

# ViT-Split: Unleashing the Power of Vision Foundation Models via Efficient Splitting Heads

Yifan Li<sup>1\*</sup>; Xin Li<sup>2</sup>, Tianqin Li<sup>2,3</sup>, Wenbin He<sup>2</sup>, Yu Kong<sup>1</sup>, Liu Ren<sup>2</sup>

<sup>1</sup>Michigan State University

<sup>2</sup>Bosch Research North America & Bosch Center for Artificial Intelligence (BCAI)

<sup>3</sup>Carnegie Mellon University

{liyifall, yukong}@msu.edu, {xin.li9, tianqin.li2, Wenbin.He2, liu.ren}@us.bosch.com

## Abstract

Vision foundation models (VFMs) have demonstrated remarkable performance across a wide range of downstream tasks. While several VFM adapters have shown promising results by leveraging the prior knowledge of VFMs, we identify two inefficiencies in these approaches. First, the interaction between convolutional neural network (CNN) and VFM backbone triggers early layer gradient backpropagation. Second, existing methods require tuning all components, adding complexity. Besides, these adapters alter VFM features, underutilizing the prior knowledge. To tackle these challenges, we propose a new approach called ViT-Split, based on a key observation: **the layers of several VFMs, like DINOv2, can be divided into two components: an extractor for learning low-level features and an adapter for learning task-specific features.** Leveraging this insight, we eliminate the CNN branch and introduce two heads, task head and prior head, to the frozen VFM. The task head is designed to learn task-specific features, mitigating the early gradient propagation issue. The prior head is used to leverage the multi-scale prior features from the frozen VFM, reducing tuning parameters and overfitting. Extensive experiments on various tasks (e.g., segmentation, detection, depth estimation, and visual question answering) validate the effectiveness and efficiency of ViT-Split. Specifically, ViT-Split reduces training time up to 4× while achieving comparable or even better results on ADE20K, compared to other VFM adapters. Codes are available: <https://jackyfl.github.io/vitsplit.github.io/>.

## 1. Introduction

Recent studies reveal that the foundation models have the remarkable ability to acquire *prior knowledge* from large-scale datasets [78], which enhances the performance

\*Work done during intern at Bosch

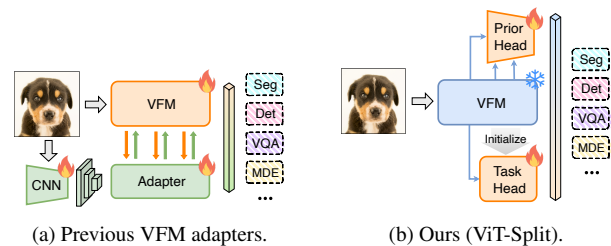


Figure 1. Comparison between previous VFM adapters and ours. Previous VFM adapters integrate low-level features learned by a CNN branch into a learnable VFM through an adapter. Our method exploits VFM prior knowledge with two heads: a prior head for multi-scale prior feature learning from a frozen VFM, and a task head for task-specific feature learning, initialized by the last few layers of the VFM.

in downstream tasks. For vision tasks, vision foundation models (VFMs) acquire prior knowledge from large-scale datasets through self-supervised learning [26], utilizing techniques such as masked image modeling (MIM) [4, 31, 79], contrastive learning [6, 11, 25, 30], or hybrid approaches (MIM + contrastive) [1, 57]. They also leverage vision-language alignment [23, 59] and dense prediction tasks [39, 68], among others. VFMs exhibit remarkable zero-shot and transfer learning capabilities across a variety of downstream tasks, e.g., classification, detection, segmentation, monocular depth estimation (MDE), and visual question answering (VQA), etc.

To leverage prior knowledge from VFMs, previous VFM adapters such as ViT-Adapter [12] or ViT-CoMer [65] primarily adopt a two-branch architecture (see Fig. 1a). Such a design enables the adapter to integrate low-level features from a convolutional neural network (CNN) with global features from a vision transformer (ViT)-based VFM. While this architecture has demonstrated promising results across various downstream tasks, certain design aspects may affect training efficiency. From Fig. 1a, we identify two main is-

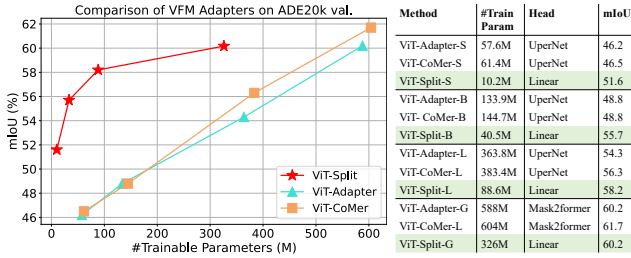


Figure 2. Comparison with previous VFM adapters (ViT-Adapter [12] and ViT-CoMer [65]) on ADE20K val. The results indicate that by leveraging the potential of VFMs (DINOv2 in this task), ViT-Split can achieve competitive results compared to previous VFM adapters. Notably, ViT-Split accomplishes this with only a single linear head and a small number of trainable parameters.

sues of inefficiency. First, the interaction between the CNN and ViT branches across multiple stages requires gradients to be back-propagated through all layers of the model during training. This results in increased computational and memory costs as the size of the VFM grows. Second, all components need to be tuned during training to achieve optimal performance. Specifically, for tasks like segmentation, a large head such as Mask2Former [14] is tuned, and its size is nearly equivalent to that of the VFM backbone.

To address the training inefficiency issue, parameter-efficient fine-tuning (PEFT) methods are proposed to reduce training parameters. These methods include prompt-tuning approaches like VPT [36], adapter-based methods like AdaptFormer [10], and low-rank weight tuning like LoRA [32] or FacT [38]. However, these methods still encounter the issue of early-layer gradient back-propagation, as learnable parameters are appended to each layer’s visual tokens (prompt tuning), or low-rank weights are inserted into the layers (adapter-based methods) or added to the original weights (low-rank weight tuning). Moreover, these PEFT methods do not incorporate low-level features as VFM adapters do, and their performance is either slightly inferior to or generally on par with traditional fine-tuning. Furthermore, despite their proven effectiveness across various tasks [57], the pretrained prior features are not fully leveraged by either PEFT methods or the VFM adapters.

To tackle the aforementioned challenges, we propose a method called *ViT-Split* (see Fig. 1b). ViT-Split is built upon the observation that *the layers of a VFM like DINOv2 [57] can be divided into two components: a low-level feature extractor and a task-specific feature adapter*. Consequently, an additional CNN branch for local feature extraction becomes unnecessary, allowing us to remove it to resolve the early layer gradient propagation issue. Additionally, we propose a task-specific adapter, named “task head”, tailored for downstream tasks. This adapter is initialized from the last few layers of the VFM, further avoid-

ing gradient propagation problems in early layers. To effectively leverage prior features learned by VFM from large-scale datasets, we introduce an additional “prior head” that integrates multi-scale prior features instead of tuning the entire VFM. Such a head reduces the number of trainable parameters and helps mitigate overfitting in the task head (see Appendix). Additionally, we explore two layer selection strategies to identify the most relevant layer features. Experiments on segmentation task (see Fig. 2) demonstrate that our ViT-Split, using only a single linear head, can achieve competitive performance compared with previous VFM adapters with larger segmentation heads like Mask2former [14] or UperNet [66], while tuning fewer parameters and reducing training time (see Fig. 8).

Furthermore, ViT-Split is both *adaptive and memory efficient for multiple tasks* (see Fig. 7). Previous VFM adapters require separate modules (VFM+CNN+adapter+heads) for each task, leading to high computational and memory overhead. In contrast, ViT-Split shares a pre-trained VFM backbone, requiring only a task-specific adapter and the corresponding task head to be learned. Our approach introduces a new paradigm for designing both computation and memory efficient VFM adapters across multiple tasks. In summary, the contributions of this paper are threefold:

- We observe that several VFMs, especially DINOv2, can be divided into two distinct components: an extractor for learning low-level features and an adapter for learning task-specific features.
- We propose an efficient and effective adapter ViT-Split for VFMs. Specifically, ViT-Split introduces two heads, a task head and a prior head. The task head is for learning task-specific features. The prior head is a lightweight CNN for extracting multi-scale prior features from a frozen VFM. We also explore two layer selection methods for selecting prior features from all the layers: uniform sampling and sparse gate.
- We perform extensive experiments and detailed ablations on various downstream tasks to validate the efficiency and effectiveness of our method, including segmentation, detection, MDE, and VQA.

## 2. Related Work

### 2.1. Vision foundation models

Vision foundation models (VFMs) [2] are trained on large-scale datasets in a self-supervised, weakly-supervised, or supervised manner, making them adaptable to a wide range of downstream tasks. Benefiting from the scalability of the transformer architecture, recent ViT-based [22] VFMs demonstrate remarkable zero-shot and transfer ability across various downstream tasks. Self-supervised pre-training paradigm learns discriminative features solely from vision data at the image and pixel level, including con-

trastive learning (MoCo [30], SimCLR [11]), masked image modeling (BEiT [4], MAE [31], iBoT [79]) or hybrid approaches (DINOv2 [57], I-JEPA [1]). Weakly-supervised pretraining paradigm leverages text guidance, aligning visual representations with language space, such as CLIP [59], ALIGN [35], EVA2 [23], SigLip [74], *etc.* Supervised pretraining paradigm learns from different task labels, such as classification (DeiT [62]), segmentation (SAM [39]), and monocular depth estimation (DAM [68]), *etc.*

## 2.2. PEFT and VFM adapters

As the size of transformer-based foundation models continues to grow, such as large language models in language [5, 76], large vision models in vision [21, 70], and multi-modal large language model [3, 13] for multi-modal learning, training efficiency becomes increasingly crucial. To address this challenge, PEFT methods have gained significant popularity in recent years.

Current PEFT approaches for vision [69] generally fall into three categories: prompt tuning, adapter tuning, and parameter tuning. Prompt tuning involves learning a small number of prompt tokens, either in the first layer (CoOp [81], CoCoOp [80]) or in every layer (VPT [36]), making it lightweight and easy to implement. Adapter tuning inserts additional blocks into a frozen model either in a sequential manner (Res-adapt [61], ST-Adapter [58]) or in parallel (AdaptFormer [10], ConvPass [37], LoSA [56]), which shows good adaptability and generalizability. Parameter tuning modifies part of the model parameters, either by adjusting the weight (LoRA [32], FacT [38]) or tuning the bias (Bitfit [73]), resulting in effective and efficient tuning.

Current VFM adapters (ViT-Adapter [12], ViT-CoMer [65]) aim to enhance full fine-tuning performance by incorporating the inductive bias from the CNN branch with spatial prior. These adapters typically require tuning the whole backbone to achieve optimal performance, resulting in better performance than PEFT methods. The interaction between CNN and ViT features is achieved through cross-attention [12], self-attention [65] or mixed [75] across several layers. By contrast, our ViT-Split keeps the entire backbone frozen, introducing two lightweight heads for separate tuning, which is efficient and effective across various tasks.

## 3. Method

### 3.1. The observation in VFMs

We observe that in some VFMs, the layers can be broadly partitioned into two groups with similar features: the earlier and later layers. First, we plot the Centered Kernel Alignment (CKA) [40] across different layers for several VFMs, as shown in Fig. 3. The results reveal that features in the earlier layers are more similar to each other, as are those in the later layers, particularly in DINOv2 [57]. We attribute

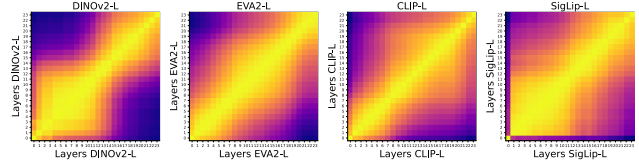


Figure 3. The CKA comparison of layer features across different VFMs, including a self-supervised method DINOv2-L [57], and three image-text alignment methods EVA2-L [23], CLIP-L [59] and SigLip-L [74]. For most of these VFMs, especially DINOv2, the features in the early and later layers show distinct similarities within their respective groups.

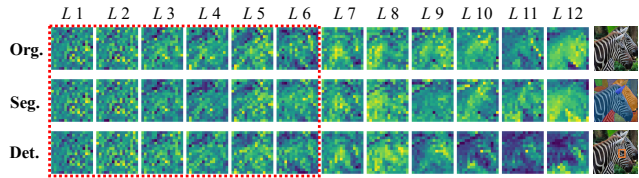


Figure 4. Comparison of DINOv2-S layer features across different tasks, including pretraining (org.), segmentation (seg.), and detection (det.). Notably, the segmentation and detection models are fine-tuned from the DINOv2-S. The features within the red dotted boxes across the three tasks exhibit similar patterns, emphasizing detailed representations. In the later layers, however, the features diverge, becoming more specialized for each task.

this phenomenon to the “encoder-decoder” architecture intrinsic to VFMs: the earlier layers function as an encoder (feature extractor) to capture features from the visual data, while the later layers act as a decoder (task-specific adapter) that generates features for downstream tasks.

A research question is raised: *what do these two groups of layers actually learn?* To answer this question, we visualize the features of each layer in DINOv2-S (Fig. 4) using the first channel of the visual tokens. To further explore feature differences across downstream tasks, we fine-tune the same DINOv2-S on segmentation and detection tasks by adding a linear head and a Mask R-CNN [29] head, respectively. As shown in Fig. 4, we observe that in the early layers (say layer 1-6), all three models exhibit similar feature patterns, focusing more on low-level features like texture and edges. This observation is also supported in [60], which demonstrates that ViT can learn low-level features through large-scale pretraining. While in the later layers, the features diverge for different tasks. Specifically, for the original DINOv2 and segmentation features (row 1 and row 2), the focus shifts towards the semantic information of objects. Whereas in the detection task, the feature attention gradually moves to the object corners or edges (row 3, L7-L12). We attribute this phenomenon to the intrinsic characteristics of each task: DINOv2’s pretraining objective is to reconstruct missing parts of the original features, which

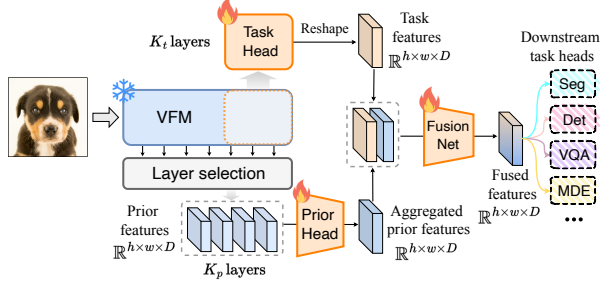


Figure 5. The framework of ViT-Split. ViT-Split introduces two splitting heads, one prior head for aggregating multi-scale prior features from VFM and a task head for learning task-specific features. These features are then combined using a fusion network, enabling effective performance across various downstream tasks.

requires the semantic level understanding as the segmentation task does. In detection, the goal is to predict object bounding boxes, which necessitates focusing more on the corners. This phenomenon also highlights the difference between dense prediction and detection task.

Based on the findings, we divide layers of VFMs into two groups with similar features: a feature extractor for learning low-level features and a task-specific adapter for learning task-related features.

### 3.2. ViT-Split

The framework of ViT-Split is illustrated in Fig. 5, which includes three trainable components: a task head, a prior head and a fusion net. The task head, initialized with the last few layers of the VFM, is designed to learn task-specific features. The prior head integrates multi-scale prior features from the VFM, which are learned from large-scale, diverse datasets. Finally, the fusion net combines both task-specific and prior features to support various downstream tasks.

When an input image with a shape of  $H \times W$  is fed into a frozen VFM (e.g., DINOv2),  $h \cdot w$  vision tokens with  $D$  channels will be obtained from each layer. The vision tokens from  $(L - K_t)$  layer are passed through a task head, which is copied from the last  $K_t$  layers of the VFM, where  $L$  is the number of the total layers. The task features are then reshaped to  $h \times w \times D$ . Meanwhile,  $K_p$  layers of prior features from the frozen VFM are sampled using selection strategies, then concatenated and reshaped into a feature map of size  $h \times w \times (K_p \cdot D)$ . The feature map is then passed through a prior head, a two-layer CNN, resulting in a prior feature map of shape  $h \times w \times D$ . Finally, the task and prior feature maps are concatenated along the channel dimension and fused by a fusion net, which has a similar architecture to the prior head. The final fusion feature map is provided for different downstream heads.

**Task Head.** Based on the observation in Sec. 3.1 that early layers of VFMs are capable of learning low-level features which are similar for different tasks, we avoid fine-

tuning the entire backbone by sharing these early layers. Meanwhile, to retain the prior features of the VFM, we replicate the final  $K_t$  layers separately, utilizing them as a task-specific adapter for downstream tasks. The hyperparameter  $K_t$  controls the adapter’s size, balancing between model capacity and training efficiency.

We observe that the benefits of increasing  $K_t$  diminish, particularly for segmentation tasks, allowing us to choose a smaller  $K_t$  to enhance efficiency (see hyper-parameter analysis in Appendix). Additionally, we find that a large segmentation head may be unnecessary, as the task-specific head is sufficient to capture the downstream dataset’s specific knowledge. Let the features from the  $(L - K_t)$ -th layer of the VFM be denoted as  $f_{L-K_t}$ . Consequently, the task-specific features are given by:

$$f_t = g_{\theta_t}(f_{L-K_t}), \quad (1)$$

where  $g_{\theta_t}$  represents the task head. After obtaining the task feature  $f_t \in \mathbb{R}^{(h \cdot w + 1) \times D}$ , we drop the class token and reshape it from the sequence dimension to form a feature map  $f'_t \in \mathbb{R}^{h \times w \times D}$ .

**Prior Head.** The prior features learned by VFMs have demonstrated strong performance across a range of downstream tasks [57, 59]. However, most current VFM adapters and PEFT methods modify these prior features during training. In contrast, our ViT-Split approach fully leverages the prior knowledge embedded in the multi-scale features of the VFM through a dedicated prior head. Our rationale for utilizing these prior features is to harness the knowledge learned by VFMs to enhance task-specific features while mitigating the risk of overfitting downstream tasks.

Specifically, the architecture of the prior head is shown in Fig. 6, consisting of two CNN layers, a  $1 \times 1$  convolution layer and a  $3 \times 3$  deformable convolution layer. The  $1 \times 1$  convolution layer is used to compress the channels of the multi-scale feature maps, providing efficiency when dealing with larger scales. Meanwhile, the deformable convolution layer [19] enhances low-level features and models geometric transformations within the feature map.

**Layer Selection.** *How to select suitable prior features from all the VFM layers?* To address this, we explore two techniques for selecting  $K_p$  layers from a total of  $L$  layers: uniform sampling and sparse gate. We delineate sparse gate in the Appendix. Uniform sampling involves selecting  $K_p$  prior features uniformly from  $L$  layers. This design is motivated by two factors: first, mitigating the high similarity between features of neighboring layers (see Fig. 3), and second, promoting greater diversity among the selected features. Specifically, the set of sampled indices,  $\mathcal{S}$ , is defined as follows:

$$\delta = \frac{L - b - 1}{K_p - 1}, \mathcal{S} = \{b + \text{round}(i \cdot \delta) | i = 0, \dots, K_p - 1\}, \quad (2)$$

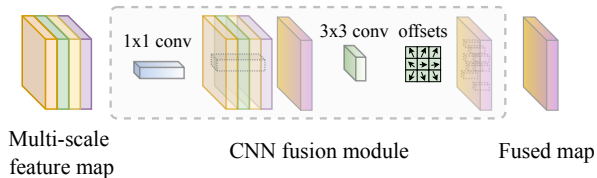


Figure 6. The illustration of the CNN fusion architecture. It is used to fuse multi-scale feature maps and serves as the architecture for both the prior head and fusion net. This module consists of two CNN layers: a  $1 \times 1$  convolution layer followed by a  $3 \times 3$  deformable convolution layer.

where  $b$  is the starting index, used to skip the first few layers, as these layers tend to contain more noise. In most experiments, we set  $b = 2$  or  $b = 3$ .  $\text{round}$  indicates the rounding to the nearest integer, and  $\delta$  represents the sampling interval.

After obtaining the selected prior features  $f_p^i \in \mathbb{R}^{(h \cdot w + 1) \times D}$ ,  $i = \{0, \dots, K_p - 1\}$ , we drop the class tokens, reshape and concatenate them to a multi-scale prior feature map  $f_p \in \mathbb{R}^{h \times w \times (K_p \cdot D)}$ . Finally, the aggregated prior map  $f'_p \in \mathbb{R}^{h \times w \times D}$  can be denoted as:

$$f'_p = g_{\theta_p}(f_p), \quad (3)$$

where  $g_{\theta_p}$  is the prior head.

**Fusion net.** Fusion net is utilized to fuse prior feature map  $f'_p$  and the task-specific feature map  $f'_t$  for different downstream tasks. This network has a similar architecture as the prior head (see Fig. 6). Let  $[f'_p; f'_t] \in \mathbb{R}^{h \times w \times (2D)}$  be the concatenated feature map of  $f'_p$  and  $f'_t$  along the channel dimension. The rationale of using concatenation to fuse two feature maps is to preserve more information (see Tab. 6). The final fused map  $f_o \in \mathbb{R}^{h \times w \times D}$  is given by:

$$f_o = g_{\theta_f}([f'_p; f'_t]), \quad (4)$$

where  $g_{\theta_f}$  is the fusion net.

We then apply different transformations based on the type of downstream task. Specifically, for the segmentation task, we upsample  $f_o$  by a factor of 4 using two transposed convolution layers. For the detection task, we transform  $f_o$  into four scales, *i.e.*,  $4 \times$ ,  $2 \times$ ,  $1 \times$  and  $0.5 \times$  to match the input requirements of the detection head (MaskRCNN). For the VQA task, we reshape  $f_o$  along the sequence dimension to  $(h \cdot w) \times D$  for the LLM decoder.

## 4. Experiments

We conduct experiments on three tasks, semantic segmentation, object detection, and VQA, using well-established benchmarks, *e.g.*, COCO [47], ADE20K [77], CityScapes [18], among others. We also present MDE results in the Appendix. Next, we perform ablation studies to further evaluate ViT-Split’s performance. A uniform selection strategy is applied to all experiments in this section, while results for the sparse gate are provided in the Appendix.

Method	Head	#Train Param	mIoU	Iters
PVT-S [63]	UperNet	54.5M	43.7	160k
Swin-T [51]	UperNet	59.9M	44.5	160k
Twins-SVT-S [16]	UperNet	54.4M	46.2	160k
ViT-S [43]	UperNet	53.6M	44.6	160k
LoSA-S [56]	UperNet	54.9M	45.8	160k
ViT-Adapter-S [12]	UperNet	57.6M	46.2	160k
ViT-CoMer-S [65]	UperNet	61.4M	46.5	160k
DINOv2S <sup>‡</sup> [57]	UperNet	52.2M	50.6	40k
DINOv2-S <sup>‡</sup> [57]	Linear	22.1M	49.6	40k
<b>ViT-Split-S<sup>‡</sup> (ours)</b>	<b>Linear</b>	<b>10.2M</b>	<b>51.6</b>	<b>40k</b>
Swin-B [51]	UperNet	121.0M	48.1	160k
Twins-SVT-L [16]	UperNet	133.0M	48.8	160k
ViT-B [43]	UperNet	127.3M	46.1	160k
LoSA-B [56]	UperNet	131.2M	47.3	160k
ViT-Adapter-B [12]	UperNet	133.9M	48.8	160k
ViT-CoMer-B [65]	UperNet	144.7M	48.8	160k
DINOv2-B <sup>‡</sup> [57]	UperNet	120.7M	54.8	40k
DINOv2-B <sup>‡</sup> [57]	Linear	91.4M	53.8	40k
<b>ViT-Split-B<sup>‡</sup> (ours)</b>	<b>Linear</b>	<b>40.5M</b>	<b>55.7</b>	<b>40k</b>
Swin-L <sup>†</sup> [51]	UperNet	234.0M	52.1	160k
LoSA-L <sup>†</sup> [56]	UperNet	338.5M	53.0	160k
ViT-Adapter-L <sup>†</sup> [12]	UperNet	363.8M	53.4	160k
ViT-CoMer-L <sup>†</sup> [65]	UperNet	383.4M	54.3	160k
DINOv2-L <sup>†</sup> [57]	UperNet	341.2M	57.1	40k
DINOv2-L <sup>†</sup> [57]	Linear	312.9M	56.2	40k
<b>ViT-Split-L<sup>†</sup> (ours)</b>	<b>Linear</b>	<b>88.6M</b>	<b>58.2</b>	<b>40k</b>

Table 1. **Semantic segmentation results on the ADE20K val with  $512 \times 512$  resolution image.** † represents the DINOv2 initialization. ‡ denotes the use of ImageNet-22K pre-trained weight, while the default is to use ImageNet-1K pre-training.

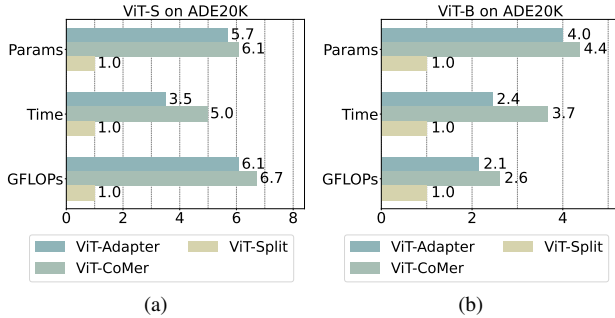
### 4.1. Semantic segmentation

**Settings.** We conduct the semantic segmentation task on ADE20K [77] and Cityscapes [18], using MMSegmentation [17]. We employ AdamW [54] with a learning rate of  $2e-4$  and a weight decay of  $1e-2$ . The training process uses a total batch size of 16. The learning rate for the task head is further reduced by a factor of 0.1. Unlike previous baselines, we use a simple linear head with two-layer deconvolutional blocks ( $\times 4$ ) for segmentation, with a total of 40k iterations (50k for DINOv2-g). We provide the hyperparameter analysis of  $K_p$  and  $K_t$  in the Appendix.

**ADE20K val with  $512 \times 512$  image.** As shown in Tab. 1, we can see that our ViT-Split surpasses all other baselines on ADE20K with  $512 \times 512$  resolution input image by fully leveraging the potential of the VFM. The results demonstrate the superiority of the DINOv2 compared to ImageNet pretrained models. Additionally, ViT-Split requires tuning only about 1/5 to 1/4 of the parameters and trains for just 1/4 of the iterations compared to previous baselines. The parameter efficiency is because of: 1) the efficient adaptation architecture of ViT-Split and 2) the lightweight linear head. The fast convergence speed attributes to effec-

Method	Head	#Train	mIoU (SS/MS)	Pretrain	Extra Pre-train	Iters
ConvNeXt-XL [53]	Mask2former [14]	588M	57.1/58.4	IN-22k	COCO-Stuff	80k
Swin-L [51]	Mask2former [14]	434M	57.3/58.3	IN-22k	COCO-Stuff	80k
SwinV2-G [52]	UperNet [66]	3B	59.3/59.9	IN-22K	Ext-70M	160k
Swin-L [51]	MaskDINO [41]	223M	59.5/60.8	IN-22k	Object365	160k
ViT-CoMer-L [65]	Mask2former [57]	604M	<b>61.7/62.1</b>	MM, BEiTv2	COCO-Stuff	80k
DINOv2-L <sup>†</sup> [57]	Linear	312.9M	58.1/58.5	LVD-142M	-	40k
ViT-Split-L <sup>‡</sup> (ours)	Linear	86.2M	59.0/59.6	LVD-142M	-	40k
ViT-Adapter-G <sup>‡</sup> [12]	Mask2former [57]	588M	-/60.2	LVD-142M	-	80k
ViT-Split-G <sup>‡</sup> (ours)	Linear	326M	60.2/60.8	LVD-142M	-	50k

Table 2. Compared with previous SOTA segmentic segmentation methods on ADE20K val with 896\*896 resolution image. † are initialized with DINOv2. \* is implemented without tuning the whole backbone [57]. “MS” means multi-scale testing. “MM” indicates multi-modal pretraining.



Types	Train Params	Train Time (10,000 iters)	GFLOPs
ViT-Split-S	10.1M	9m25s	129.9
ViT-Split-B	33.2M	17m41s	508.6

Figure 8. Comparison of time complexity for VFM adapters on ADE20K using two different sizes of ViT: (a) ViT-S and (b) ViT-B. For a fair evaluation, we reimplemented the other adapters under the same conditions, i.e., 4xA6000 Ada, over 10,000 iterations.

tive utilization of the prior knowledge embedded in VFMs. Moreover, compared to fine-tuning the entire DINOv2 baseline, our ViT-Split adjusts only 1/4 to 1/2 of the parameters while achieving an average improvement of 2% across three model sizes. Since most tunable parameters come from the tuned head, which represents a small portion of the entire VFM, the overall parameter count for tuning remains low. The performance gains can be attributed to the utilization of the multi-scale prior features from the VFM.

#### ADE20K and Cityscapes val with 896×896 image.

Additionally, we also compare with other SOTA methods on ADE20K (Tab. 2) and Cityscapes (Tab. 3) using images of 896×896 resolution. As shown in Tab. 2, we can see that ViT-Split achieves results comparable to current SOTA methods on ADE20K val. It is worth mentioning that ViT-Split uses only a small linear head and does not rely on extra pretraining data. For a fair comparison, we benchmark against ViT-Adapter-G, which trains only the adapter and the Mask2former head based on the DINOv2 backbone. Our ViT-Split not only delivers better performance

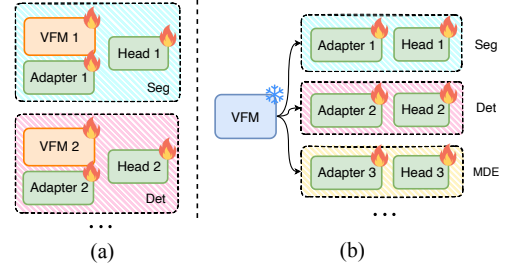


Figure 7. Inference comparison: (a) Previous VFM adapters vs. (b) Our ViT-Split. ViT-Split is efficient during inference for multiple tasks.

Method	Head	#Train Param	mIoU (SS/MS)	Iters
DINOv2-B <sup>‡</sup> [57]	Linear	127.0M	81.2/82.3	20k
ViT-Split-B <sup>‡</sup> (ours)	Linear	55.2M	84.2/85.2	20k
Swin-L [51]	Oneformer [34]	219M	83.0/84.4	90k
ViT-Adapter-L <sup>†</sup> [12]	Mask2former	571M	84.9/85.8	80k
DINOv2-L <sup>†</sup> [57]	Linear	312.9M	83.5/84.3	20k
ViT-Split-L <sup>‡</sup> (ours)	Linear	164.1M	<b>85.8/86.7</b>	20k

Table 3. Semantic segmentation results on Cityscapes val with 896\*896 resolution image. “†” indicates that the model is initialized with BEiTv2 then pretrained on the Mapillary dataset. “‡” represents the use of DINOv2. “SS” denotes single-scale testing, and “MS” means multi-scale testing.

but also requires half the training parameters and achieves faster training speed. Specifically, according to [57], training ViT-Adapter-G requires 16 V100 GPUs for 28 hours, whereas our ViT-Split-G takes only 8 A6000 Ada GPUs for 15.7 hours. Moreover, on Cityscapes dataset (Tab. 3), our ViT-Split outperforms ViT-Adapter with only around 1/6 parameters being tuned. The results suggest that a simple linear head is enough for competitive results on semantic segmentation by fully leveraging VFM prior knowledge.

**Time complexity analysis.** As illustrated in Fig. 8, our ViT-Split achieves, on average, approximately 4× faster training speed for the small model and 3× faster for the base model compared to the other two VFM adapters. The slower training speed of the other adapters can be attributed to two factors: the early gradient backpropagation and the interaction between the CNN branch and the ViT. In contrast, our ViT-Split avoids backpropagating gradients to early layers, and reduces both the CNN branch computations and interaction overhead by fully leveraging the prior knowledge in the VFM. As shown in Fig. 7, traditional VFM adapters require training a task-specific VFM along with its corresponding adapter and head. In contrast, ViT-Split keeps the entire VFM frozen, training only a smaller adapter and the corresponding head. This design significantly reduces computational costs, making it more efficient for supporting multiple downstream tasks during inference.

Method	LLM	Image Size	Sample Size		VQAv2 [24]	VizWiz [27]	LLaVA-Wild [49]	SciQA-IMG [55]	MM-Vet [71]	POPE [45]			MMB [50]
			Pre	Ft						rand	pop	adv	
BLIP-2 [42]	Vicuna-13B	224 <sup>2</sup>	129M	-	65.0	19.6	19.6	61	22.4	<b>89.6</b>	85.5	80.9	-
InstructBLIP [20]	Vicuna-7B	224 <sup>2</sup>	129M	1.2M	-	34.5	34.5	60.5	26.2	-	-	-	36
InstructBLIP [20]	Vicuna-13B	224 <sup>2</sup>	129M	1.2M	-	33.4	33.4	63.1	25.6	87.7	77	72	-
Shikra [8]	Vicuna-13B	224 <sup>2</sup>	600K	5.5M	77.4*	-	-	-	-	-	-	-	58.8
IDEFICS-9B [33]	LLaMA-7B	224 <sup>2</sup>	353M	1M	50.9	35.5	35.5	-	-	-	-	-	48.2
IDEFICS-80B [33]	LLaMA-65B	224 <sup>2</sup>	353M	1M	60.0	36	36.0	-	-	-	-	-	54.5
Qwen-VL [3]	Qwen-7B	448 <sup>2</sup>	1.4B	50M	<b>78.8*</b>	35.2	35.2	67.1	-	-	-	-	38.2
Qwen-VL-Chat [3]	Qwen-7B	448 <sup>2</sup>	1.4B*	50M	78.2*	38.9	38.9	<u>68.2</u>	-	-	-	-	60.6
LLaVA-1.5 [48]	Vicuna-7B	336 <sup>2</sup>	<b>558K</b>	<b>665K</b>	<u>78.5*</u>	<u>50.0*</u>	<u>65.4</u>	<u>66.8</u>	<u>31.1</u>	87.3	<u>86.2</u>	<u>84.2</u>	<u>64.3</u>
LLaVA-1.5 + ViT-Split	Vicuna-7B	336 <sup>2</sup>	<b>558K</b>	<b>665K</b>	<u>78.2</u> <sub>-0.3</sub>	<u>51.7</u> <sub>+1.7</sub>	<u>71.1</u> <sub>+5.7</sub>	<u>70.4</u> <sub>+3.6</sub>	<u>31.2</u> <sub>+0.1</sub>	<u>88.5</u> <sub>+1.2</sub>	<u>87.4</u> <sub>+1.2</sub>	<u>86.1</u> <sub>+1.9</sub>	<u>66.4</u> <sub>+2.1</sub>

Table 4. **Comparison with different VLLM methods on VQA benchmarks.** ViT-Split is integrated into the vision encoder (CLIP-L) of LLaVA-1.5 (7B), tuning the penultimate block and utilizing prior feature from this layer. This adaptation can consistently enhance performance across most benchmarks, demonstrating the effectiveness and generalization of ViT-Split.

## 4.2. Detection and Instance Segmentation

**Settings.** We present detection and instance segmentation results on COCO-2017 [47] in Tab. 5, using MMDetection [7]. The AdamW optimizer is employed with an initial learning rate of 1e-4 and a weight decay of 5e-2, training for 12 epochs (1× schedule). The total batch size is set to 16 and we utilize a MaskRCNN [29] head for experiment. The setting of  $K_p$  and  $K_t$  is given in the Appendix.

As shown in Tab. 5, our ViT-Split achieves comparable performance with current SOTA VFM adapter ViT-CoMer. As discussed in 3.1, the detection task may differ significantly from the original DINOv2 pretraining task, necessitating the tuning of more parameters. Despite this, our ViT-Split still involves *fewer parameters and faster training speed (reducing 42% training time) than ViT-CoMer*, demonstrating the efficiency of our architecture.

## 4.3. Visual Question Answering

**Settings.** We also present VQA results using the popular visual large language model (VLLM) [46], LLaVA-1.5 [48]. This model comprises a CLIP-L visual encoder for encoding images, an MLP connector for projecting visual tokens into the language space, and a Vicuna-based LLM [15] for generating language tokens. In our modified LLaVA, we replace the original MLP projector with our ViT-Split. To comprehensively evaluate the effectiveness of our ViT-Split, we utilize both academic-task-oriented benchmarks (VQAv2 [24], VizWiz [27], SciQA-IMG [55]), and instruction-following LLM benchmarks (POPE [45], MMBench [50], LLaVA-Wild [49], MM-Vet [71]). Following [48], we first pretrain our ViT-Split using 558K image-text pairs, and subsequently fine-tune both ViT-Split and the LLM with 665K mixed data pairs. For more detailed information regarding the hyperparameter settings, please refer to the Appendix.

As shown in Tab. 4, our ViT-Split enhances LLaVA-1.5 performance across most benchmarks. This improvement demonstrates that ViT-Split is also applicable to other VFMs and VQA tasks. Unlike most current VLLMs that di-

Method	#Param	Mask R-CNN 1× schedule					
		AP <sup>b</sup>	AP <sup>b</sup> <sub>50</sub>	AP <sup>b</sup> <sub>75</sub>	AP <sup>m</sup>	AP <sup>m</sup> <sub>50</sub>	AP <sup>b</sup> <sub>75</sub>
ConvNeXt-T [53]	48M	44.2	66.6	48.3	40.1	63.3	42.8
Focal-T [67]	49M	44.8	67.7	49.2	41.0	64.7	44.2
SPANet-S [72]	48M	44.7	65.7	48.8	40.6	62.9	43.8
MixFormer-B4 [9]	53M	45.1	67.1	49.2	41.2	64.3	44.1
Twins-B [16]	76M	45.2	67.6	49.3	41.5	64.5	44.8
Swin-S [51]	69M	44.8	66.6	48.9	40.9	63.4	44.2
Flatten-PVT-T [28]	49M	44.2	67.3	48.5	40.2	63.8	43.0
ViT-S [43]	44M	40.2	63.1	43.4	37.1	60.0	38.8
ViTDet-S [44]	46M	40.6	63.3	43.5	37.1	60.0	38.8
ViT-Adapter-S [12]	48M	44.7	65.8	48.3	39.9	62.5	42.8
ViT-CoMer-S [65]	50M	45.8	67.0	49.8	40.5	63.8	43.3
ViT-CoMer-S <sup>‡</sup> [65]	50M	<b>48.6</b>	<b>70.5</b>	<u>53.1</u>	<b>42.9</b>	67.0	<b>45.8</b>
ViT-Split-S <sup>‡</sup> (ours)	45M	<u>48.5</u>	<b>70.5</b>	<b>53.3</b>	<u>42.8</u>	<b>67.2</b>	<u>45.6</u>
PVTv2-B5 [63]	102M	47.4	68.6	51.9	42.5	65.7	46.0
InternImage-B [64]	115M	48.8	70.9	54.0	44.0	67.8	47.4
ViT-B [43]	114M	42.9	65.7	46.8	39.4	62.6	42.0
ViTDet-B [44]	121M	43.2	65.8	46.9	39.2	62.7	41.4
ViT-Adapter-B [12]	120M	47.0	68.2	51.4	41.8	65.1	44.9
ViT-CoMer-B [65]	129M	47.6	68.9	51.9	41.8	65.9	44.9
ViT-CoMer-B <sup>‡</sup> [65]	129M	<b>52.0</b>	<b>73.6</b>	<b>57.2</b>	<b>45.5</b>	<b>70.6</b>	<b>49.0</b>
ViT-Split-B <sup>‡</sup> (ours)	118M	<u>51.8</u>	<b>73.6</b>	<u>57.1</u>	<u>45.4</u>	<u>70.3</u>	48.6
ViT-L <sup>†</sup> [43]	337M	45.7	68.9	49.4	41.5	65.6	44.6
ViTDet-L <sup>†</sup> [44]	351M	46.2	69.2	50.3	41.4	65.8	44.1
ViT-Adapter-L <sup>†</sup> [12]	348M	48.7	70.1	53.2	43.3	67.0	46.9
ViT-CoMer-L <sup>†</sup> [65]	363M	51.4	73.5	55.7	45.2	70.3	48.5
ViT-CoMer-L <sup>†</sup> [65]	363M	<b>53.4</b>	<b>75.3</b>	<b>58.9</b>	<b>46.8</b>	<b>72.0</b>	<b>50.9</b>
ViT-Split-L <sup>†</sup> (ours)	348M	<u>53.0</u>	<u>75.1</u>	<u>58.1</u>	<u>46.6</u>	<u>71.9</u>	<u>50.4</u>

Table 5. **Object detection and instance segmentation using Mask R-CNN on COCO val2017.** “†” indicates pre-training with ImageNet-22K, “‡” represents the use of DINOv2 [57], while the default setting uses ImageNet-1K pre-training.

rectly utilize features from the penultimate layer, ViT-Split leverages both the prior features of the vision encoder and the task-specific features, resulting in richer visual representations that improve the LLM’s learning process. Moreover, we tune only a small portion of the vision encoder’s parameters (specifically, one layer), which ensures efficiency for both training and inference. We believe that ViT-Split will offer new inspiration for VLLM design.

Components			Train Params		mIoU	
prior $g_{\theta_p}$	task $g_{\theta_t}$	fusion $g_{\theta_f}$	Small	Base	Small	Base
			22.1M	91.4M	49.6	53.8
			1.2M	4.84M	44.3	47.8
✓			3.2M	12.6M	46.0	51.4
	✓		6.6M	26.1M	49.5	53.2
✓	✓		8.6M	33.9M	50.4	54.6
✓	✓	✓	10.2M	40.5M	51.6	55.7

Table 6. Ablation study of the prior head ( $g_{\theta_p}$ ), task head ( $g_{\theta_t}$ ), and fusion net ( $g_{\theta_f}$ ) on ADE20K, conducted with two ViT sizes: small and base on ViT-Split<sub>u</sub>. We set  $K_t = 3$  and  $K_p = 4$  for both model sizes. The baseline model (no modules used, shown without background color) uses only the frozen features from the last layer. The baseline with a gray background indicates full fine-tuning of the entire backbone. When only  $g_{\theta_p}$  and  $g_{\theta_t}$  are used, their features are combined via addition.

#### 4.4. Ablation Study

We conduct an ablation study for each trainable component in Tab. 6 on ADE20K. The default settings are consistent with those described in Sec. 4.1.

**The effectiveness of prior head.** The results in Tab. 6 show that incorporating the prior head improves performance by 2.7% and 3.6% compared to the baseline that uses only the final-layer features. This suggests that the prior head effectively leverages multi-layer prior features from the VFM to enhance overall representation quality, surpassing the use of solely the final layer’s prior features. Additionally, our module enhances 2D local representations through the use of a CNN. Furthermore, the results demonstrate that the prior features extracted from the original VFM are highly valuable, achieving performance levels nearly equivalent to those obtained through full fine-tuning.

**The effectiveness of task head.** As shown in Tab. 6, by tuning only the task head  $g_{\theta_t}$ , the performance nearly matches that of fine-tuning the entire model, supporting the finding in Sec. 3.1. Last few layers can learn task-specific features and achieve similar performance as tuning the entire backbone. Furthermore, the experiments demonstrate that performance can be further enhanced when combined with prior features. We attribute this improvement to the combined benefits of task-specific and prior knowledge, with the latter helping to reduce task head overfitting.

**The effectiveness of fusion head.** Tab. 6 shows that using fusion net  $g_{\theta_f}$  yields a performance improvement of 1.1% for two ViT sizes. We attribute this enhancement to our CNN-based fusion module, which retains richer feature information compared to a simple addition operation. Again, the CNN component strengthens the local feature representation, contributing to improved fusion results.

**The effectiveness of uniform layer selection.** In Tab. 7,

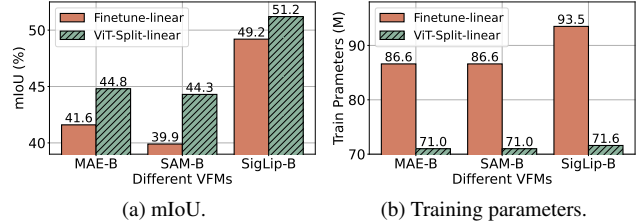


Figure 9. Segmentation results and parameters on ADE20K with different VFMs, including MAE-B [31], SAM-B [39] and SigLip-B [74]. We set  $K_p = 4$  and  $K_t = 8$  for all the VFMs.

we evaluate the effectiveness of the selection strategy for prior features. Compared to selecting features from only the last few layers, which capture mostly task-specific prior information—uniform selection allows for a more diverse set of prior features, encompassing both low-level and task-specific characteristics. This uniform selection approach becomes increasingly impactful as the backbone size grows.

**The effectiveness across different VFMs.** To evaluate the generality of our ViT-Split, we present results on various VFMs in Fig. 9, leveraging the excellent VFM-benchmark codebase<sup>1</sup>. The experiments demonstrate that ViT-Split consistently enhances performance across both weakly-supervised VFMs (SAM and SigLip) and self-supervised VFMs (MAE). These results not only validate the effectiveness of ViT-Split on multiple VFMs but also suggest that our observations may hold for a broader range of VFMs.

## 5. Conclusion

In this paper, we introduce ViT-Split, an efficient, effective, and generalized adapter, to adapt VFMs for downstream tasks. Specifically, we introduce two heads based on a frozen VFM, a prior head for multi-scale prior feature extraction and a task head for task-specific feature adaptation. Experiments on segmentation, detection, MDE, and VQA verify the effectiveness and efficiency of our method. In the future, we aim to apply ViT-Split to more VFMs and tasks. We hope our method offers a fresh perspective for efficient and effective VFM adapter design.

## Acknowledgement

Yifan Li and Yu Kong are partially supported by NSF ReD-DDoT award 2429836.

<sup>1</sup><https://github.com/tue-mps/benchmark-vfm-ss>

## References

- [1] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *CVPR*, pages 15619–15629, 2023. 1, 3
- [2] Muhammad Awais, Muzammal Naseer, Salman Khan, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Foundational models defining a new era in vision: A survey and outlook. *arXiv preprint arXiv:2307.13721*, 2023. 2
- [3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 3, 7
- [4] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. In *ICLR*, 2022. 1, 3
- [5] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 3
- [6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9650–9660, 2021. 1
- [7] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 7
- [8] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm’s referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 7
- [9] Qiang Chen, Qiman Wu, Jian Wang, Qinghao Hu, Tao Hu, Errui Ding, Jian Cheng, and Jingdong Wang. Mixformer: Mixing features across windows and dimensions. In *CVPR*, pages 5249–5259, 2022. 7
- [10] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. In *NeurIPS*, pages 16664–16678, 2022. 2, 3
- [11] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607, 2020. 1, 3
- [12] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. In *ICLR*, 2023. 1, 2, 3, 5, 6, 7
- [13] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, pages 24185–24198, 2024. 3
- [14] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, pages 1290–1299, 2022. 2, 6
- [15] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%\* chatgpt quality. See <https://vicuna.lmsys.org> (accessed 14 April 2023), 2(3):6, 2023. 7
- [16] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the design of spatial attention in vision transformers. In *NeurIPS*, pages 9355–9366, 2021. 5, 7
- [17] MMSegmentation Contributors. Mmsegmentation: Openmmlab semantic segmentation toolbox and benchmark, 2020. 5
- [18] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, pages 3213–3223, 2016. 5
- [19] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, pages 764–773, 2017. 4
- [20] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, 2023. 7
- [21] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *ICML*, pages 7480–7512, 2023. 3
- [22] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. 2
- [23] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *CVPR*, pages 19358–19369, 2023. 1, 3
- [24] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, pages 6904–6913, 2017. 7
- [25] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent—a new approach to self-supervised learning. In *NeurIPS*, pages 21271–21284, 2020. 1
- [26] Jie Gui, Tuo Chen, Jing Zhang, Qiong Cao, Zhenan Sun, Hao Luo, and Dacheng Tao. A survey on self-supervised learning: Algorithms, applications, and future trends. *IEEE TPAMI*, 2024. 1
- [27] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham.

- Vizwiz grand challenge: Answering visual questions from blind people. In *CVPR*, pages 3608–3617, 2018. 7
- [28] Dongchen Han, Xuran Pan, Yizeng Han, Shiji Song, and Gao Huang. Flatten transformer: Vision transformer using focused linear attention. In *ICCV*, 2023. 7
- [29] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017. 3, 7
- [30] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020. 1, 3
- [31] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022. 1, 3, 8
- [32] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022. 2, 3
- [33] IDEFICS. Introducing idefics: An open reproduction of state-of-the-art visual language model. <https://huggingface.co/blog/idefics>, 2023. 7
- [34] Jitesh Jain, Jiachen Li, Mang Tik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. Oneformer: One transformer to rule universal image segmentation. In *CVPR*, pages 2989–2998, 2023. 6
- [35] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pages 4904–4916, 2021. 3
- [36] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, pages 709–727, 2022. 2, 3
- [37] Shibo Jie and Zhi-Hong Deng. Convolutional bypasses are better vision transformer adapters. *arXiv preprint arXiv:2207.07039*, 2022. 3
- [38] Shibo Jie and Zhi-Hong Deng. Fact: Factor-tuning for lightweight adaptation on vision transformer. In *AAAI*, pages 1060–1068, 2023. 2, 3
- [39] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pages 4015–4026, 2023. 1, 3, 8
- [40] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *ICML*, pages 3519–3529, 2019. 3
- [41] Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M Ni, and Heung-Yeung Shum. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In *CVPR*, pages 3041–3050, 2023. 6
- [42] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, pages 19730–19742, 2023. 7
- [43] Yanghao Li, Saining Xie, Xinlei Chen, Piotr Dollár, Kaiming He, and Ross Girshick. Benchmarking detection transfer learning with vision transformers. *arXiv preprint arXiv:2111.11429*, 2021. 5, 7
- [44] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *ECCV*, pages 280–296, 2022. 7
- [45] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 7
- [46] Yifan Li, Zhixin Lai, Wentao Bao, Zhen Tan, Anh Dao, Kewei Sui, Jiayi Shen, Dong Liu, Huan Liu, and Yu Kong. Visual large language models for generalized and specialized applications. *arXiv preprint arXiv:2501.02765*, 2025. 7
- [47] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014. 5, 7
- [48] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, pages 26296–26306, 2024. 7
- [49] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2024. 7
- [50] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *ECCV*, pages 216–233, 2024. 7
- [51] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 5, 6, 7
- [52] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *CVPR*, pages 12009–12019, 2022. 6
- [53] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, pages 11976–11986, 2022. 6, 7
- [54] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5
- [55] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *NeurIPS*, 2022. 7
- [56] Otniel-Bogdan Mercea, Alexey Gritsenko, Cordelia Schmid, and Anurag Arnab. Time-memory-and parameter-efficient visual adaptation. In *CVPR*, pages 5536–5545, 2024. 3, 5
- [57] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2023. 1, 2, 3, 4, 5, 6, 7
- [58] Junting Pan, Ziyi Lin, Xiatian Zhu, Jing Shao, and Hongsheng Li. St-adapter: Parameter-efficient image-to-video transfer learning. In *NeurIPS*, pages 26462–26477, 2022. 3
- [59] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,

- Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 1, 3, 4
- [60] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *NeurIPS*, 34:12116–12128, 2021. 3
- [61] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Learning multiple visual domains with residual adapters. In *NeurIPS*, 2017. 3
- [62] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, pages 10347–10357, 2021. 3
- [63] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022. 5, 7
- [64] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In *CVPR*, pages 14408–14419, 2023. 7
- [65] Chunlong Xia, Xinliang Wang, Feng Lv, Xin Hao, and Yifeng Shi. Vit-comer: Vision transformer with convolutional multi-scale feature interaction for dense predictions. In *CVPR*, pages 5493–5502, 2024. 1, 2, 3, 5, 6, 7
- [66] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *ECCV*, pages 418–434, 2018. 2, 6
- [67] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, and Jianfeng Gao. Focal self-attention for local-global interactions in vision transformers. *arXiv preprint arXiv:2107.00641*, 2021. 7
- [68] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, pages 10371–10381, 2024. 1, 3
- [69] Bruce XB Yu, Jianlong Chang, Haixin Wang, Lingbo Liu, Shijie Wang, Zhiyu Wang, Junfan Lin, Lingxi Xie, Haojie Li, Zhouchen Lin, et al. Visual tuning. *ACM Computing Surveys*, 56(12):1–38, 2024. 3
- [70] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, 2022. 3
- [71] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 7
- [72] Guhnoo Yun, Juhan Yoo, Kijung Kim, Jeongho Lee, and Dong Hwan Kim. Spanet: Frequency-balancing token mixer using spectral pooling aggregation modulation. In *ICCV*, pages 6113–6124, 2023. 7
- [73] Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *ACL*, pages 1–9, 2022. 3
- [74] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, pages 11975–11986, 2023. 3, 8
- [75] Dong Zhang, Rui Yan, Pingcheng Dong, and Kwang-Ting Cheng. Memory efficient transformer adapter for dense predictions. In *ICLR*, 2025. 3
- [76] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023. 3
- [77] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *IJCV*, 127:302–321, 2019. 5
- [78] Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, et al. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *arXiv preprint arXiv:2302.09419*, 2023. 1
- [79] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. Image bert pre-training with online tokenizer. In *ICLR*, 2022. 1, 3
- [80] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, pages 16816–16825, 2022. 3
- [81] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 130(9):2337–2348, 2022. 3