigating the role of supervision in vision transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7486–7496 (June 2023)
49. Wang, S., Xu, Y., Fang, Y., Liu, Y., Sun, S., Xu, R., Zhu, C., Zeng, M.: Training data is more valuable than you think: A simple and effective method by retrieving from training data. arXiv preprint arXiv:2203.08773 (2022)

50. Wang, X., Zhang, R., Kong, T., Li, L., Shen, C.: Solov2: Dynamic and fast instance segmentation. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 17721–17732. Curran Associates, Inc. (2020), https://proceedings.neurips.cc/paper_files/paper/2020/file/cd3afef9b8b89558cd56638c3631868a-Paper.pdf
51. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2. https://github.com/facebookresearch/detectron2 (2019)
52. Xie, C.W., Sun, S., Xiong, X., Zheng, Y., Zhao, D., Zhou, J.: Ra-clip: Retrieval augmented contrastive language-image pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19265–19274 (2023)
53. Yang, Z., Ping, W., Liu, Z., Korthikanti, V., Nie, W., Huang, D.A., Fan, L., Yu, Z., Lan, S., Li, B., et al.: Re-vilm: Retrieval-augmented visual language model for zero and few-shot image captioning. arXiv preprint arXiv:2302.04858 (2023)
54. Yanwei Li, Hengshuang Zhao, X.Q.L.W.Z.L.J.S., Jia, J.: Fully convolutional networks for panoptic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
55. Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., Lipson, H.: Understanding neural networks through deep visualization. arXiv preprint arXiv:1506.06579 (2015)
56. Yu, Q., Wang, H., Kim, D., Qiao, S., Collins, M., Zhu, Y., Adam, H., Yuille, A., Chen, L.C.: Cmt-deeplab: Clustering mask transformers for panoptic segmentation. In: CVPR (2022)
57. Yu, Q., Wang, H., Qiao, S., Collins, M., Zhu, Y., Adam, H., Yuille, A., Chen, L.C.: k-means Mask Transformer. In: ECCV (2022)
58. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13. pp. 818–833. Springer (2014)
59. Zhang, M., Tseng, C., Kreiman, G.: Putting visual object recognition in context. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12985–12994 (2020)
60. Zhang, W., Pang, J., Chen, K., Loy, C.C.: K-net: Towards unified image segmentation. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems. vol. 34, pp. 10326–10338. Curran Associates, Inc. (2021), https://proceedings.neurips.cc/paper_files/paper/2021/file/55a7cf9c71f1c9c495413f934dd1a158-Paper.pdf
61. Zheng, Y., Duan, Y., Lu, J., Zhou, J., Tian, Q.: Hyperdet3d: Learning a scene-conditioned 3d object detector (2022)
62. Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ade20k dataset. International Journal of Computer Vision **127**, 302–321 (2019)
63. Zhou, C., Loy, C.C., Dai, B.: Extract free dense labels from clip. In: European Conference on Computer Vision. pp. 696–712. Springer (2022)
64. 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)