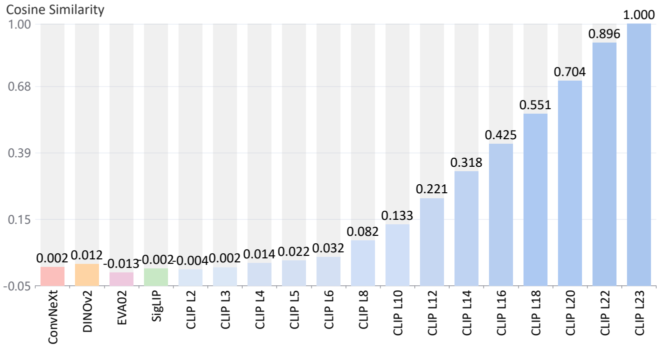
arXiv:2410.11829v1 [cs.CV] 15 Oct 2024

MMFuser: Multimodal Multi-Layer Feature Fuser for Fine-Grained Vision-Language Understanding

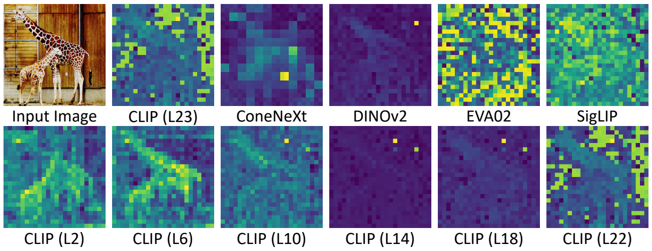
Yue Cao, Yangzhou Liu, Zhe Chen, Guangchen Shi, Wenhai Wang, Danhuai Zhao, and Tong Lu

***Abstract*—Despite significant advancements in Multimodal Large Language Models (MLLMs) for understanding complex human intentions through cross-modal interactions, capturing intricate image details remains challenging. Previous methods integrating multiple vision encoders to enhance visual detail introduce redundancy and computational overhead. We observe that most MLLMs utilize only the last-layer feature map of the vision encoder for visual representation, neglecting the rich fine-grained information in shallow feature maps. To address this issue, we propose MMFuser, a simple yet effective multi-layer feature fuser that efficiently integrates deep and shallow features from Vision Transformers (ViTs). Specifically, it leverages semantically aligned deep features as queries to dynamically extract missing details from shallow features, thus preserving semantic alignment while enriching the representation with fine-grained information. Applied to the LLaVA-1.5 model, MMFuser achieves significant improvements in visual representation and benchmark performance, providing a more flexible and lightweight solution compared to multi-encoder ensemble methods. The code and model have been released at https://github.com/yuecao0119/MMFuser.**

***Index Terms*—Multimodal large language model, visual perception, feature fusion, Transformer.**



(a) Cosine similarity between different feature maps.



(b) Visualization of different feature maps.

Fig. 1. Comparison of feature maps from different vision encoders and various layers of CLIP-ViT. (a) Cosine similarity is computed between the feature maps from various vision encoders, including CLIP-ViT-L [17], ConvNeXt-XXL [18], DINOv2-L [19], EVA02-L [20], and SigLIP-L [21], and the final-layer feature map of CLIP-ViT-L. (b) Visualization of different feature maps. These results indicate significant feature differences not only between various vision encoders but also across different layers within the same vision encoder. This observation motivates us to fully explore the potential of individual vision encoders for developing MLLMs.

I. INTRODUCTION

IN recent years, Multimodal Large Language Models (MLLMs) [1]–[7] have emerged as a research hotspot in the field of Artificial General Intelligence (AGI). These models have made significant strides in understanding and expressing complex human intent through cross-modal interaction and learning. Building on rapid advancements in Large Language Models (LLMs) [8]–[16], MLLMs utilize pre-trained vision encoders to extract image features and integrate them with advanced LLMs, demonstrating remarkable capabilities across a wide range of vision-language tasks.

Currently, the mainstream approach [2]–[5], [22]–[24] in the community involves using a pre-trained Vision Transformer (ViT) [17], [25] or its variants [18], [19], [26] as the vision encoder, feeding the outputs from its final or penultimate layer into LLMs as visual representations. In this manner, these features with rich high-level semantic information are effectively transformed from an image space to a semantic text space. However, due to the loss of low-level image information in deep features, current MLLMs encounter challenges in accurately interpreting details, resulting in issues such as Optical Character Recognition (OCR) errors and object hallucinations.

To address these issues, recent studies [24], [27] indicate that the ability of vision encoders to learn fine-grained image features has become a bottleneck for MLLMs. Consequently, some researchers [15], [28]–[30] believe that solely relying on features of a single vision encoder may not be optimal. They propose integrating multiple pre-trained vision encoders, such as CLIP [17], DINOv2 [19], and ConvNext [18], to enhance fine-grained visual representations in a complementary manner. While these ensemble-based methods achieve promising

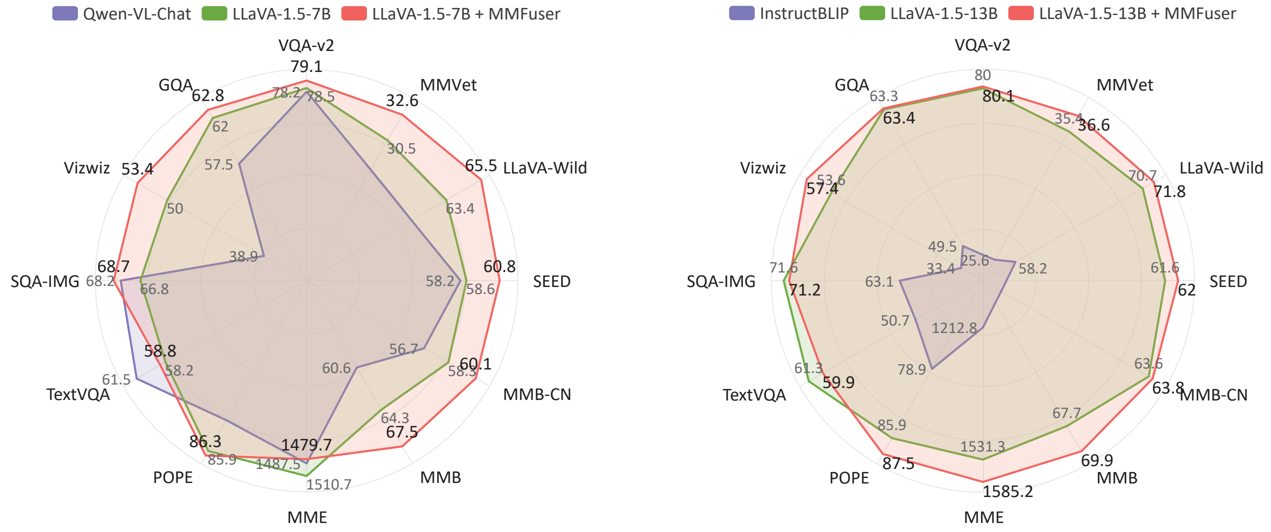
---

This work was supported in part by the China Mobile Zijin Innovation Insititute under Grant NR2310J7M, in part by the National Natural Science Foundation of China under Grant 62372223, and in part by the Youth Student Basic Research Project of the National Natural Science Foundation for PhD Students under Grant 623B2050. *(Yue Cao and Yangzhou Liu are co-first authors.) (Corresponding author: Tong Lu.)*

Yue Cao, Yangzhou Liu, Zhe Chen, Guangchen Shi, and Tong Lu are with the State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China, (e-mail: caoyue0119@gmail.com; lyzlll2343@gmail.com; czcz94cz@gmail.com; guangchenshi@smail.nju.edu.cn; lutong@nju.edu.cn).

Wenhai Wang is with the Chinese University of Hong Kong, Hong Kong 999077, China, (e-mail: wangwenhai362@163.com).

Danhuai Zhao is with the China Mobile Zijin Innovation Insititute, Nanjing 211899, China, (e-mail: zhaodanhuai@chinamobile.com).



(a) Comparison of 7B models

(b) Comparison of 13B models

Fig. 2. Performance comparison across different model sizes. (a) Among 7B models, including Qwen-VL-Chat [5], LLaVA-1.5-7B [3], our model surpasses LLaVA-1.5-7B on 11 out of 12 benchmarks, with an average score of 61.8 compared to LLaVA-1.5-7B's 60.3. (b) Among 13B models, including InstructBLIP [1] and LLaVA-1.5-13B [3], our model also outperforms LLaVA-1.5-13B on 10 out of 12 benchmarks, achieving an average score of 64.1 compared to LLaVA-1.5-13B's 63.2. These results indicate that MMFuser can effectively improve the performance of LLaVA-1.5 models.

results, they unfortunately introduce model redundancy and increase computational overhead. Therefore, the necessity of employing multiple vision encoders remains contentious.

As a matter of fact, even with a single vision encoder, learned visual representations are diverse, as shown in Fig. 1. The visualizations indicate that deep features are effective at extracting high-level semantic information, whereas shallow features are better suited for capturing low-level details such as edges and textures, which have not been fully leveraged in current MLLMs. Looking back at classic image and video tasks such as object detection and semantic segmentation, multi-layer features are widely used [31]–[35], where the combination of shallow and deep visual features provides a more comprehensive understanding of images or videos. However, the application of multi-layer features in this manner remains relatively uncommon within the field of MLLMs.

It is natural to extend this idea to the visual representation of MLLMs. We have experimented with some straightforward methods to combine multiple layers of features from a single ViT [17] for enhancing image detail, such as element-wise averaging or channel-wise concatenation. However, these simple fusion methods have a little improvement in performance. Through further exploration, we find that while shallow features capture richer low-level details, their semantic alignment with the text feature space is insufficient. This deficiency impairs the model's ability to understand images and offsets the benefits introduced by combining shallow and deep visual features, indicating that multi-layer feature fusion in MLLMs is vital and requires more thoughtful designs.

Further, previous work [36] suggests that LLMs excel in understanding deep features of ViTs, which are fully aligned with text feature spaces. Conversely, while shallow features are rich in details, they exhibit poor semantic alignment, making it challenging for LLMs to effectively interpret these features. This insight inspired us to propose a simple yet effective method called **MMFuser** (see Fig. 3), which uses deep features as queries to dynamically extract missing details from shallow features. It minimizes the risk of shallow features disrupting semantic alignment, maintaining the coherence of deep features while enriching them with fine-grained information. By leveraging multi-layer features, MMFuser can enhance the overall performance of MLLM in processing images and videos.

To validate the effectiveness of MMFuser, we applied it to the recent well-known model, LLaVA-1.5 [3]. As shown in Fig. 2, our MMFuser significantly enhances the visual representations input into the MLLM, thereby improving LLaVA-1.5's performance on most multimodal benchmarks. Specifically, our 7B model outperformed LLaVA-1.5-7B in 10 out of 12 benchmarks, and our 13B model outperformed LLaVA-1.5-13B in 10 out of 12 benchmarks. Besides, our model demonstrated superior performance in fine-grained recognition tasks, including OCR and visual grounding.

In summary, our main contributions are as follows:

- We reveal that the expressive potential of single vision encoders in MLLMs is underutilized. Shallow features, rich in detail, suffer from poor semantic alignment with text features, indicating that simple fusion methods are inadequate and require more advanced design.
- We introduce MMFuser, which enhances the visual representations of a single vision encoder by dynamically integrating fine-grained details from shallow features while maintaining the semantic coherence of deep features.
- Applying MMFuser to LLaVA-1.5 models, we achieve significant performance improvements. Our 13B model surpasses LLaVA-1.5 by 3.8, 53.9, and 2.2 points on the VizWiz, MME, and MMBench-EN, respectively, demonstrating the efficacy of our method.

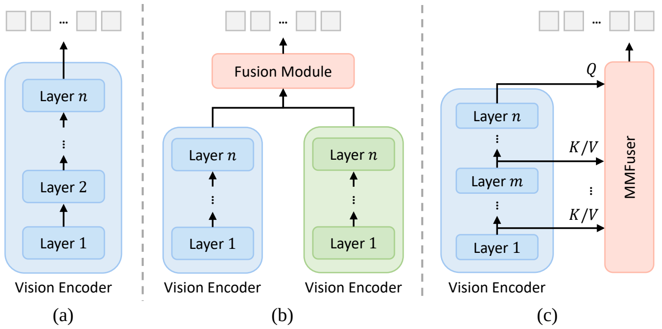


Fig. 3. Previous methods *vs.* the proposed MMFuser. (a) Previous methods typically utilize visual features from the final or penultimate layer of the vision encoder. For example, the LLaVA series [2], [3] adopted this approach. (b) Some models integrate visual features from multiple vision encoders, such as MouSi [47], DeepSeek-VL [15], and LLaVA-HR [28]. (c) Our MMFuser fuses visual features from different layers of a single vision encoder, providing richer detail and better semantic alignment with text.

II. RELATED WORK

A. *Multimodal Large Language Model*

Multimodal Large Language Models (MLLMs) integrate visual representations from images with linguistic embeddings from text, thereby enhancing the models' capabilities in comprehending and generating language descriptions of visual content. Most open-source MLLMs employ architectures that include a pre-trained vision encoder, an LLM, and a cross-modal connector. Early models, such as the BLIP series [1], [36], utilized the Q-Former module to align text and images, thus improving multimodal capabilities. Flamingo [37] employed a gated cross-attention mechanism to integrate images and text. LLaVA-1.5 [3] adopted an MLP projector to connect the pre-trained vision encoder with the LLM. InternVL [4], [6] employed a dynamic resolution strategy, segmenting images into tiles and encoding both the tiles and the thumbnail view together. It then uses a pixel shuffle operation to reduce the number of visual tokens before integrating these features with the LLM through an MLP projector.

Additionally, private MLLMs such as the Gemini series [38], [39], GPT-4V [7], and Claude-3V series [40], along with open-source MLLMs like MiniGPT-4 [23], Qwen-VL [5], CogVLM [41], the VisionLLM series [42], [43], and the All-Seeing series [44], [45], among others [46], have demonstrated robust multimodal capabilities. These models exemplify the powerful ability of MLLMs to understand, generalize, and reason with multimodal information, consistently setting new benchmarks in multimodal tasks.

B. *Vision Encoder in MLLMs*

The vision encoder plays a pivotal role in MLLMs, where notable models like CLIP-ViT [17] have been widely used in this field. CLIP [17] leveraged contrastive learning on large-scale image-text pairs for pre-training, resulting in a vision encoder that learns rich and general visual representations. This capability enhances the understanding of the relationship between vision and language. Several models, including the LLaVA series [2], [3], PaLI [48], and Qwen-VL [5], adopted CLIP-ViT [17], [49] as their default vision encoder.

Additionally, other vision foundation models are employed to construct MLLMs. For instance, CogVLM [41] leveraged the pre-trained EVA2-CLIP-E [50] model for visual representations. ConvLLaVA [26] incorporated ConvNeXt [18], a convolution-based hierarchical model, as its vision encoder. In DeepSeek-VL [15], SigLIP [21] and SAM-ViT [51] were utilized as vision encoders. Furthermore, InternVL [4], [6] employed InternViT-6B, a vision foundation model trained on web-scale image-text data. These works typically use the feature map from the final layer of the vision encoder as the visual representation, as shown in Fig. 3 (a). In contrast, our approach aims to explore the potential benefits of using feature maps from the shallow and intermediate layers of the vision encoder for vision-language tasks.

C. *Enhanced Visual Representation in MLLMs*

Many works are dedicated to enhancing visual representations in MLLMs, including:

1) Scaling Up the Vision Encoder: PaLI [48] increased the parameters of its vision encoder to 4 billion. In PaLI-17B, the vision encoder, ViT-e, accounts for approximately 25% of the total parameters. InternVL [4] scaled its vision foundation model to 6 billion parameters, progressively aligning it with a large language model. PaLM-E [52] achieved a scale of 562 billion parameters by integrating the 540 billion parameter PaLM [53] LLM with the ViT-22B [54].

2) Integrating Multiple Vision Encoders: As shown in Fig. 3 (b), this method enhances visual representations by integrating multiple vision encoders. For example, MMVP [27] employed a Mixture of Features (MoF) approach to integrate image features from CLIP-ViT and DINOv2 [19]. Notably, DINOv2 was a self-supervised vision model trained without any language guidance. Similarly, MouSi [47] utilized an ensemble technique to synergize the capabilities of individual vision encoders. This method introduced a fusion network to unify the processing of outputs from different vision encoders, including CLIP, DINOv2, and SAM [51]. LLaVA-HR [28] integrated image features from two vision encoders of different resolutions: 336px from CLIP-ViT and 1024px from CLIP-ConvNeXt [18]. This approach leveraged the strengths of both resolution inputs to enhance visual understanding. DeepSeek-VL [15] employed a hybrid vision encoder design, effectively encoding images by combining SigLIP-L [21] for low-resolution inputs and SAM-B [51] for high-resolution inputs.

3) Feature Fusion: MEP3P [55] enhanced the original visual features input into MLLMs with image depth features and pseudo-3D positions. VCMR [56] utilized deformable attention to process multi-granularity image features and obtain fine-grained information, thereby improving performance in subsequent cross-modal tasks. Our concurrent work, Dense Connector [57], integrated features from multiple layers, enriching the visual inputs for MLLMs by capturing multi-level representations from the vision encoder.

Overall, these methods demonstrated significant performance boosts for MLLMs. However, the potential of a single vision encoder remained underexplored. To address this, we proposed MMFuser, to integrate feature maps from multiple

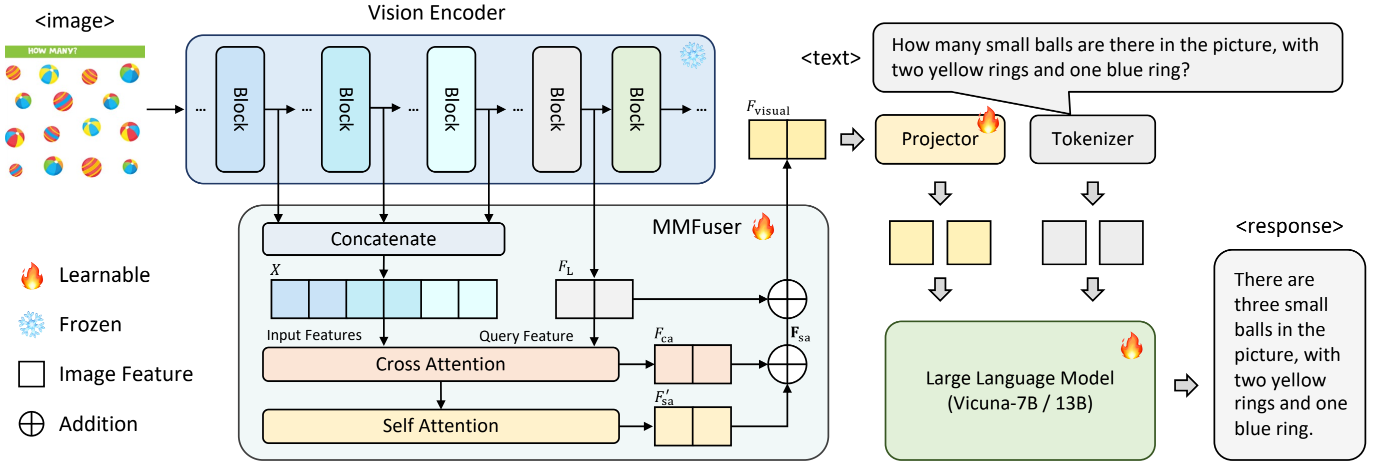


Fig. 4. Overview of MMFuser. In MMFuser, feature maps from different layers of the vision encoder are strategically integrated to enhance the visual representations. Deep feature maps are employed as query elements, while shallow and intermediate feature maps are concatenated to form key and value elements. Through a dynamic attention-based fusion, MMFuser combines fine-grained details and higher-level semantic information. The fused features are then aligned with text using a projector and subsequently passed as inputs to LLMs.

layers of the vision encoder, as shown in Fig. 3 (c). This approach allowed us to obtain more powerful visual representations, thereby enhancing the performance of MLLMs.

III. METHOD

In this section, we address the challenge of missing detailed information in current visual representations for MLLMs. Initial attempts using some simple fusion methods demonstrate that shallow features suffer from inadequate semantic alignment. To overcome this, we propose the MMFuser for MLLMs, designed to effectively integrate multi-layer features while maintaining semantic alignment.

TABLE I
COMPARISON BETWEEN THE LLAVA-1.5 BASELINE, FOUR SIMPLE FEATURE FUSION METHODS, AND OUR MMFUSER.

Method	VizWiz [61]	POPE [62]	MME [63]	MMB^{CN} [64]	MMVet [65]	Avg.
LLaVA-1.5-13B [3]	53.6	85.9	1531.3	63.6	35.4	63.0
w/ Concatenation	52.1	86.9	1537.5	63.7	35.8	63.5
w/ Average	54.7	87.1	1527.9	63.6	35.7	63.0
w/ Weighted Average	54.4	87.0	1532.8	62.5	34.6	63.1
w/ FPN [31]	53.7	87.3	1553.2	63.4	37.3	63.9
w/ MMFuser (Ours)	57.4	87.5	1585.2	63.8	36.6	**64.9**

A. Analysis of Visual Representations for MLLMs

Currently, most mainstream MLLMs [2]–[4], [23], [58], [59] adopt CLIP-ViT [17] as their vision encoder, typically selecting a single feature map from the final layers as the visual representation. Prior studies [60] suggest that in deeper layers of ViT, the receptive fields of attention heads become predominantly global, while the shallow layers retain both local and global information. Consequently, the lack of local details in deep feature maps can lead to suboptimal performance in fine-grained visual recognition tasks.

Different from existing approaches that integrate multiple encoders [15], [28]–[30], we consider that the visual information captured by the CLIP-ViT itself is not fully leveraged in MLLMs. As illustrated in Fig. 1, shallow layers capture fine-grained details, which are often underutilized. Therefore, we argue that combining shallow and deep features can significantly enhance MLLM performance.

To validate our point, we explored several feature map fusion methods. We selected L feature maps from various depths of ViT, denoted as $\mathcal{F} = [F_1, F_2, ..., F_{\mathrm{L}}]$, where $F_i \in \mathbb{R}^{N \times D}$, with $i \in [1, L]$. Here, N represents the number of ViT patches, and D denotes the dimension of the image feature embeddings. The four fusion methods considered are: *(1) Concatenation*: Concatenate feature maps along the channel dimension to create the fused feature map, *i.e.* $\mathrm{Concat}(F_1, F_2, ..., F_{\mathrm{L}})$. *(2) Average*: Compute the element-wise average of all feature maps to obtain the fused feature map, *i.e.* $\frac{1}{L}\sum_{i=1}^{L} F_i$. *(3) Weighted Average*: Assign learnable weights to each feature map, then compute the weighted average as $\sum_{i=1}^{L} w_i F_i$, where w_i is the learnable weight associated with F_i. *(4) Feature Pyramid Network (FPN)* [31]: Feed all feature maps into FPN for multi-scale feature learning, then compute the weighted average of the FPN outputs to obtain the fused feature map.

We applied the fused feature maps as the visual representations in LLaVA-1.5 [3] and followed its original settings to evaluate model performance. However, as shown in Table I, none of the four fusion methods consistently improved model performance. We attribute this to the semantic misalignment between deep and shallow features. As shown in Fig. 1, although shallow features capture more fine-grained details, their alignment with text is considerably weaker than that of deep features. In deeper layers, the features that correspond to the text are prominently highlighted, while shallow features lack this clear correspondence. By simply fusing the two types of image features, the model struggles to effectively leverage the complementary strengths of each feature.

TABLE II

Comparison with state-of-the-art MLLMs on 12 general multimodal benchmarks. The benchmarks include: VQAv2 [67], GQA [68], VizWiz [61], ScienceQA-IMG [69]; TextVQA [70], POPE [62], MME [63], MMBench-EN [64], MMBench-CN [64], SEED-Bench [71], LLaVA-Bench-In-the-Wild [2], and MMVet [65]. * The training annotations of the datasets are observed during training. The best results are marked in **bold**, and the second best results are underlined.

Method	LLM	VQA$^{\text{v2}}$	GQA	VizWiz	SQA^{I}	VQA^{T}	POPE	MME	MMB	MMB^{CN}	SEED	LLaVA^{W}	MMVet
InstructBLIP [1]	Vicuna-7B	–	49.2	34.5	60.5	50.1	–	–	36.0	23.7	53.4	60.9	26.2
IDEFICS-9B [72]	LLaMA-7B	50.9	38.4	35.5	–	25.9	–	–	48.2	25.2	–	–	–
Qwen-VL [5]	Qwen-7B	78.8*	59.3*	35.2	67.1	**63.8***	–	–	38.2	7.4	56.3	–	–
Qwen-VL-Chat [5]	Qwen-7B	78.2*	57.5*	38.9	68.2	61.5*	–	1487.5	60.6	56.7	58.2	–	–
LLaVA-1.5-7B [3]	Vicuna-7B	78.5*	62.0*	50.0	66.8	58.2	85.9	**1510.7**	64.3	58.3	58.6	63.4	30.5
LLaVA-1.5-7B + MMFuser (Ours)	Vicuna-7B	**79.1***	**62.8***	**53.4**	**68.7**	58.8	**86.3**	1479.7	**67.5**	**60.1**	**60.8**	**65.5**	**32.6**
BLIP-2 [36]	Vicuna-13B	65.0	41.0	19.6	61.0	42.5	85.3	1293.8	–	–	46.4	38.1	22.4
InstructBLIP [1]	Vicuna-13B	–	49.5	33.4	63.1	50.7	78.9	1212.8	–	–	–	58.2	25.6
IDEFICS-80B [72]	LLaMA-65B	60.0	45.2	36.0	–	30.9	–	–	54.5	38.1	–	–	–
Shikra [73]	Vicuna-13B	77.4*	–	–	–	–	–	–	58.8	–	–	–	–
LLaVA-1.5-13B [3]	Vicuna-13B	80.0*	63.3*	53.6	**71.6**	**61.3**	85.9	1531.3	67.7	63.6	61.6	70.7	35.4
LLaVA-1.5-13B + MMFuser (Ours)	Vicuna-13B	**80.1***	**63.4***	**57.4**	71.2	59.9	**87.5**	**1585.2**	**69.9**	**63.8**	**62.0**	**71.8**	**36.6**

B. MMFuser: Multimodal Multi-Layer Feature Fuser

Building on the insights from the previous sections, we observe that the shallow and deep features of ViT can complement each other. To harness this potential, we propose a multi-layer feature fusion module, MMFuser. It can serve as a bridge between the vision encoder and the LLM. The overall architecture of MMFuser is shown in Fig. 4.

Specifically, we extract L feature maps from the ViT, denoted as $\mathcal{F} = [F_1, F_2, ..., F_{\text{L}}]$. Since the strong semantic alignment between deep visual features and text space, we use the deep feature F_{L} as queries to dynamically extract missing details from shallow features $X = \text{Concat}(F_1, F_2, ..., F_{\text{L}-1})$, through a cross-attention operation. This results in a visual feature $F_{\text{ca}} \in \mathbb{R}^{N\times D}$ with richer fine-grained features. This process can be formulated as:

$$F_{\text{ca}} = \text{Attention}(\text{norm}(F_{\text{L}}),\ \text{norm}(X)), \tag{1}$$

where $\text{Attention}(\cdot)$ denotes the attention mechanism, $\text{norm}(\cdot)$ means layer normalization [66], and $\text{Concat}(\cdot)$ represents the concatenation operation.

To effectively facilitate feature interaction and emphasize salient features, we incorporate a self-attention mechanism into the feature map F_{ca}, formulated as:

$$\begin{aligned} F'_{\text{sa}} &= \text{Attention}(\text{norm}\,(F_{\text{ca}}),\ \text{norm}\,(F_{\text{ca}})), \\ F_{\text{sa}} &= F_{\text{ca}} + \gamma_2 F'_{\text{sa}}, \end{aligned} \tag{2}$$

where $\gamma_2 \in \mathbb{R}^D$ is a learnable vector that adjusts the contribution of F_{ca} relative to F'_{sa}. Subsequently, for the resulting feature map F_{sa}, we introduce another learnable vector $\gamma_1 \in \mathbb{R}^D$ to modulate the integration of F_{L} and F_{sa}:

$$F_{\text{visual}} = F_{\text{L}} + \gamma_1 F_{\text{sa}}. \tag{3}$$

Through the aforementioned steps, we derive the enhanced visual feature F_{visual}. Unlike the original visual feature F_{L}, F_{visual} integrates richer fine-grained information, making it a superior alternative for the visual input to the LLM.

C. Overall Framework Design

Our proposed MMFuser can be integrated into mainstream open-source MLLMs, particularly within the "ViT-MLP-LLM" architecture [2]–[4], [23]. As a case study, we demonstrate this integration using the LLaVA-1.5 model. In this framework, MMFuser is positioned between the ViT and the MLP projector to fuse multi-layer feature maps from the ViT. The overall architecture, illustrated in Fig. 4, consists of four key components: the vision encoder (ViT), MMFuser, MLP projector, and the large language model.

The input image is first fed into the ViT to extract multi-layer visual features. Then, our MMFuser leverages these multi-layer features to obtain a fused feature with richer local information. Following the pipeline in LLaVA-1.5 [3], the fused features are then aligned with the text embedding space through a trainable MLP projector. Similarly, the input text is transformed into a text embedding via a tokenizer. Finally, the image and text embeddings are concatenated and fed into an LLM, such as Vicuna-7B [8]. The LLM then answers the user's questions based on the image features. This framework can also be readily adapted for processing video data. Overall, existing mainstream MLLMs can easily adopt our MMFuser to enhance their visual feature extraction capabilities.

IV. EXPERIMENT

A. Implementation Details

We adopt LLaVA-1.5 [3] as the baseline to study the visual representations of MLLMs. The model comprises three components: the pre-trained vision encoder CLIP-ViT-L-336px [17], the pre-trained LLM Vicuna-v1.5 [8], and a two-layer MLP projector. To fully leverage the potential of a single vision encoder, we use the proposed MMFuser to fuse multi-layer features from the vision encoder, replacing the original single-layer image feature for the LLM.

1) Architecture Settings: In MMFuser, the number of feature layers L selected from the ViT is set to 5 by default. The parameters γ_1 and γ_2, which control the weights of features from different layers, are both initialized to 0. Deformable

TABLE III
RESULT ON OCRBENCH [82]. RECOG.: TEXT RECOGNITION, VQA^{S}: SCENE TEXT-CENTRIC VQA; VQA^{D}: DOCUMENT-ORIENTED VQA; KIE: KEY INFORMATION EXTRACTION; HMER: HANDWRITTEN MATHEMATICAL EXPRESSION; FINAL: OVERALL SCORE ACROSS ALL FIVE CATEGORIES.

Method	Recog.	VQA^{S}	VQA^{D}	KIE	HMER	Final
MiniGPT4V2 [83]	124	29	4	0	0	157
BLIP2 [36]	154	71	10	0	0	235
InstructBLIP [1]	168	93	14	1	0	276
BLIVA [84]	165	103	22	1	0	291
LLaVA-1.5-7B [3]	160	117	15	5	0	297
+ MMFuser (Ours)	159	128	20	8	0	**315**$^{\uparrow 18}$
LLaVA-1.5-13B [3]	176	129	19	7	0	331
+ MMFuser (Ours)	171	136	25	11	0	**343**$^{\uparrow 12}$

attention [74] is used as the default attention mechanism in MMFuser. In this setup, the number of sampling points is fixed at 4, with the attention layer employing 16 heads.

2) Training Settings: For a fair comparison, we adopt the same two-stage training recipes as LLaVA-1.5 [3]:

Pre-training. During the pre-training stage, we utilize the LLaVA-LCS-558K dataset [2], which comprises 558K image-caption pairs. In this phase, the vision encoder and the LLM are kept frozen. Training is focused solely on the MLP projector and MMFuser, with the objective of aligning the visual features with the input space of the LLM.

Fine-tuning. During the fine-tuning stage, we employ the LLaVA-Instruct-665K dataset [3], which includes 665K instruction-following data from sources such as GQA [68], TextCaps [75], ShareGPT [76], and others [67], [77]–[81]. In this phase, the MLP projector, MMFuser, and LLM are trained end-to-end, while the vision encoder remains frozen.

We train our model using the same experimental settings and hyperparameters as LLaVA-1.5. We adopt the AdamW optimizer and use a cosine decay learning rate scheduler with an initial warmup ratio of 0.03 and no weight decay. The global batch size is set to 256 for pre-training and 128 for fine-tuning. The learning rate is set to 1e-3 during the pre-training phase and reduced to 2e-5 during the fine-tuning phase. Both phases are trained for a single epoch.

B. Results on General Multimodal Benchmarks

We evaluated MMFuser using a comprehensive suite of 12 benchmarks, including both academic visual question answering (VQA) benchmarks and comprehensive multimodal benchmarks, to assess its performance across multiple dimensions. As shown in Table II, our model exhibits substantial performance improvements on these benchmarks.

1) Results on Academic VQA Benchmarks: On the academic VQA benchmarks, our 7B model consistently outperforms LLaVA-1.5-7B across all five benchmarks. Similarly, the 13B version of our model surpasses LLaVA-1.5-13B on the VQAv2, GQA, and VizWiz benchmarks, with a particularly notable improvement of 3.8 points on VizWiz. Furthermore, our model achieves comparable performance on the ScienceQA and TextVQA benchmarks.

TABLE IV
RESULTS OF REGION CAPTIONING. RESULTS ARE REPORTED WITH THE CIDER SCORE.

Model	RefCOCO			RefCOCO+			RefCOCOg		Avg.
	val	testA	testB	val	testA	testB	val	test	
LLaVA-1.5-7B	30.4	16.0	42.0	30.2	20.3	39.1	60.5	58.9	37.2
+ MMFuser	33.6	17.7	45.9	33.6	21.2	42.6	61.5	61.6	**39.7**
LLaVA-1.5-13B	33.1	16.7	45.2	33.4	19.8	41.6	61.6	59.9	38.9
+ MMFuser	38.2	19.5	53.6	36.8	22.8	43.8	64.5	63.4	**42.8**

TABLE V
RESULTS OF REFERRING EXPRESSION COMPREHENSION. RESULTS ARE REPORTED WITH THE PRECISION@0.5 SCORE.

Model	RefCOCO			RefCOCO+			RefCOCOg		Avg.
	val	testA	testB	val	testA	testB	val	test	
LLaVA-1.5-7B	56.2	64.4	47.5	50.0	59.2	39.0	48.8	48.4	51.7
+ MMFuser	62.0	70.7	52.4	55.6	65.1	44.2	54.0	54.8	**57.4**
LLaVA-1.5-13B	66.5	73.9	55.7	59.8	67.9	48.7	57.3	56.0	60.7
+ MMFuser	66.6	73.9	56.3	61.3	68.8	49.2	56.5	56.5	**61.1**

2) Results on Comprehensive Multimodal Benchmarks: In the multimodal benchmarks, our 7B and 13B models demonstrate significant performance improvements compared to the corresponding LLaVA-1.5 models. Notably, our 13B model substantially outperforms LLaVA-1.5-13B across seven different benchmarks. Specifically, our 13B model achieves scores of 1585.2 on the MME benchmark and 69.9 on the MMBench benchmark, representing enhancements of 53.9 and 2.2 points over LLaVA-1.5-13B, respectively. Additionally, our model exhibits strong performance across other multimodal benchmarks, such as POPE, SEED-Bench, and MMVet.

C. Results on OCRBench

OCRBench [82] serves as a comprehensive OCR benchmark, comprising 1,000 manually curated and corrected OCR-related VQA instructions. The benchmark is systematically divided into five distinct categories: Text Recognition (Recog.), Scene Text-Centric VQA (VQA^{S}), Document-Oriented VQA (VQA^{D}), Key Information Extraction (KIE), and Handwritten Mathematical Expression Recognition (HMER).

As detailed in Table III, our models, with 7B and 13B parameters, exhibit an average improvement of 15 points over LLaVA-1.5. This substantial gain underscores MMFuser's enhanced capability in refining the granularity of visual representations, thereby contributing to more accurate text recognition and superior OCR performance.

D. Results on Region-level Benchmarks

To assess regional understanding and grounding capabilities, we evaluate MMFuser on two representative regional-level tasks. (1) Regional Captioning [81], [85]: This task requires the model to generate a description for an object in the image based on a given region. (2) Referring Expression Comprehension [80], [85]: This task requires the model to locate target objects in an image based on a given description.

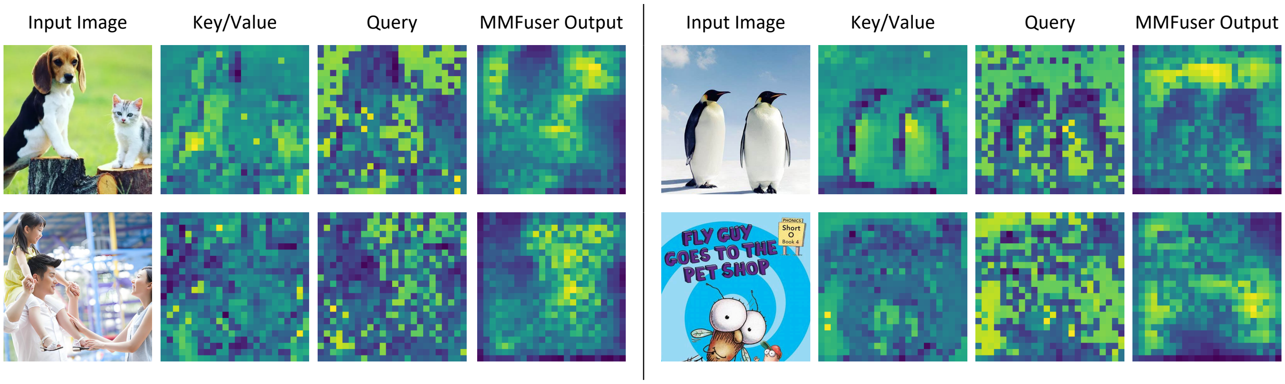


Fig. 5. Feature map visualization of MMFuser. For each image, we provide three types of output feature maps. The term "Key/Value" refers to the averaged feature maps from four selected shallow and intermediate layers of the ViT—specifically, layers 3, 8, 13, and 18—used as the key and value inputs in MMFuser. "Query" denotes the feature map from the penultimate layer of the ViT, serving as the query input in MMFuser and as the visual representations in prior MLLMs. "MMFuser Output" represents the feature map generated after applying the proposed MMFuser. As can be seen, the proposed MMFuser captures fine-grained details from shallow and intermediate ViT layers, enriching the visual representations for the LLM.

TABLE VI
ABLATION ON LAYER COMBINATION.

Selected Layers		VizWiz	POPE	MME	MMB	SEED	MMVet	Avg.
Q	K, V							
–	–	53.6	85.9	1531.3	67.7	61.6	35.4	63.5
23	[1, 3, 5, 7]	54.3	87.3	1582.2	69.5	61.8	35.1	64.5
23	[9, 11, 13, 15]	52.4	87.0	1560.3	69.1	61.7	35.6	64.0
23	[17, 19, 21, 24]	54.7	87.3	**1591.2**	69.3	62.0	35.8	64.8
23	[5, 8, 11, 20]	54.7	87.2	1584.0	69.3	**62.5**	36.2	64.9
23	[3, 8, 13, 18]	**57.4**	**87.5**	1585.2	**69.9**	62.0	**36.6**	**65.4**

1) Results of Region Captioning: On region captioning tasks, our model shows significant improvements. As shown in Table IV, compared to LLaVA-1.5, the 7B model of MMFuser surpasses LLaVA-1.5 by 2.5 points on average, while the 13B version improves by 3.9 points. This indicates that MMFuser captures fine-grained information, enhancing caption accuracy and richness.

2) Results of Referring Expression Comprehension (REC): We also employ REC tasks to evaluate the model's grounding capabilities. As shown in Table V, our model consistently outperforms LLaVA-1.5 models across all benchmarks, with an especially notable average improvement of 5.7 points for the 7B model compared to LLaVA-1.5-7B. This highlights that the visual representations generated by MMFuser are more detailed and comprehensive, enhancing spatial localization and significantly boosting performance in grounding tasks.

E. Ablation Study

1) Ablation Study on Layer Combination: To evaluate the contribution of feature maps from different layers of ViT [17] to model performance, we conduct an ablation study on the MMFuser-13B model by sampling feature maps from various depths. As detailed in Table VI, we utilize the feature map from the 23rd layer of ViT as the query, while the key and value are formed by concatenating feature maps sampled from different layers: shallow (row 2), intermediate (row 3), deep (row 4), non-uniform (row 5), and uniform sampling (row 6).

TABLE VII
ABLATIONS ON ATTENTION MECHANISMS. "CPLX." INDICATES THE COMPLEXITY TYPE OF THE ATTENTION MECHANISM, ENCOMPASSING BOTH QUADRATIC (QUAD.) AND LINEAR VARIANTS.

Attention Type	Cplx.	VizWiz	POPE	MME	MMB	SEED	MMVet	Avg.
LLaVA-1.5-13B [3]	–	53.6	85.9	1531.3	67.7	61.6	35.4	63.5
Global Attn [86]	Quad.	52.9	**87.6**	1566.3	68.6	**62.2**	35.3	64.2
Linear SRA [87]	Linear	54.3	87.0	1581.6	68.9	61.9	34.7	64.3
Deformable Attn [74]	Linear	**57.4**	87.5	**1585.2**	**69.9**	62.0	**36.6**	**65.4**

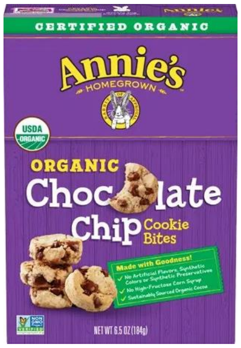
TABLE VIII
ABLATIONS ON INTERNAL MODULE DESIGNS.

Module	VizWiz	POPE	MME	MMB	SEED	MMVet	Avg.
LLaVA-1.5-13B	53.6	85.9	1531.3	67.7	61.6	35.4	63.5
+ Cross-Attn	54.6	87.2	1557.1	68.2	**62.2**	36.1	64.4
++ Self-Attn	**57.4**	**87.5**	**1585.2**	**69.9**	62.0	**36.6**	**65.4**

The results demonstrate that incorporating feature maps from multiple layers significantly improves the model's visual representation capabilities. The feature maps from different layers capture information at varying receptive field sizes, each offering distinct advantages depending on the task. Notably, uniform sampling across layers, which aggregates information from diverse receptive fields, achieves the highest performance. Therefore, MMFuser employs this strategy to optimize visual representation for visual-language tasks.

2) Ablation on Attention Mechanisms: The attention mechanism in MMFuser is modular and can be replaced with different variants. We experiment with three types of attention mechanisms using the 13B model of MMFuser. As shown in Table VII, our framework consistently enhances the visual representation capability of the MLLM, regardless of the attention mechanism employed. This confirms the effectiveness and adaptability of our feature fusion module.

Compared to global attention mechanisms [86], sparse attention mechanisms with linear complexity not only deliver superior performance in MMFuser but also offer substantial

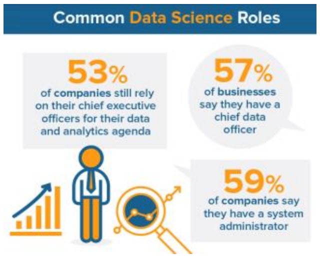


User: Please provide a detailed description of the image.

MMFuser (Ours): The image features a box of Annie's Homegrown Organic Chocolate Chip Cookie Bites. The box is purple and has a picture of a cookie on it. The cookies are made with organic ingredients and are free from synthetic flavors, synthetic colors, and synthetic preservatives. The box contains a total of 12 cookies, making it a perfect snack option for those who enjoy organic and healthier alternatives.

LLaVA-1.5: The image features a box of Annie's Organic Chocolate Chip Cookie Bites. The box is purple and has a picture of a cookie on it. The cookies are made with organic ingredients and are gluten-free, making them a healthier alternative to traditional chocolate chip cookies. The box contains a dozen cookies, each with a bite-sized portion, perfect for snacking or sharing with friends and family.

(a) Qualitative example of the complex captioning capability.

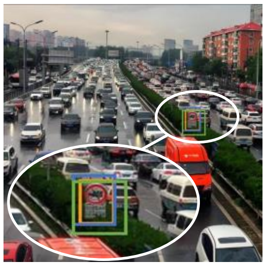

Common Data Science Roles
53% of companies still rely on their chief executive officers for their data and analytics agenda
57% of businesses say they have a chief data officer
59% of companies say they have a system administrator

User: Please identify the text in the picture.

MMFuser (Ours): The image displays a series of statistics related to common data science roles. The first statistic shows that 53% of companies still rely on their chief executive officer (CEO) for their data and analytics strategy. The second statistic indicates that 57% of businesses say they have a chief data officer (CDO) to lead their data and analytics efforts. The third statistic highlights that 58% of companies believe their system administrator should be responsible for their data and analytics strategy.

LLaVA-1.5: The image features a pie chart with various data points related to common data science roles. The chart is divided into four sections. The first section is titled "53% of companies still rely on their chief executive officer for their data and analytics agenda." The second section is titled "57% of companies say they have a system administrator." The third section is titled "59% of companies say they have a business intelligence officer." The fourth section is titled "58% of companies say they have a chief data officer."

(b) Qualitative example of the OCR capability.

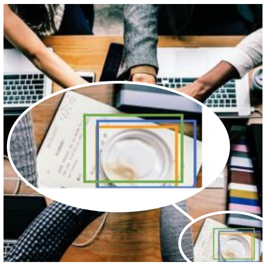


Ground-truth:
[0.70, 0.41, 0.76, 0.49]

User:
Please provide the bounding box coordinates of the area described in this sentence: Traffic signs.

MMFuser (Ours): IoU: 0.76
[0.69, 0.40, 0.76, 0.49]

LLaVA-1.5: IoU: 0.53
[0.69, 0.41, 0.78, 0.51]



Ground-truth:
[0.83, 0.89, 0.96, 1.00]

User:
Please provide the bounding box coordinates of the area described in this sentence: Glass water cup.

MMFuser (Ours): IoU: 0.74
[0.83, 0.88, 0.99, 1.00]

LLaVA-1.5: IoU: 0.63
[0.81, 0.87, 0.97, 0.99]

(c) Qualitative example of the visual grounding capability.

Fig. 6. Qualitative comparisons between MMFuser and LLaVA-1.5. For the complex captioning and OCR tasks, the text generated by each model is color-coded to match the model name, indicating content that accurately reflects the information present in the image. Text in red denotes errors or hallucinations. For visual grounding examples, the predicted bounding boxes are also color-coded according to the respective model names, while yellow bounding boxes denote the ground truth. The Intersection over Union (IoU) metric is used to evaluate the overlap between predicted bounding boxes and the ground truth. A higher IoU value indicates a more accurate prediction, as it reflects a larger intersection area relative to the union area.

improvements in computational efficiency. Among the linear attention mechanisms, deformable attention [74] provides the most significant performance gains. Therefore, deformable attention is adopted as the default mechanism in MMFuser. It is noteworthy that future research could explore more advanced attention mechanisms to further enhance the visual representation capabilities of MLLMs.

3) Ablation on Internal Module Designs: To validate the key contributions of our MMFuser architecture, we incrementally enhance the LLaVA-1.5-13B baseline [3] with our proposed designs. As illustrated in Table VIII, integrating cross-attention mechanisms to extract fine-grained information from various ViT layers leads to a marked performance improvement. Specifically, our model outperforms the baseline by 1.3 points on POPE and 0.7 points on MMVet. Moreover, the addition of self-attention to these fine-grained features further enhances the model's capability to capture relevant information, resulting in gains of 2.8 points on VizWiz, 28.1 points on MME, and 1.7 points on MMB. These results collectively demonstrate that our design significantly enhances the visual representation abilities of MLLMs, highlighting the critical role of each component in our model.

4) Visual Representation Visualization: To intuitively validate the impact of MMFuser on visual features, we present the input and output feature map visualizations for four example images in Fig. 5. For the input image, the shallow feature maps convey richer fine-grained information, but this information is messy and the semantic information is difficult to distinguish. For instance, the lower-left image contains complex information, making it challenging to intuitively discern semantic content from the shallow features. But after the attention mechanism of MMFuser, the foreground characters in the picture are well highlighted, and the semantic information is aligned with the last layer feature map of ViT. The aligned detailed features can effectively enhance the fine-grained perception capability of MLLMs.

F. Qualitative Comparison

In Fig. 6, we compare MMFuser and LLaVA-1.5 across three case studies related to fine-grained analysis, including complex captioning, OCR, and visual grounding. The visualization results show that the rich fine-grained information in the shallow feature maps significantly enhances the ability of MLLMs to capture and process image details.

For the complex captioning example, MMFuser shows great improvement in recognizing small text and small objects. Compared to LLaVA-1.5, MMFuser even successfully identifies the small and blurred green background text describing the healthy ingredients of the cookie: “free from synthetic flavors, synthetic colors, and synthetic preservatives.”

For the OCR task, MMFuser shows enhanced recognition of small texts, effectively mitigating hallucinations that often occur in LLaVA-1.5. The latter struggles with misaligned and hallucinated content due to inadequate text recognition capabilities, leading it to supplement details from its prior knowledge rather than the image itself.

In the visual grounding example, demonstrates a more accurate delineation of object boundaries, especially when adjacent objects have similar colors. This improvement is more pronounced with small objects, such as the positioning of the traffic sign in the left-side example, where MMFuser shows a 23% increase in the IoU metric.

In summary, these results further validate the effectiveness of MMFuser in addressing the issue of missing fine-grained image details in MLLMs.

V. CONCLUSION

In this paper, we introduced MMFuser, a novel multimodal multi-layer feature fuser designed to improve visual representation in Multimodal Large Language Models (MLLMs). By integrating both shallow and deep features from the vision encoder, MMFuser addresses the limitations of relying solely on deep features from a single vision encoder, which often results in the loss of fine-grained details. Our experiments show that MMFuser enhances the performance of the LLaVA-1.5 model across various benchmarks, enriching visual representation without the redundancy and computational overhead of ensemble models. This approach maximizes the potential of a single ViT encoder, offering an efficient and flexible solution for MLLMs. Overall, MMFuser improves fine-grained detail capture and semantic understanding in MLLMs, and we hope it will contribute to the community’s efforts in developing more robust and efficient multimodal models.

REFERENCES

[1] W. Dai, J. Li, D. Li, A. Huat, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi, “Instructblip: Towards general-purpose vision-language models with instruction tuning,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2023.

[2] H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual instruction tuning,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2024.

[3] H. Liu, C. Li, Y. Li, and Y. J. Lee, “Improved baselines with visual instruction tuning,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 26 296–26 306.

[4] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu *et al.*, “Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 24 185–24 198.

[5] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, “Qwen-vl: A frontier large vision-language model with versatile abilities,” *arXiv preprint arXiv:2308.12966*, 2023.

[6] Z. Chen, W. Wang, H. Tian, S. Ye, Z. Gao, E. Cui, W. Tong, K. Hu, J. Luo, Z. Ma *et al.*, “How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites,” *arXiv preprint arXiv:2404.16821*, 2024.

[7] OpenAI, “Gpt-4v(ision) system card,” 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:263218031

[8] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez *et al.*, “Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality,” Mar. 2023. [Online]. Available: https://lmsys.org/blog/2023-03-30-vicuna/

[9] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, “Llama: Open and efficient foundation language models,” *arXiv preprint arXiv:2302.13971*, 2023.

[10] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, “Llama 2: Open foundation and fine-tuned chat models,” *arXiv preprint arXiv:2307.09288*, 2023.

[11] OpenAI, “Chatgpt,” 2022. [Online]. Available: https://chat.openai.com/

[12] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, “Gpt-4 technical report,” *arXiv preprint arXiv:2303.08774*, 2023.

[13] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan *et al.*, “The llama 3 herd of models,” *arXiv preprint arXiv:2407.21783*, 2024.

[14] I. Team, “Internlm: A multilingual language model with progressively enhanced capabilities,” 2023. [Online]. Available: https://github.com/InternLM/InternLM

[15] H. Lu, W. Liu, B. Zhang, B. Wang, K. Dong, B. Liu, J. Sun, T. Ren, Z. Li, Y. Sun *et al.*, “Deepseek-vl: towards real-world vision-language understanding,” *arXiv preprint arXiv:2403.05525*, 2024.

[16] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang *et al.*, “Qwen technical report,” *arXiv preprint arXiv:2309.16609*, 2023.

[17] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021.

[18] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, “A convnet for the 2020s,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 11 976–11 986.

[19] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, “Dinov2: Learning robust visual features without supervision,” *Trans. Mach. Learn. Res.*, pp. 1–31, 2024.

[20] Y. Fang, Q. Sun, X. Wang, T. Huang, X. Wang, and Y. Cao, “Eva-02: A visual representation for neon genesis,” *Image Vision Comput.*, vol. 149, p. 105171, 2024.

[21] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, “Sigmoid loss for language image pre-training,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 11 975–11 986.

[22] K. Zheng, X. He, and X. E. Wang, "Minigpt-5: Interleaved vision-and-language generation via generative vokens," *arXiv preprint arXiv:2310.02239*, 2023.

[23] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," *arXiv preprint arXiv:2304.10592*, 2023.

[24] R. Xu, Y. Yao, Z. Guo, J. Cui, Z. Ni, C. Ge, T.-S. Chua, Z. Liu, M. Sun, and G. Huang, "Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images," *arXiv preprint arXiv:2403.11703*, 2024.

[25] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.

[26] C. Ge, S. Cheng, Z. Wang, J. Yuan, Y. Gao, J. Song, S. Song, G. Huang, and B. Zheng, "Convllava: Hierarchical backbones as visual encoder for large multimodal models," *arXiv preprint arXiv:2405.15738*, 2024.

[27] S. Tong, Z. Liu, Y. Zhai, Y. Ma, Y. LeCun, and S. Xie, "Eyes wide shut? exploring the visual shortcomings of multimodal llms," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 9568–9578.

[28] G. Luo, Y. Zhou, Y. Zhang, X. Zheng, X. Sun, and R. Ji, "Feast your eyes: Mixture-of-resolution adaptation for multimodal large language models," *arXiv preprint arXiv:2403.03003*, 2024.

[29] D. Jiang, Y. Liu, S. Liu, J. Zhao, H. Zhang, Z. Gao, X. Zhang, J. Li, and H. Xiong, "From clip to dino: Visual encoders shout in multi-modal large language models," *arXiv preprint arXiv:2310.08825*, 2024.

[30] Y. Li, Y. Zhang, C. Wang, Z. Zhong, Y. Chen, R. Chu, S. Liu, and J. Jia, "Mini-gemini: Mining the potential of multi-modality vision language models," *arXiv preprint arXiv:2403.18814*, 2024.

[31] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 2117–2125.

[32] S. Xu, H. Zhang, X. Xu, X. Hu, Y. Xu, L. Dai, K.-S. Choi, and P.-A. Heng, "Representative feature alignment for adaptive object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 2, pp. 689–700, 2023.

[33] W. Gao, G. Liao, S. Ma, G. Li, Y. Liang, and W. Lin, "Unified information fusion network for multi-modal rgb-d and rgb-t salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 2091–2106, 2022.

[34] Z. Chen, H. Ji, Y. Zhang, Z. Zhu, and Y. Li, "High-resolution feature pyramid network for small object detection on drone view," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 1, pp. 475–489, 2024.

[35] S. Wang, H. Lu, and Z. Deng, "Fast object detection in compressed video," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 7104–7113.

[36] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2023, pp. 19 730–19 742.

[37] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, "Flamingo: a visual language model for few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 23 716–23 736.

[38] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth *et al.*, "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.

[39] M. Reid, N. Savinov, D. Teplyashin, D. Lepikhin, T. Lillicrap, J.-b. Alayrac, R. Soricut, A. Lazaridou, O. Firat, J. Schrittwieser *et al.*, "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," *arXiv preprint arXiv:2403.05530*, 2024.

[40] Anthropic, "The claude 3 model family: Opus, sonnet, haiku," 2024. [Online]. Available: https://www.anthropic.com

[41] W. Wang, Q. Lv, W. Yu, W. Hong, J. Qi, Y. Wang, J. Ji, Z. Yang, L. Zhao, X. Song *et al.*, "Cogvlm: Visual expert for pretrained language models," *arXiv preprint arXiv:2311.03079*, 2023.

[42] W. Wang, Z. Chen, X. Chen, J. Wu, X. Zhu, G. Zeng, P. Luo, T. Lu, J. Zhou, Y. Qiao *et al.*, "Visionllm: Large language model is also an open-ended decoder for vision-centric tasks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2024.

[43] J. Wu, M. Zhong, S. Xing, Z. Lai, Z. Liu, W. Wang, Z. Chen, X. Zhu, L. Lu, T. Lu *et al.*, "Visionllm v2: An end-to-end generalist multimodal large language model for hundreds of vision-language tasks," *arXiv preprint arXiv:2406.08394*, 2024.

[44] W. Wang, M. Shi, Q. Li, W. Wang, Z. Huang, L. Xing, Z. Chen, H. Li, X. Zhu, Z. Cao *et al.*, "The all-seeing project: Towards panoptic visual recognition and understanding of the open world," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2024.

[45] W. Wang, Y. Ren, H. Luo, T. Li, C. Yan, Z. Chen, W. Wang, Q. Li, L. Lu, X. Zhu *et al.*, "The all-seeing project v2: Towards general relation comprehension of the open world," *arXiv preprint arXiv:2402.19474*, 2024.

[46] Z. Peng, W. Wang, L. Dong, Y. Hao, S. Huang, S. Ma, and F. Wei, "Kosmos-2: Grounding multimodal large language models to the world," *arXiv preprint arXiv:2306.14824*, 2023.

[47] X. Fan, T. Ji, C. Jiang, S. Li, S. Jin, S. Song, J. Wang, B. Hong, L. Chen, G. Zheng *et al.*, "Mousi: Poly-visual-expert vision-language models," *arXiv preprint arXiv:2401.17221*, 2024.

[48] X. Chen, X. Wang, S. Changpinyo, A. Piergiovanni, P. Padlewski, D. Salz, S. Goodman, A. Grycner, B. Mustafa, L. Beyer *et al.*, "Pali: A jointly-scaled multilingual language-image model," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2023.

[49] G. Ilharco, M. Wortsman, R. Wightman, C. Gordon, N. Carlini, R. Taori, A. Dave, V. Shankar, H. Namkoong, J. Miller, H. Hajishirzi, A. Farhadi, and L. Schmidt, "Openclip," Jul. 2021. [Online]. Available: https://doi.org/10.5281/zenodo.5143773

[50] Q. Sun, Y. Fang, L. Wu, X. Wang, and Y. Cao, "Eva-clip: Improved training techniques for clip at scale," *arXiv preprint arXiv:2303.15389*, 2023.

[51] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 4015–4026.

[52] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu *et al.*, "Palm-e: An embodied multimodal language model," in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 202, 2023, pp. 8469–8488.

[53] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann *et al.*, "Palm: Scaling language modeling with pathways," *J. Mach. Learn. Res.*, vol. 24, no. 240, pp. 1–113, 2023.

[54] M. Dehghani, J. Djolonga, B. Mustafa, P. Padlewski, J. Heek, J. Gilmer, A. P. Steiner, M. Caron, R. Geirhos, I. Alabdulmohsin *et al.*, "Scaling vision transformers to 22 billion parameters," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2023, pp. 7480–7512.

[55] J. Zhu, H. Wang, and M. Shi, "Multi-modal large language model enhanced pseudo 3d perception framework for visual commonsense reasoning," *IEEE Trans. Circuits Syst. Video Technol.*, pp. 1–1, 2024.

[56] X. Zhang, P. Zhao, J. Ji, X. Lu, and Y. Yin, "Video corpus moment retrieval via deformable multigranularity feature fusion and adversarial training," *IEEE Trans. Circuits Syst. Video Technol.*, pp. 1–1, 2023.

[57] H. Yao, W. Wu, T. Yang, Y. Song, M. Zhang, H. Feng, Y. Sun, Z. Li, W. Ouyang, and J. Wang, "Dense connector for mllms," *arXiv preprint arXiv:2405.13800*, 2024.

[58] B. McKinzie, Z. Gan, J.-P. Fauconnier, S. Dodge, B. Zhang, P. Dufter, D. Shah, X. Du, F. Peng, F. Weers *et al.*, "Mm1: Methods, analysis & insights from multimodal llm pre-training," *arXiv preprint arXiv:2403.09611*, 2024.

[59] Z. Li, B. Yang, Q. Liu, Z. Ma, S. Zhang, J. Yang, Y. Sun, Y. Liu, and X. Bai, "Monkey: Image resolution and text label are important things for large multi-modal models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 26 763–26 773.

[60] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?" in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 12 116–12 128.

[61] D. Gurari, Q. Li, A. J. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, and J. P. Bigham, "Vizwiz grand challenge: Answering visual questions from blind people," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 3608–3617.

[62] Y. Li, Y. Du, K. Zhou, J. Wang, W. X. Zhao, and J.-R. Wen, "Evaluating object hallucination in large vision-language models," *Proc. Empirical Methods Natural Lang. Process.*, pp. 292–305, 2023.

[63] C. Fu, P. Chen, Y. Shen, Y. Qin, M. Zhang, X. Lin, Z. Qiu, W. Lin, J. Yang, X. Zheng *et al.*, "Mme: A comprehensive evaluation benchmark for multimodal large language models," *arXiv preprint arXiv:2306.13394*, 2023.

[64] Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu *et al.*, "Mmbench: Is your multi-modal model an all-around player?" *arXiv preprint arXiv:2307.06281*, 2023.

[65] W. Yu, Z. Yang, L. Li, J. Wang, K. Lin, Z. Liu, X. Wang, and L. Wang, "Mm-vet: Evaluating large multimodal models for integrated capabilities," *arXiv preprint arXiv:2308.02490*, 2023.

[66] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[67] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the v in vqa matter: Elevating the role of image understanding in visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 6325–6334.

[68] D. A. Hudson and C. D. Manning, "Gqa: A new dataset for real-world visual reasoning and compositional question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 6700–6709.

[69] P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, O. Tafjord, P. Clark, and A. Kalyan, "Learn to explain: Multimodal reasoning via thought chains for science question answering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022.

[70] A. Singh, V. Natarajan, M. Shah, Y. Jiang, X. Chen, D. Batra, D. Parikh, and M. Rohrbach, "Towards vqa models that can read," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 8317–8326.

[71] B. Li, R. Wang, G. Wang, Y. Ge, Y. Ge, and Y. Shan, "Seed-bench: Benchmarking multimodal llms with generative comprehension," *arXiv preprint arXiv:2307.16125*, 2023.

[72] IDEFICS, "Introducing idefics: An open reproduction of state-of-the-art visual language model," 2023. [Online]. Available: https://huggingface.co/blog/idefics

[73] K. Chen, Z. Zhang, W. Zeng, R. Zhang, F. Zhu, and R. Zhao, "Shikra: Unleashing multimodal llm's referential dialogue magic," *arXiv preprint arXiv:2306.15195*, 2023.

[74] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020.

[75] O. Sidorov, R. Hu, M. Rohrbach, and A. Singh, "Textcaps: a dataset for image captioning with reading comprehension," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Springer, 2020, pp. 742–758.

[76] ShareGPT, 2023. [Online]. Available: https://sharegpt.com/

[77] K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi, "Ok-vqa: A visual question answering benchmark requiring external knowledge," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 3195–3204.

[78] A. Mishra, S. Shekhar, A. K. Singh, and A. Chakraborty, "Ocr-vqa: Visual question answering by reading text in images," in *Int. J. Doc. Anal. Recog.*, 2019, pp. 947–952.

[79] D. Schwenk, A. Khandelwal, C. Clark, K. Marino, and R. Mottaghi, "A-okvqa: A benchmark for visual question answering using world knowledge," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 146–162.

[80] S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg, "Referitgame: Referring to objects in photographs of natural scenes," in *Proc. Empirical Methods Natural Lang. Process.*, 2014, pp. 787–798.

[81] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vis.*, pp. 32–73, 2017.

[82] Y. Liu, Z. Li, M. Huang, B. Yang, W. Yu, C. Li, X. Yin, C. lin Liu, L. Jin, and X. Bai, "Ocrbench: On the hidden mystery of ocr in large multimodal models," *arXiv preprint arXiv:2305.07895*, 2024.

[83] J. Chen, D. Zhu, X. Shen, X. Li, Z. Liu, P. Zhang, R. Krishnamoorthi, V. Chandra, Y. Xiong, and M. Elhoseiny, "Minigpt-v2: large language model as a unified interface for vision-language multi-task learning," *arXiv preprint arXiv:2310.09478*, 2023.

[84] W. Hu, Y. Xu, Y. Li, W. Li, Z. Chen, and Z. Tu, "Bliva: A simple multimodal llm for better handling of text-rich visual questions," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 38, no. 3, 2024, pp. 2256–2264.

[85] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. Murphy, "Generation and comprehension of unambiguous object descriptions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 11–20.

[86] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017.

[87] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pvtv2: Improved baselines with pyramid vision transformer," *Computational Visual Media*, vol. 8, no. 3, pp. 415–424, 2022.