Playing to Vision Foundation Model's Strengths in Stereo Matching

Chuang-Wei Liu, Qijun Chen, *Senior Member, IEEE*, and Rui Fan, *Senior Member, IEEE*

***Abstract*—Stereo matching has become a key technique for 3D environment perception in intelligent vehicles. For a considerable time, convolutional neural networks (CNNs) have remained the mainstream choice for feature extraction in this domain. Nonetheless, there is a growing consensus that the existing paradigm should evolve towards vision foundation models (VFM), particularly those developed based on vision Transformers (ViTs) and pre-trained through self-supervision on extensive, unlabeled datasets. While VFMs are adept at extracting informative, general-purpose visual features, specifically for dense prediction tasks, their performance often lacks in geometric vision tasks. This study serves as the first exploration of a viable approach for adapting VFMs to stereo matching. Our ViT adapter, referred to as ViTAS, is constructed upon three types of modules: spatial differentiation, patch attention fusion, and cross-attention. The first module initializes feature pyramids, while the latter two aggregate stereo and multi-scale contextual information into fine-grained features, respectively. ViTAStereo, which combines ViTAS with cost volume-based stereo matching back-end processes, achieves the top rank on the KITTI Stereo 2012 dataset and outperforms the second-best network StereoBase by approximately 7.9% in terms of the percentage of error pixels, with a tolerance of 3 pixels. Additional experiments across diverse scenarios further demonstrate its superior generalizability compared to all other state-of-the-art approaches. We believe this new paradigm will pave the way for the next generation of stereo matching networks. Our source code and supplementary material are publicly available at https://mias.group/ViTAS.**

***Index Terms*—stereo matching, intelligent vehicle, vision foundation model, geometry vision task, attention.**

I. Introduction

STEREO matching, which mimics human binocular depth perception, has long been a key technique in intelligent vehicles and mobile robots [1]–[4]. Vision foundation models (VFMs) have rapidly emerged as a focal point in the field of computer vision [5], [6]. From models like Segmentation Anything [7] and DINOv2 [8] by Meta AI to the more recent Depth Anything [9], VFMs have garnered significant attention and interest. Surprisingly, despite its fundamental role in 3D computer vision, stereo matching has not received adequate attention amidst the wave of VFMs yet. Therefore, this article serves as the first attempt to navigate stereo matching into this new continent, with a specific emphasis on adapting VFMs for more generalizable stereo matching.

Recent advancements in stereo matching have primarily focused on the back-end processes, including cost volume construction [10], cost aggregation [11], and disparity refinement [12], while relatively overlooking the development of deep feature extractors. This is largely attributed to the effectiveness of traditional backbone networks, such as ResNet [13] and MobileNet [14], in extracting rich deep features for matching cost computation. Specifically, the limited progress in this process has focused on employing cross-attention layers after the traditional backbone network for further feature quality improvements [15]–[18]. Nevertheless, recent VFMs, generally built upon a Vision Transformer (ViT), have demonstrated greater effectiveness in learning informative, general-purpose deep features across various related computer vision tasks, when pre-trained in a self-supervised fashion on large curated datasets [7]–[9]. Therefore, a key focus of this article lies in the development of ViT adapters to selectively leverage these general-purpose deep features to improve stereo matching.

Despite the extensive application of pre-trained VFMs with task-specific adapters for scene parsing tasks [19]–[21], their utilization in 3D geometric vision tasks solved by pixel-wise dense matching, such as stereo matching, remains unexplored. This is primarily because VFMs trained for image segmentation (pixel-level classification) and monocular depth estimation (pixel-level regression) are not capable of producing features that are sufficiently distinct for similarity measurement in the cost volume construction stage [22]. The significant domain gap of geometry information between VFM features and those preferred by 3D geometric vision tasks renders existing VFM adapters infeasible for stereo matching. Thus, the primary objective in designing our VFM adapter is to further enhance feature distinctiveness, thereby reducing ambiguities in stereo matching.

Additionally, it is noteworthy that there is a potential trend among state-of-the-art (SoTA) networks [22], [23] to shift away from constructing cost volumes for stereo matching. These networks generally employ a Transformer with an encoder-decoder architecture to aggregate stereo knowledge into features from a single view. These features are then taken as input by a dense prediction Transformer [24] for disparity regression. This new design transforms disparity estimation from pixel-wise matching process into a regression task, thereby sacrificing the explicit constraint on the absolute scale of disparity offered by epipolar geometry. Consequently, we are intrigued by the generalizability of such a network

This research was supported in part by the National Natural Science Foundation of China under Grant 62473288 and Grant 62233013, the Science and Technology Commission of Shanghai Municipal under Grant 22511104500, the Fundamental Research Funds for the Central Universities, and the Xiaomi Young Talents Program. (*Corresponding author: Rui Fan*)

Chuang-Wei Liu, Qijun Chen, and Rui Fan are with the College of Electronics & Information Engineering, Shanghai Research Institute for Intelligent Autonomous Systems, the State Key Laboratory of Intelligent Autonomous Systems, and Frontiers Science Center for Intelligent Autonomous Systems, Tongji University, Shanghai 201804, China (e-mail: {cwliu, qjchen, rfan}@tongji.edu.cn).

design in unseen scenarios and have conducted extensive experiments across various public datasets. Regrettably, the comprehensive experimental results suggest a shortfall in its performance on unseen data, particularly evident in its tendency to estimate disparity based on the disparity range encountered during training. This limitation could possibly be attributed to the reduced explainability of disparity estimation without the use of cost volumes. This observation further reinforces our motivation for playing to VFM's strengths by developing an effective adapter to fully exploit the general-purpose deep features for cost volume construction, rather than simply regressing disparities from these features without any interpretability.

Therefore, in this article, we introduce **ViT Adapter for Stereo (ViTAS)**, playing to the strengths of VFMs in stereo matching. Our proposed ViTAS incorporates three types of modules: (1) the spatial differentiation module (SDM), which captures multi-scale contextual information by initializing feature pyramids, akin to the studies presented in [11], [12], [16], [25], (2) the patch attention fusion module (PAFM), which aggregates multi-scale contextual information into fine-grained features, and (3) the cross-attention module (CAM), which aggregates stereo contextual information into extracted features via cross-view interactions. Notably, our newly developed PAFM employs local patch attention and quasi-global attention, devised in accordance with the pixel-to-patch and squeeze-and-excitation manners, to learn the local and global feature weighting parameters, respectively. The PAFM significantly reduces computational complexity and memory usage compared to the conventional global attention mechanism [26]–[28], which learns these features simultaneously. Combining ViTAS with cost volume-based stereo matching back-end processes yields **ViTAStereo**, a SoTA, powerful, and highly generalizable stereo matching network. ViTAStereo achieves **top ranking** on the KITTI Stereo 2012 dataset and **second-best performance** on the KITTI Stereo 2015 dataset [29], outperforming StereoBase, the current SoTA stereo matching network, by approximately 5.2-11.3% in the percentage of error pixels.

We conclude the contributions of this study as follows:

- We introduce ViTAS, marking the first research endeavor to fully exploit the informative, general-purpose features extracted by VFMs for stereo matching.
- We develop a novel, lightweight PAFM that learns local and global feature weighting parameters separately, effectively, and efficiently.
- We argue that stereo matching networks relying solely on cross-attention mechanism have limited generalizability, primarily due to the absence of cost volumes.
- We conduct extensive experiments to demonstrate the SoTA performance and superior generalizability of ViTAStereo across various public datasets.

The remainder of this article is structured as follows: related works, including ViT adapters and stereo matching networks, are presented in Sect. II. Sect. III details our proposed ViTAS. Comprehensive ablation studies and comparative experiments are presented in Sect. IV. Finally, in Sect. V, we summarize the results and provide recommendations for future work.

II. Related Work

A. ViT Adapters

SoTA VFMs generally utilize a plain ViT as their backbone network. To date, ViT adapters have found widespread application in 2D computer vision tasks. For instance, ViTDet [19], [20] enables the plain, non-hierarchical ViT architecture to undergo fine-tuning for object detection without the need for redesigning a hierarchical backbone for pre-training. Similarly, ViT-Adapter [21] injects image priors into the ViT using an additional attention path, resulting in superior accuracy in both object detection and semantic/instance segmentation tasks. On the other hand, recent researches on stereo matching task have explored various training strategies, including contrastive learning and self-supervision, to leverage the ViT architecture [30], [31]. However, there has been relatively less focus on the development of adapters in this field. It is likely that the deep features utilized for scene understanding tasks are not inherently suitable or compatible with geometric vision tasks. Therefore, developing an adapter compatible with the plain ViT architecture for geometric vision tasks is a promising area of research that requires more attention.

B. Stereo Matching

1) Cost Volume-based Networks: Recent SoTA stereo matching networks [11], [12], [16], [25] based on cost volumes have largely overlooked pyramid feature extraction and instead focused mainly on back-end stages that process features and costs. Since the introduction of RAFT-Stereo [25], its core component–a multi-level gated recurrent unit, has become prevalent in stereo matching. This unit takes feature pyramids extracted by a conventional deep feature extractor as input for cost aggregation, enabling the incorporation of semantic information at various scales. RAFT-Stereo has significantly influenced subsequent advancements in stereo matching networks, such as those seen in CREStereo [16], IGEV-Stereo [11], and GMStereo [12]. These networks have further extended and refined this multi-scale structure, often integrating attention mechanism to improve feature representation and address local matching ambiguities in challenging regions. For instance, CREStereo [16] follows LoFTR [32] and incorporates an attention module at the lowest resolution to aggregate global contextual and stereo information in single or cross-view feature maps. IGEV-Stereo [11], on the other hand, combines multi-scale correlation volumes with a geometry encoding volume obtained through 3D convolutions, aiming at addressing local matching ambiguities in ill-posed regions. Moreover, GMStereo [12] further extends this multi-scale structure into both stereo matching and optical flow estimation tasks, with the exactly same learnable parameters. Similar to CreStereo, GMStereo also employs attention modules prior to cost volume construction, but across all spatial scales. In contrast to these SoTA methods, we leverage recent VFMs for feature extraction and design our adapter, drawing inspiration from the existing pyramid feature structure and attention mechanism.

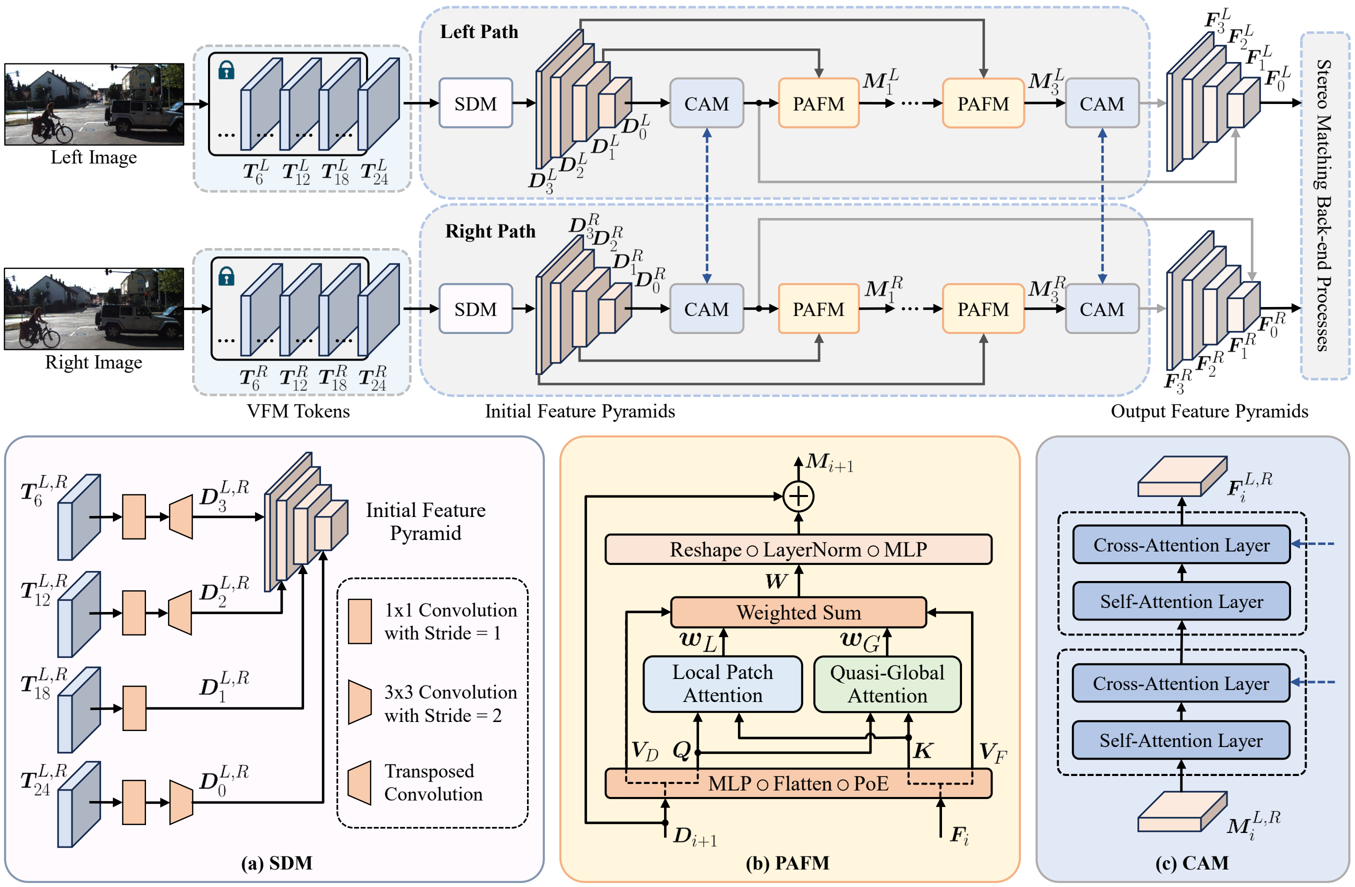


Fig. 1. An illustration of our proposed **ViTAS**. ViTAS employs a Siamese architecture and each sub-network is comprised of a SDM, four CAMs, and three PAFMs. The output feature pyramids are passed through the back-end processes of a stereo matching network for disparity estimation. The superscripts "L" and "R" denote the left and right views, respectively.

2) Cost Volume-free Network: CroCo-Stereo [22] is a seminal contribution in this domain. It utilizes an encoder-decoder Transformer, initially pre-trained on the ImageNet database [33] for cross-view completion [23], and subsequently fine-tuned for stereo matching by incorporating a dense prediction Transformer (DPT) head [31]. Specifically, stereo information is aggregated into the left-view features through cross-attention via a plain ViT decoder, followed by a DPT head to regress disparities. However, due to the utilization of a large VFM encoder and extensively deployed attention modules in the decoder, CroCo-Stereo incurs even greater computational and memory overhead compared to cost volume-based networks. More importantly, abandoning the cost volume considerably reduces its generalizability and interoperability, as demonstrated by our extensive experiments. These limitations underscore the importance of prioritizing compatibility with cost volume-based networks.

III. Methodology

The overall task is first formulated in Sect. III-A. Then, the three modules in ViTAS are detailed in Sects. III-B1, III-B2, and III-B3, respectively.

A. Task Formulation

Pyramid feature extraction has been prevalently used in stereo matching [12], [16], [25], [34], primarily due to its capability to handle objects of various scales while maintaining high computational efficiency [35], [36]. As demonstrated in recent works [11], [12], [16], [25], [37], [38], features at four scales ($1/32$, $1/16$, $1/8$, and $1/4$ of the original image resolution) have been shown to be sufficient and effective for stereo matching. However, the SoTA VFMs [7], [8] generally adopt the plain ViT architecture [39] for feature extraction, resulting in extracted features at a single resolution. Specifically, a VFM consists of a patch embedding layer and N consecutive Transformer encoders. An input image $\boldsymbol{I} \in \mathbb{R}^{h_0 \times w_0 \times 3}$ is first divided into a collection of $p \times p$ non-overlapping patches by the patch embedding layer. These patches are then sequentially projected into N tokens $\boldsymbol{T} \in \mathbb{R}^{h_0/p \times w_0/p \times c_T}$ through the Transformer encoders. Therefore, to fully exploit the general-purpose VFM features, the most fundamental task of our ViTAS is to transform the tokens into a collection of pyramid features $\mathcal{F} = \{\boldsymbol{F}_0, \boldsymbol{F}_1, \boldsymbol{F}_2, \boldsymbol{F}_3\}$, where $\boldsymbol{F}_k \in \mathbb{R}^{\frac{h_0}{2^{5-k}} \times \frac{w_0}{2^{5-k}} \times c_k}$.

In this study, we use DINOv2 [8] as our backbone VFM, in which N is set to 24. Nonetheless, a limitation arises as it sets the parameter p to 14, causing a misalignment between

Algorithm 1: ViTAS workflow

```
Input: VFM tokens $\mathcal{T}^{L,R}$
Output: Output feature pyramids $\mathcal{F}^{L,R}$
Generating initialized feature pyramids from VFM tokens
  via: $\mathcal{D}^{L,R} \leftarrow \mathrm{SDM}(\mathcal{T}^{L,R})$;
Aggregating stereo contextual information for the deepest
  initial features via: $\{\boldsymbol{F}_0^L, \boldsymbol{F}_0^R\} \leftarrow \mathrm{CAM}(\boldsymbol{D}_0^L, \boldsymbol{D}_0^R)$ ;
Aggregating stereo and multi-scale contextual information
  hierarchically via: for $i \leftarrow 1$ to $3$ do
    $\boldsymbol{M}_i^{L,R} \leftarrow \mathrm{PAFM}(\boldsymbol{F}_{i-1}^{L,R}, \boldsymbol{D}_i^{L,R})$;
    $\boldsymbol{F}_i^{L,R} \leftarrow \mathrm{CAM}(\boldsymbol{M}_i^{L,R}, \boldsymbol{M}_i^{R,L})$;
```

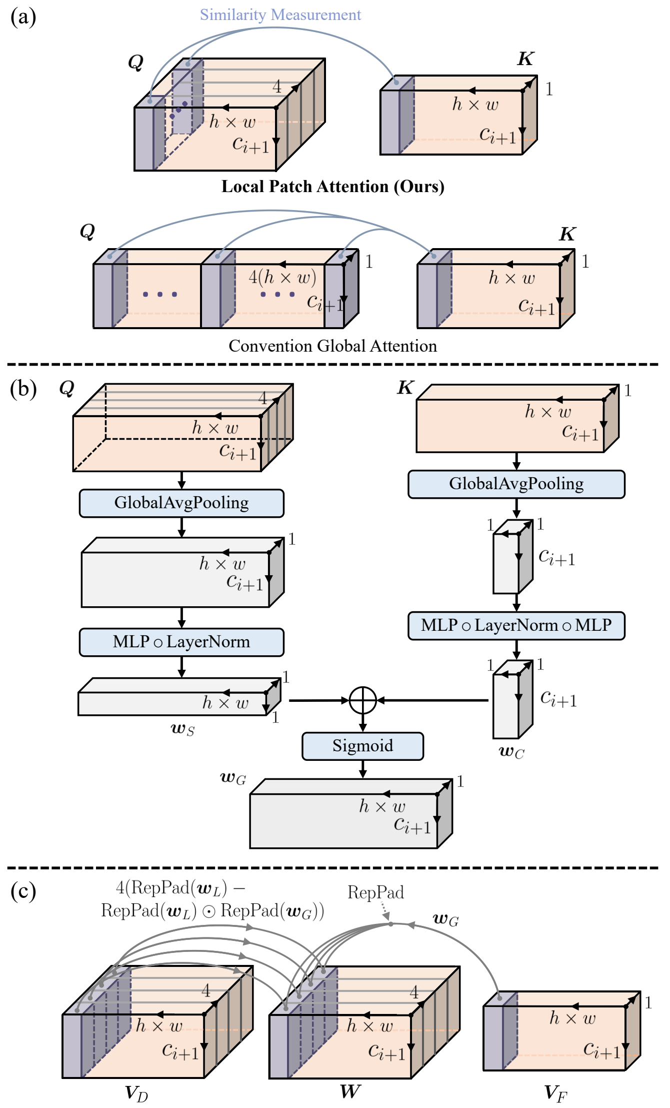


Fig. 2. Illustrations of (a) local path attention versus conventional global attention, (b) quasi-global attention, and (c) multi-scale feature aggregation within PAFM.

token scales and the preferred pyramid feature scales. To address this issue, we first adjust the input stereo images by a factor of $\frac{14}{16}$, resulting in tokens at $\frac{1}{16}$ of the original image resolution. Subsequently, following recent advancements in ViT adapters [19], [21], we split the Transformer encoders into four groups. From each group, we select tokens generated by the final Transformer block to serve as input for our ViTAS. Consequently, the essential goal of our ViTAS is to transform the $\mathcal{T}^{L,R} = \{\boldsymbol{T}_6^{L,R}, \boldsymbol{T}_{12}^{L,R}, \boldsymbol{T}_{18}^{L,R}, \boldsymbol{T}_{24}^{L,R}\}$ into $\mathcal{F}^{L,R}$, where the superscripts L and R correspond to the left and right images, respectively.

B. ViTAS Architecture

As depicted in Fig. 1, our proposed ViTAS adopts a Siamese architecture comprising two weight-sharing subnetworks. Each sub-network is dedicated to processing one view of the stereo images and consists of a SDM followed by three PAFMs and four CAMs arranged alternately. The SDM generates initial feature pyramids, while the PAFMs and CAMs hierarchically aggregate stereo and multi-scale contextual information into fine-grained features, respectively. Each module type is designed to accomplish an independent task, making our ViTAS highly modular and adaptable to future updates with more advanced techniques. The workflow of our proposed ViTAS is detailed in Algorithm 1.

1) SDM: Recent studies [40], [41] have demonstrated the complementarity between ViTs and convolutional neural networks (CNNs). The former excels at capturing global contextual information, while the latter enriches the local spatial patterns of the ViT tokens. Therefore, we introduce the SDM at the beginning of ViTAS to re-scale the ViT tokens, as illustrated in Fig. 1(a). This process enables ViTAS to capture multi-scale contextual information, resulting in significantly improved stereo matching accuracy, particularly for small objects and boundaries.

The input ViT tokens $\mathcal{T}^{L,R}$ are first assembled into initial feature pyramids $\mathcal{D}^{L,R} = \{\boldsymbol{D}_0^{L,R}, \boldsymbol{D}_1^{L,R}, \boldsymbol{D}_2^{L,R}, \boldsymbol{D}_3^{L,R}\}$ using two SDMs. Each SDM consists of four blocks of convolutions and transpose convolutions, where $\boldsymbol{D}_k^{L,R}$ is at a resolution equal to $1/2^{5-k}$ of the original image. Since deeper ViT tokens contain richer global context and shallower ones focus on fine-grained details [24], we assemble ViT tokens from deeper to shallower layers with gradually increasing resolutions. In addition, we reduce the number of channels in the initial feature pyramids to further alleviate computational and memory pressure. A hierarchical refinement process is then performed with CAMs and PAFMs from $\boldsymbol{D}_0^{L,R}$ to $\boldsymbol{D}_3^{L,R}$, as detailed in the remainder of this section.

2) PAFM: With the feature pyramids $\mathcal{D}^{L,R}$ initialized by SDM, we perform multi-scale feature fusion to aggregate $\boldsymbol{F}_i$ and $\boldsymbol{D}_{i+1}$. Although recent Transformer-based multi-scale feature fusion approaches [28], [42] have demonstrated superior performance over CNN-based methods [26], [27], [43], [44] in dense prediction tasks [27], their computational complexity and memory consumption, inherent in the global attention mechanism, pose significant challenges. Furthermore, a common limitation persists wherein lower-resolution feature maps have to be upsampled (typically via bilinear interpolation) to align with the higher-resolution feature maps. However, such simplistic feature upsampling operations fail to preserve fine-grained details in low-resolution features [45], [46]. To overcome these limitations, we design PAFM, a lightweight yet effective Transformer-based multi-scale feature fusion module. As illustrated in Fig. 1(b), our proposed PAFM consists of a local patch attention and a quasi-global attention, capable of learning local weights $\boldsymbol{w}_L \in \mathbb{R}^{(h\times w)\times 4\times 1}$ and global weights $\boldsymbol{w}_G \in \mathbb{R}^{(h\times w)\times 1\times c_{i+1}}$, respectively. $\boldsymbol{F}_i$ is aggregated based on $\boldsymbol{w}_G$, while $\boldsymbol{D}_{i+1}$ is aggregated based on both $\boldsymbol{w}_G$ and $\boldsymbol{w}_L$.

The local patch attention measures the fine-grained feature similarity between a given pixel in $\boldsymbol{F}_i \in \mathbb{R}^{h\times w\times c_i}$ and its corresponding patch with a resolution of 2×2 pixels in $\boldsymbol{D}_{i+1} \in \mathbb{R}^{2h\times 2w\times c_{i+1}}$. To this end, we project $\boldsymbol{D}_{i+1}$ into query $\boldsymbol{Q} \in \mathbb{R}^{(h\times w)\times 4\times c_{i+1}}$ and value $\boldsymbol{V}_D \in \mathbb{R}^{(h\times w)\times 4\times c_{i+1}}$, and project $\boldsymbol{F}_i$ into key $\boldsymbol{K} \in \mathbb{R}^{(h\times w)\times 1\times c_{i+1}}$ and another value $\boldsymbol{V}_F \in \mathbb{R}^{(h\times w)\times 1\times c_{i+1}}$, where the second dimension, referred to as the "patch dimension" in this article, depicts the number of pixels inside a feature patch. Compared to conventional global attention, our local patch attention operates by measuring feature similarity in a pixel-to-patch manner, as illustrated in Fig. 2(a). This approach dramatically reduces the computational demands, lowering the complexity from $\mathcal{O}(N^2)$ to $\mathcal{O}(N)$. The local weights $\boldsymbol{w}_L$ are computed through the following process:

$$\boldsymbol{w}_L = \mathrm{Softmax}\left(\frac{\boldsymbol{Q}\odot \mathrm{RepPad}(\boldsymbol{K})\times \boldsymbol{O}}{\sqrt{c_{i+1}}}\right), \tag{1}$$

where $\odot$ denotes the element-wise multiplication operation, RepPad denotes replication padding operation, $\boldsymbol{O} \in \mathbb{R}^{c_{i+1}\times 1}$ is a matrix storing ones, and the $\mathrm{Softmax}$ operation is performed at the patch dimension to normalize the weights within a feature patch.

The quasi-global attention emphasizes both the informative spatial areas in $\boldsymbol{D}_{i+1}$ and the prominent feature channels in $\boldsymbol{F}_i$, yielding the global weights $\boldsymbol{w}_G$, as illustrated in Fig. 2(b). To this end, we first employ a global average pooling layer to squeeze $\boldsymbol{Q}$ along the patch dimension, thereby aligning its size with $\boldsymbol{K}$. We then aggregate the squeezed features along the channel dimension using a multi-layer perceptron (MLP) to prioritize informative spatial areas while suppressing redundant ones, thereby producing the spatial weights $\boldsymbol{w}_S \in \mathbb{R}^{(h\times w)\times 1\times 1}$. In the meantime, we follow the design of squeeze-and-excitation block [47], [48] and produce the context weights $\boldsymbol{w}_C \in \mathbb{R}^{1\times 1\times c_{i+1}}$ from $\boldsymbol{K}$ (containing rich global contextual information), highlighting prominent feature channels while de-emphasizing the less important ones. This is achieved through a combination of a global average pooling layer and two MLPs. Finally, the global weights $\boldsymbol{w}_G$ are calculated as follows:

$$\boldsymbol{w}_G = \mathrm{Sigmoid}\left(\mathrm{RepPad}\left(\boldsymbol{w}_C\right) \oplus \mathrm{RepPad}\left(\boldsymbol{w}_S\right)\right), \tag{2}$$

where $\oplus$ denotes the element-wise summation operation. Afterwards, as depicted in Fig. 2(c), we combine $\boldsymbol{w}_L$ and $\boldsymbol{w}_G$, the local and global weights, to adaptively fuse $\boldsymbol{V}_D$ and $\boldsymbol{V}_F$ as follows:

$$\begin{aligned}\boldsymbol{W} &= \mathrm{RepPad}\left(\boldsymbol{w}_G \odot \boldsymbol{V}_F\right) \oplus \\ &4\left(\mathrm{RepPad}\left(\boldsymbol{w}_L\right) - \mathrm{RepPad}\left(\boldsymbol{w}_L\right) \odot \mathrm{RepPad}\left(\boldsymbol{w}_G\right)\right) \odot \boldsymbol{V}_D,\end{aligned} \tag{3}$$

where $\boldsymbol{W} \in \mathbb{R}^{(h\times w)\times 4\times c_{i+1}}$ denotes the fused features. Consequently, multi-scale contextual information is aggregated into $\boldsymbol{W}$ with weight parameters balancing between $\boldsymbol{Q}$ and $\boldsymbol{V}$ (summing to 4 in each feature patch). With fully fused information from $\boldsymbol{D}_{i+1}$ and $\boldsymbol{F}_i$, the final output of our PAFM is derived as follows:

$$\boldsymbol{M}_{i+1} = \boldsymbol{D}_{i+1} + \mathrm{Reshape} \circ \mathrm{LayerNorm} \circ \mathrm{MLP}(\boldsymbol{W}), \tag{4}$$

which is subsequently fed into a CAM for stereo contextual information aggregation.

3) CAM: CAMs have been commonly utilized in stereo matching networks [12], [16], [49] to aggregate stereo contextual information into features through cross-view feature interactions. In this study, CAMs are strategically positioned after the SDM and interleaved with PAFMs, before generating the output feature pyramids $\mathcal{F}^{L,R}$. As depicted in Fig. 1(c), each CAM contains two attention blocks, each of which consists of a self-attention layer and a cross-attention layer, respectively. The former aggregates global contextual information, whereas the latter enhances feature distinctiveness, thereby reducing disparity ambiguities, particularly in textureless and occluded regions. Despite having similar structures and sharing the same query feature sources, these two layers diverge in key and value feature sources: the self-attention layer uses features from the same view, whereas the cross-attention layer uses features from the other view. The existing networks for stereo matching suffer from a crucial limitation due to the absence of cross-scale feature interaction, leading to an over-reliance on self-attention layers to capture global contextual information. To address this issue, GMStereo [12] utilizes six attention blocks within each CAM to independently process features across different layers, albeit at a notable increase in computational complexity. In contrast, our proposed ViTAS has a progressive architecture, wherein PAFMs collaborate with CAMs to aggregate both global and stereo contextual information from deeper layers into shallower ones. As a result, ViTAS utilizes markedly more lightweight CAMs to process features, significantly reducing the computational complexity and memory demands in comparison to GMStereo.

IV. Experiments

This section comprehensively analyzes the effectiveness of our proposed ViTAS in improving both disparity accuracy and network generalizability. The following subsections delve into details on datasets and implementations, evaluation metrics, ablation studies, and a thorough performance evaluation.

A. Datasets and Implementation Details

Five public stereo matching datasets are utilized in our experiments for model pre-training and fine-tuning. The following two synthetic, large-scale datasets with dense disparity ground truth are employed for the first purpose:

1) **SceneFlow** [50] consists of a training set (containing 35,454 stereo image pairs) and a test set (often known as the Flying 3D test set, containing 4,370 stereo image pairs) with the image resolution of 960×540 pixels. We

use the “finalpass” version rather than the “cleanpass” version because it is more realistic.

2) **Virtual KITTI** [51] contains 21,260 stereo image pairs (resolution: $1,242 \times 375$ pixels), generated from five different virtual scenarios (created using the Unity game engine and a real-to-virtual cloning method) in urban settings under different imaging and weather conditions.

The following three real-world, small datasets are used to fine-tune networks and evaluate their performance:

1) **KITTI Stereo** contains two subsets: 2012 [29] and 2015 [52], with 192 and 200 training pairs, respectively, and 194 and 200 test pairs, respectively. The image resolution is around $1,240 \times 370$ pixels and the sparse disparity ground truth is generated using a LIDAR.
2) **Middlebury** [53]–[55] contains five subsets: 2005, 2006, 2014, 2021, and MiddEval3, with 45, 171, 132, 335, and 14 pairs of high-resolution indoor stereo images and their corresponding disparity ground truth provided by structured light, respectively.
3) **ETH3D** [56] contains 27 pairs of stereo gray-scale images of both indoor and outdoor scenes (resolution: around 930×490 pixels). A high-precision laser scanner is used to provide the disparity ground truth.

In our ablation studies (see Sect. IV-C), each network is first trained on the SceneFlow training set and Virtual KITTI dataset for 50 epochs. The pre-trained networks are determined based on their performance on the test set of the SceneFlow dataset [50], with the networks demonstrating the best performance being selected.

When comparing our proposed ViTAStereo with SoTA networks on the KITTI test set (see Sect. IV-D), we first pre-train our network on the combined training set of the aforementioned five datasets for 100 epochs. Subsequently, an additional fine-tuning stage is conducted on the KITTI training set for 200 epochs.

When evaluating the generalizability of our proposed ViTAStereo (see Sect. IV-E), we adopt the same pre-training strategy used in the ablation studies. Specifically, we split the original KITTI training set into two subsets: **KITTI Train** and **KITTI Eval**, for model fine-tuning and generalizability evaluation, respectively. Similarly, we divide the original Middlebury dataset into two subsets: **Midd Train** (excluding the MiddEval3 dataset) and **Midd Eval** (identical to the MiddEval3 dataset) for model fine-tuning and generalizability evaluation, respectively. The entire **ETH3D** dataset is used only for generalizability evaluation.

All experiments are conducted on four NVIDIA RTX 4090 GPUs. During model training, we randomly crop images to 320×720 pixels and apply conventional data augmentation techniques, including random changes in image color, random rescaling, and random erasing, to further enhance model performance. The back-end components in IGEV-Stereo [11] are used to build our ViTAStereo, primarily due to the similar hardware requirement (both are capable of training and testing on a GPU with 24 GB GDDR6X memory). The parameters of the VFM, excluding those of the last five ViT encoder blocks, are frozen. In our generalizability evaluation experiments, we further demonstrate the compatibility of our proposed ViTAStereo with three additional SoTA stereo matching networks, GMStereo [12], CREStereo [16], and CroCo-Stereo [22]. The experimental results on CroCo-Stereo are presented in the supplementary material, demonstrating that stereo matching networks relying solely on cross-attention mechanisms have limited generalizability, primarily due to the absence of cost volumes. The loss function, learning rate, and optimizer used in our experiments are identical to the settings reported in their publications [11], [12], [16], [22].

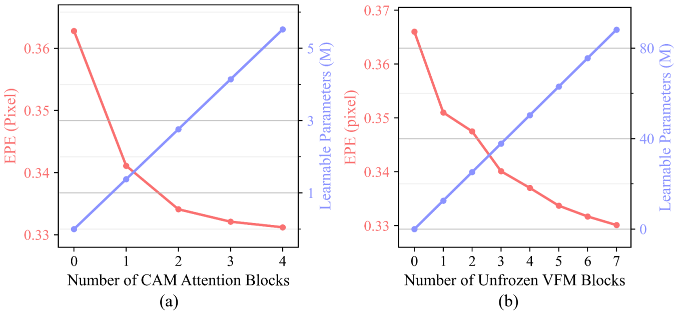


Fig. 3. Ablation studies on (a) the optimal configuration for CAM attention blocks and (b) the most suitable number of unfrozen VFM blocks.

TABLE I
ABLATION STUDY ON THE EFFECTIVENESS OF EACH COMPONENT WITHIN ViTAS.

SDM	Fusion Methods			CAM	EPE (pixel)	D1-all (%)	Runtime (s)
	PAFM	SDFA [27]	VAF [28]				
✔	✔			✔	**0.334**	**1.109**	0.278
	✔			✔	0.351	1.206	0.275
✔				✔	0.388	1.349	0.273
✔	✔				0.362	1.281	0.266
✔					0.427	1.389	0.262
	✔				0.383	1.266	0.265
				✔	0.417	1.348	0.272
					0.435	1.394	**0.261**
✔		✔		✔	0.351	1.155	0.274
✔			✔	✔	0.335	1.122	0.287

B. Evaluation Metrics

The following three metrics are computed to quantify stereo matching accuracy (lower values indicate better performance):

- **end-point error (EPE)**, indicating the average disparity estimation error;
- **percentage of error pixels (PEP)**, indicating the percentage of incorrect disparities with respect to a tolerance of δ pixels;
- **D1**, indicating the percentage of disparities for which the estimation error exceeds both three pixels and 5% of the ground-truth disparity.

C. Ablation Studies

We first investigate the optimal configuration for CAM and determine the most suitable number of unfrozen VFM blocks,

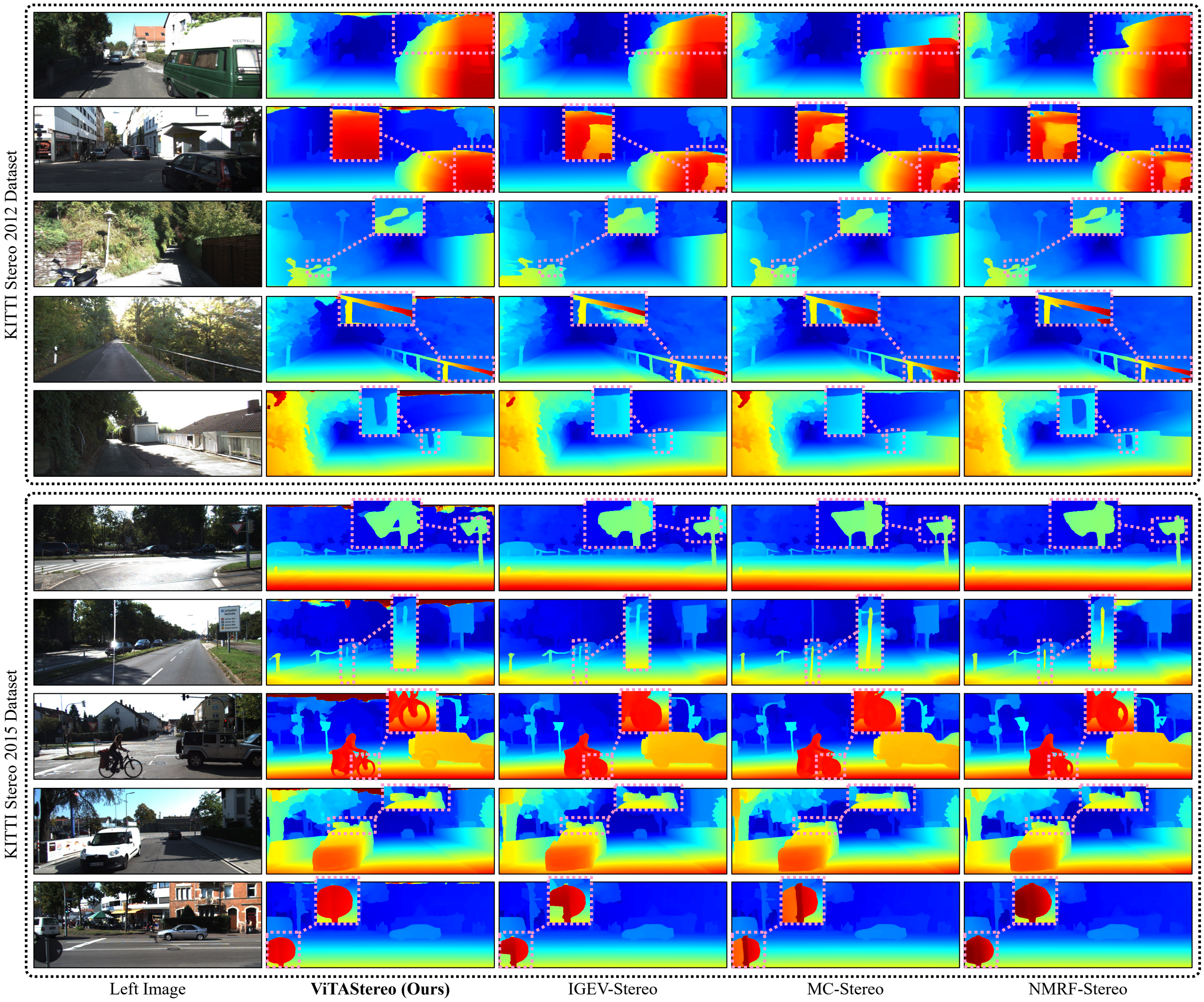


Fig. 4. Qualitative experimental results of ViTAStereo, IGEV-Stereo [11], MC-Stereo [57], and NMRF-Stereo [58] on the KITTI Stereo datasets [29], [52], where significantly improved regions are shown with pink dashed boxes.

TABLE II
THE AMOUNTS OF LEARNABLE PARAMETERS, MEMORY DEMANDS, AND FUSION INFERENCE TIME OF PAFM AND ANOTHER TWO METICULOUSLY DESIGNED FEATURE FUSION APPROACHES.

Feature Fusion Methods	Parameters (M)	Memory (MB)	Inference Time (ms)
PAFM	**0**.22	118	6.11
SDFA [27]	1.38	**98.3**	**5.67**
VAF [28]	1.31	360	17.1

TABLE III
ABLATION STUDY ON RECENT VFMS FOR BUILDING VITASTEREO.

VFM	EPE (pixel)	PEP w.r.t Different δ			D1 (%)
		1	2	3	
DINOv2 [8]	**0.334**	**3.05**	**1.82**	**1.36**	**1.11**
Depth Anything [9]	0.341	3.12	1.87	1.41	1.13

as detailed in Fig. 3. It is evident that incorporating a greater number of attention blocks and unfreezing additional VFM blocks both contribute to reductions in EPE, albeit at the cost of a significant increase in the model's learnable parameters. Therefore, in subsequent experiments, we build our CAM with only two attention blocks and unfreeze the last five VFM blocks, so as to minimize the trade-off between disparity accuracy and network complexity.

In an additional ablation study conducted to validate the effectiveness of each component within our ViTAS in terms of both disparity estimation accuracy and the runtime of ViTAStereo. The findings, as detailed in Table I, demonstrate that the inclusion of any single module leads to improved disparity accuracy. Specifically, the incorporation of SDM, PAFM, and CAM independently leads to reductions in the EPE by 1.8%, 11.9%, and 4.1%, respectively. When all three modules are incorporated, ViTAS achieves the most significant decrease in EPE. This investigation further indicates that

TABLE IV

COMPARISONS WITH SOTA STEREO MATCHING NETWORKS PUBLISHED ON THE KITTI STEREO 2012 DATASET [29]. "δ-NOC" DENOTES PEP FOR NON-OCCLUDED PIXELS W.R.T. δ, AND "ALL" DENOTES PEP FOR ALL PIXELS W.R.T. δ.

Network	PEP w.r.t Different δ								Runtime (s)
	2-noc (%)	2-all (%)	3-noc (%)	3-all (%)	4-noc (%)	4-all (%)	5-noc (%)	5-all (%)	
LEAStereo [59]	1.90	2.39	1.13	1.45	0.83	1.08	0.67	0.88	0.30
HITNet [60]	2.00	2.65	1.41	1.89	1.14	1.53	0.96	1.29	**0.02**
ACVNet [61]	1.83	2.34	1.13	1.47	0.86	1.12	0.71	0.91	0.20
CREStereo [16]	1.72	2.18	1.14	1.46	0.90	1.14	0.76	0.95	0.40
PCWNet [62]	1.69	2.18	1.04	1.37	0.78	1.01	0.63	0.81	0.44
IGEV-Stereo [12]	1.71	2.17	1.12	1.44	0.88	1.12	0.73	0.94	0.18
UCFNet [63]	1.67	2.17	1.09	1.45	0.85	1.12	0.69	0.91	0.21
ICVP [64]	1.72	2.21	1.06	1.39	0.80	1.05	0.66	0.86	0.17
MC-Stereo [57]	1.55	1.99	1.04	1.35	0.82	1.05	0.68	0.87	0.40
NMRF-Stereo [58]	1.59	2.07	1.01	1.35	0.78	1.03	0.64	0.84	0.09
StereoBase [65]	1.54	1.95	1.00	1.26	0.76	0.97	0.62	0.80	0.24
ViTAStereo (ours)	**1.46**	**1.80**	**0.93**	**1.16**	**0.71**	**0.87**	**0.58**	**0.71**	0.22

TABLE V

COMPARISONS WITH SOTA STEREO MATCHING NETWORKS PUBLISHED ON THE KITTI STEREO 2015 DATASET [52]. D1-BG, D1-FG, AND D1-ALL DENOTE D1 FOR BACKGROUND, FOREGROUND, AND ALL PIXELS, RESPECTIVELY. ALL VALUES ARE EXPRESSED IN PERCENTAGES (%).

Network	All Pixels			Non-Occluded Pixels		
	D1-bg	D1-fg	D1-all	D1-bg	D1-fg	D1-all
LEAStereo [59]	1.40	2.91	1.65	1.29	2.65	1.51
HITNet [60]	1.74	3.20	1.98	1.54	2.72	1.74
CREStereo [16]	1.45	2.86	1.69	1.33	2.60	1.54
ACVNet [61]	1.37	3.07	1.65	1.26	2.84	1.52
UCFNet [63]	1.57	3.33	1.86	1.41	2.93	1.66
GMStereo [12]	1.49	3.14	1.77	1.34	2.97	1.61
CroCo-Stereo [22]	1.38	2.65	1.59	1.30	2.56	1.51
IGEV-Stereo [12]	1.38	2.67	1.59	1.27	2.62	1.49
MC-Stereo [57]	1.36	2.51	1.55	1.24	2.55	1.46
NMRF-Stereo [58]	1.28	3.13	1.59	1.17	2.95	1.46
StereoBase [65]	1.28	**2.26**	**1.44**	1.17	**2.23**	**1.35**
ViTAStereo (ours)	**1.21**	2.99	1.50	**1.12**	2.90	1.41

excluding any single module from the complete ViTAS yields a performance deterioration comparable to the impact observed when the module is used in isolation. This observation underscores the modular independence within our ViTAS. Notably, the PAFM is identified as the most influential component, primarily attributed to its capability of aggregating both global and stereo contextual information throughout different feature layers, thereby underlining its significance in enhancing the model's overall performance.

Moreover, we compare our PAFM with two other meticulously designed multi-scale feature fusion methods: the CNN-based self-distilled feature aggregation (SDFA) [27] and the Transformer-based vertical attention fusion (VAF), to underscore the efficacy of PAFM. As shown in Table II, PAFM dramatically reduces the number of learnable parameters by 84.1% and 83.2% compared to SDFA and VAF, respectively. While PAFM has marginally higher memory requirements than SDFA, its memory demands are significantly lower—by 32.8%—than those of VAF. Table I further illustrates that ViTAStereo, equipped with PAFM, outperforms another two feature fusion methods in disparity estimation accuracy. Furthermore, our proposed PAFM achieves a feature fusion inference time that is similar but slightly higher compared to SDFA, and is 64.3% faster than VAF. These comprehensive experiments collectively demonstrate that PAFM achieves not only the highest disparity accuracy but also the fewest learnable parameters. In addition, it markedly reduces both computational complexity and memory consumption compared to VAF, demonstrating its exceptional capacity to minimize the trade-off between accuracy and efficiency.

We further validate the universality of our proposed ViTAS in using the general-purpose VFM features for stereo matching. Two recent VFMs, DINOv2 [8] and Depth Anything [9], are employed to build ViTAStereo and the quantitative results are presented in Table III. Although ViTAStereo built with DINOv2 achieves higher stereo matching accuracy across all metrics, the performance of ViTAStereo built with Depth Anything is only marginally lower. The minor performance gap between ViTAStereo built with DINOv2 and Depth Anything underscores the effectiveness of ViTAS in adapting general-purpose VFM features for stereo matching and its potential compatibility with future, more advanced VFMs.

D. Comparisons with SoTA Networks

Upon submitting our best results[1] (achieved without extensive hyperparameter tuning) to the KITTI Stereo 2012 and 2015 benchmark suites, we conduct a comparative analysis with other SoTA stereo matching networks published on these benchmarks. The results presented in Table IV suggest that on the KITTI Stereo 2012 dataset, our ViTAStereo outperforms all other SoTA stereo matching networks across all evaluation metrics. Surprisingly, ViTAStereo outperforms StereoBase [65], the second-best network, by up to 7.00% and 11.25% in PEP for non-occluded pixels and all pixels, respectively. Moreover, compared to IGEV-Stereo [11], which uses an identical stereo matching back-end structure, our ViTAStereo reduces PEP by up to 24.47%. These significant performance gains underscore the effectiveness of adapting a pre-trained VFM to

[1] These results can be accessed at https://cvlibs.net/datasets/kitti/eval_stereo_flow.php?benchmark=stereo and https://cvlibs.net/datasets/kitti/eval_scene_flow.php?benchmark=stereo.

TABLE VI
COMPARISONS AMONG SOTA STEREO MATCHING NETWORKS WITHOUT AND WITH OUR PROPOSED VITAS APPLIED.

Network	Dataset for Model Fine-tuning	ViTAS	KITTI Eval		Midd Eval		ETH3D	
			EPE (px)	D1-all (%)	EPE (px)	D1-all (%)	EPE (px)	D1-all (%)
IGEV-Stereo [11]	KITTI Train	w/o	0.55	1.71	5.27	16.8	1.15	5.23
		w/	**0.49**	**1.36**	**3.05**	**10.9**	**1.01**	**5.08**
	Midd Train	w/o	1.07	5.27	2.14	10.8	4.20	5.49
		w/	**0.87**	**3.45**	**1.34**	**6.00**	**2.65**	**3.68**
GMStereo [12]	KITTI Train	w/o	0.59	1.82	2.96	14.1	1.08	8.99
		w/	**0.56**	**1.62**	**2.65**	**12.4**	**0.82**	**2.30**
	Midd Train	w/o	1.14	6.14	2.65	13.3	1.72	10.1
		w/	**1.02**	**4.41**	**1.99**	**10.2**	**0.67**	**4.34**
CREStereo [16]	KITTI Train	w/o	0.70	2.32	5.20	**16.1**	3.45	18.8
		w/	**0.66**	**2.02**	**4.99**	16.6	**1.75**	**14.8**
	Midd Train	w/o	1.18	6.24	3.98	14.4	**28.6**	35.0
		w/	**1.07**	**4.91**	**2.59**	**12.7**	29.1	**25.9**

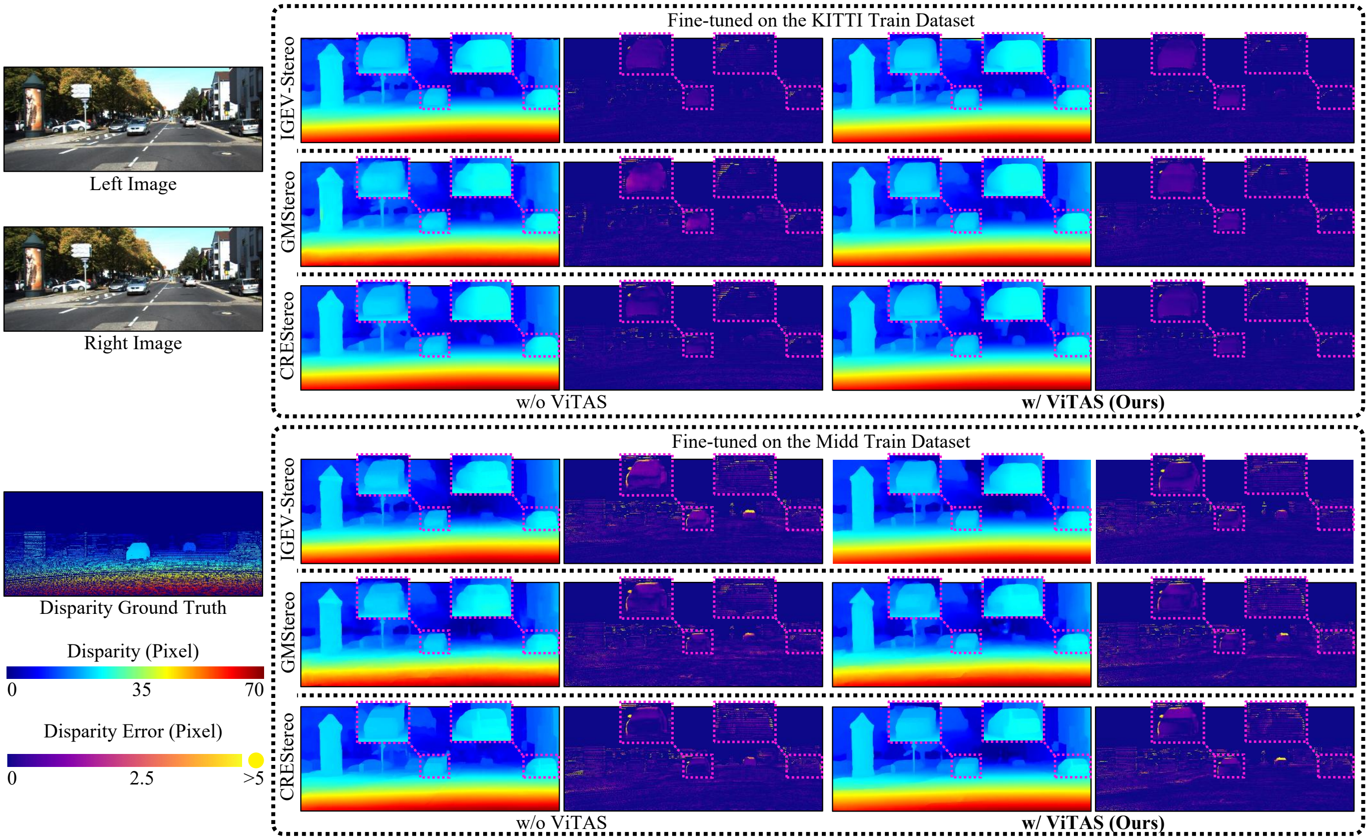


Fig. 5. Qualitative experimental results on the KITTI Eval dataset. Significantly improved regions are shown in pink dashed boxes.

stereo matching using ViTAS, as opposed to traditional CNN-based backbones for feature extraction. However, ViTAStereo shows a 22.2% increase in runtime compared to IGEV-Stereo. This increase is primarily due to the additional computations introduced by the VFM and ViTAS, as opposed to the original CNN-based feature extractor used in IGEV-Stereo. On the other hand, ViTAStereo remains highly efficient, with an 8.3% reduction in runtime compared to StereoBase, the second most accurate stereo matching network on the KITTI Stereo 2012 dataset. Furthermore, the results presented in Table V indicate that ViTAStereo achieves the best performance in terms of D1 in background areas (lower by approximately 5.47% compared to StereoBase and about 12.3% compared to IGEV-Stereo) on the KITTI Stereo 2015 dataset. Moreover, our ViTAStereo decreases D1-all by around 5.66% compared to IGEV-Stereo.

The qualitative experimental results on these two datasets, as illustrated in Fig. 4, also suggest that ViTAStereo outperforms other SoTA networks in handling challenging scenarios. This superior performance is evident in both large-scale, textureless regions (illustrated in rows 1 and 2) as well as small-scale

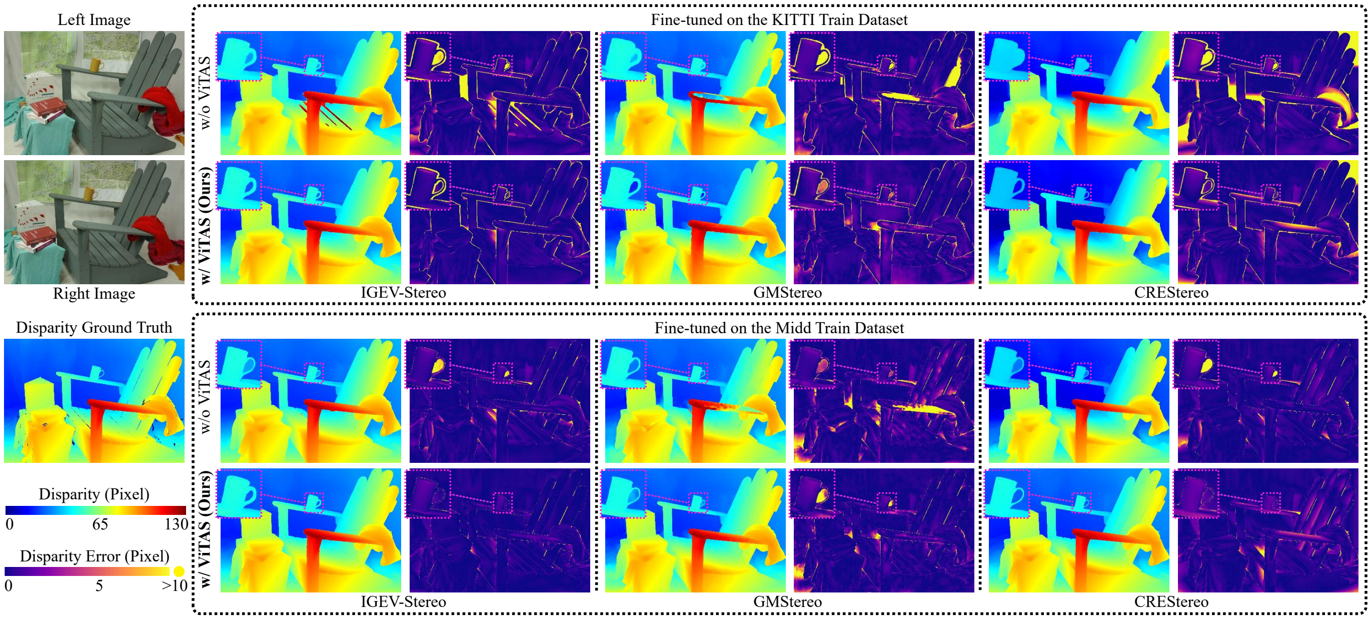


Fig. 6. Qualitative experimental results on the Midd Eval dataset. Significantly improved regions are shown in pink dashed boxes.

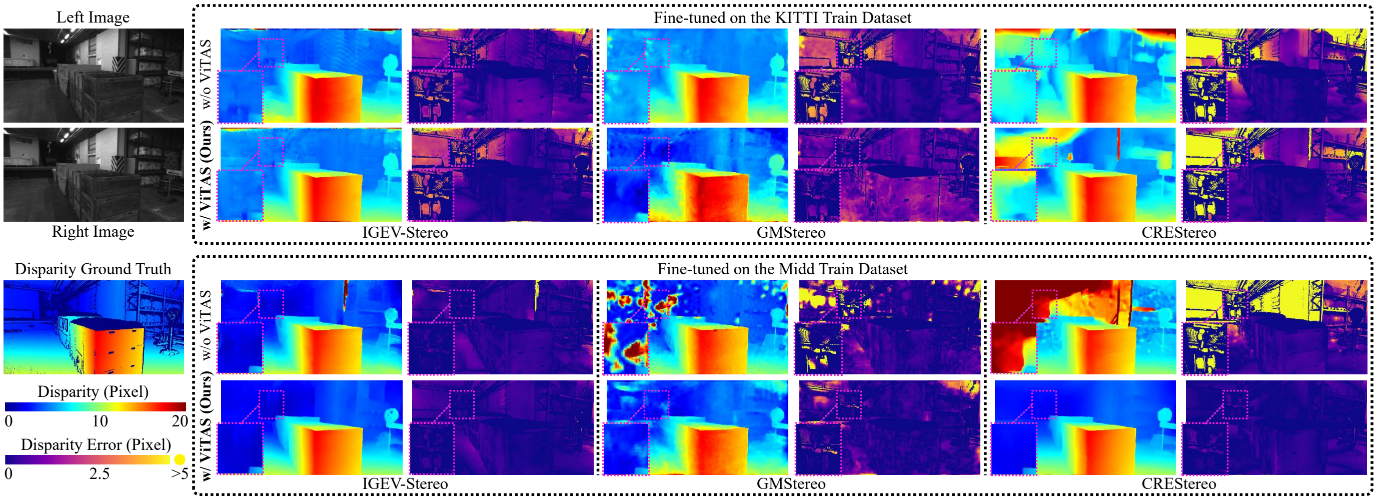


Fig. 7. Qualitative experimental results on the ETH3D dataset. Significantly improved regions are shown in pink dashed boxes.

areas rich in details (shown in rows 3 to 9). We attribute these improvements to the multi-scale feature aggregation process performed by our proposed PAFM. Additionally, ViTAStereo achieves improved disparity accuracy within occluded areas (shown in row 10), further validating the robustness of our approach in complex environments.

E. Generalizability Evaluation

VFMs are renowned for their remarkable generalizability across diverse scenarios. Therefore, we conduct a series of additional experiments on three public, real-world datasets to further evaluate the generalizability of our proposed ViTAS, when combined with three SoTA cost volume-based stereo matching networks. The quantitative results are given in Table VI, while the qualitative results on the KITTI Eval, Midd Eval, and ETH3D datasets are presented in Figs. 5, 6, and 7, respectively. Our results suggest that leveraging ViTAS for visual feature extraction enables both IGEV-Stereo and GMStereo to consistently achieve superior performance across all evaluation metrics on each dataset. While CREStereo achieves comparable, albeit slightly less favorable, results in terms of D1-all (when trained on the KITTI Train dataset and evaluated on the Midd Eval dataset) and EPE (when trained on the Midd Train dataset and evaluated on the ETH3D dataset), it gains performance improvements in the remaining experiments. These extensive and comprehensive experimental results validate the compatibility of our proposed ViTAS as well as its effectiveness in adapting to new, unseen datasets.

V. Conclusion and Future Work

This article introduced ViTAS, a pioneering research effort to fully exploit the general-purpose VFM features for stereo matching. Our study has yielded several key findings: (1) the prevalent use of CNN-based feature extractors in existing stereo matching networks has been identified as a critical bottleneck, limiting these networks from achieving higher levels of stereo matching accuracy; (2) a pre-trained VFM combined with an appropriate adapter, demonstrates superior performance in terms of both stereo matching accuracy and generalizability, when compared to conventional CNN-based backbones; (3) merely aggregating stereo contextual information via the cross-attention mechanism falls short in addressing the scale ambiguity problem, underscoring the indispensable role of cost volumes in developing generalizable stereo matching networks. Our ViTAStereo sets a new standard of performance on the KITTI Stereo datasets, establishing itself as the SoTA in this field. While the contributions of this study are significant, it is noted that the CAM utilized herein remains unchanged and still requires substantial computational and memory resources. Therefore, we intend to investigate more efficient strategies for the aggregation of stereo contextual information in the near future, to facilitate the future application of our ViTAStereo in other correspondence matching tasks and its deployment on mobile hardware devices.

References

[1] R. Fan *et al.*, "Pothole detection based on disparity transformation and road surface modeling," *IEEE Transactions on Image Processing*, vol. 29, pp. 897–908, 2019.

[2] R. Fan *et al.*, "Road surface 3D reconstruction based on dense subpixel disparity map estimation," *IEEE Transactions on Image Processing*, vol. 27, no. 6, pp. 3025–3035, 2018.

[3] Z. Wu *et al.*, "S^{3}M-Net: Joint learning of semantic segmentation and stereo matching for autonomous driving," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 2, pp. 3940–3951, 2024.

[4] R. Fan *et al.*, "Rethinking road surface 3-D reconstruction and pothole detection: From perspective transformation to disparity map segmentation," *IEEE Transactions on Cybernetics*, vol. 52, no. 7, pp. 5799–5808, 2022.

[5] Z. Huang *et al.*, "Online, target-free LiDAR-camera extrinsic calibration via cross-modal mask matching," *IEEE Transactions on Intelligent Vehicles*, 2024, DOI: 10.1109/TIV.2024.3456299.

[6] J. Li *et al.*, "HAPNet: Toward Superior RGB-Thermal Scene Parsing via Hybrid, Asymmetric, and Progressive Heterogeneous Feature Fusion," *IEEE Transactions on Intelligent Transportation Systems*, 2024, in press.

[7] A. Kirillov *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 4015–4026.

[8] M. Oquab *et al.*, "DINOv2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[9] L. Yang *et al.*, "Depth anything: Unleashing the power of large-scale unlabeled data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 10 371–10 381.

[10] X. Guo *et al.*, "Group-wise correlation stereo network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3273–3282.

[11] G. Xu *et al.*, "Iterative geometry encoding volume for stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 21 919–21 928.

[12] H. Xu *et al.*, "Unifying flow, stereo and depth estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 11, pp. 13 941–13 958, 2023.

[13] K. He *et al.*, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[14] M. Sandler and pthers, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4510–4520.

[15] B. Huang *et al.*, "H-Net: Unsupervised attention-based stereo depth estimation leveraging epipolar geometry," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 4460–4467.

[16] J. Li *et al.*, "Practical stereo matching via cascaded recurrent network with adaptive correlation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 16 263–16 272.

[17] W. Guo *et al.*, "Context-enhanced stereo Transformer," in *European Conference on Computer Vision (ECCV)*. Springer, 2022, pp. 263–279.

[18] Z. Liu *et al.*, "Global occlusion-aware Transformer for robust stereo matching," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024, pp. 3535–3544.

[19] Y. Li *et al.*, "Benchmarking detection transfer learning with vision Transformers," *arXiv preprint arXiv:2111.11429*, 2021.

[20] Y. Li *et al.*, "Exploring plain vision Transformer backbones for object detection," in *European Conference on Computer Vision (ECCV)*. Springer, 2022, pp. 280–296.

[21] Z. Chen *et al.*, "Vision Transformer adapter for dense predictions," in *International Conference on Learning Representations (ICLR)*, 2023.

[22] P. Weinzaepfel *et al.*, "CroCo v2: Improved cross-view completion pre-training for stereo matching and optical flow," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 17 969–17 980.

[23] P. Weinzaepfel *et al.*, "CroCo: Self-supervised pre-training for 3D vision tasks by cross-view completion," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, pp. 3502–3516, 2022.

[24] R. Ranftl *et al.*, "Vision Transformers for dense prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 12 179–12 188.

[25] L. Lipson *et al.*, "RAFT-Stereo: Multilevel recurrent field transforms for stereo matching," in *2021 International Conference on 3D Vision (3DV)*. IEEE, 2021, pp. 218–227.

[26] Z. Shen *et al.*, "CFNet: Cascade and fused cost volume for robust stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 13 906–13 915.

[27] Z. Zhou and Q. Dong, "Self-distilled feature aggregation for self-supervised monocular depth estimation," in *European Conference on Computer Vision (ECCV)*. Springer, 2022, pp. 709–726.

[28] Y. Zhao *et al.*, "Semantic-aligned fusion Transformer for one-shot object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 7601–7611.

[29] A. Geiger *et al.*, "Are we ready for autonomous driving? the KITTI vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 3354–3361.

[30] Z. Xie *et al.*, "Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 16 684–16 693.

[31] X. Wang *et al.*, "Dense contrastive learning for self-supervised visual pre-training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 3024–3033.

[32] J. Sun *et al.*, "LoFTR: Detector-free local feature matching with Transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 8922–8931.

[33] J. Deng *et al.*, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2009, pp. 248–255.

[34] H. Wang *et al.*, "PVStereo: Pyramid voting module for end-to-end self-supervised stereo matching," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4353–4360, 2021.

[35] W. Wang *et al.*, "PVT v2: Improved baselines with pyramid vision Transformer," *Computational Visual Media*, vol. 8, no. 3, pp. 415–424, 2022.

[36] Y. Quan *et al.*, "Centralized feature pyramid for object detection," *IEEE Transactions on Image Processing*, vol. 32, pp. 4341–4354, 2023.

[37] Z. Ma *et al.*, "Multiview stereo with cascaded epipolar raft," in *European Conference on Computer Vision (ECCV)*. Springer, 2022, pp. 734–750.

[38] G. Xu *et al.*, "Accurate and efficient stereo matching via attention concatenation volume," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–13, 2023.

[39] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2021.

[40] N. Park and S. Kim, "How do vision Transformers work?" in *International Conference on Learning Representations (ICLR)*, 2021.

[41] Y. Fang *et al.*, "Unleashing vanilla vision Transformer with masked image modeling for object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 6244–6253.

[42] G. Wang *et al.*, "Cross-level attentive feature aggregation for change detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023, DOI: 10.1109/TCSVT.2023.3344092.

[43] R. Fan *et al.*, "SNE-RoadSeg: Incorporating surface normal information into semantic segmentation for accurate freespace detection," in *European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 340–356.

[44] J.-R. Chang *et al.*, "Attention-aware feature aggregation for real-time stereo matching on edge devices," in *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2020.

[45] W. Liu *et al.*, "Learning to upsample by learning to sample," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 6027–6037.

[46] Y. Liu *et al.*, "The devil is in the upsampling: Architectural decisions made simpler for denoising with deep image prior," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 12 408–12 417.

[47] J. Hu *et al.*, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7132–7141.

[48] L. Gao *et al.*, "Doubly-fused ViT: Fuse information from vision Transformer doubly with local representation," in *European Conference on Computer Vision (ECCV)*. Springer, 2022, pp. 744–761.

[49] Q. Su and S. Ji, "ChiTransformer: Towards reliable stereo from cues," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 1939–1949.

[50] N. Mayer *et al.*, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4040–4048.

[51] Y. Cabon *et al.*, "Virtual KITTI 2," *arXiv preprint arXiv:2001.10773*, 2020.

[52] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3061–3070.

[53] D. Scharstein and C. Pal, "Learning conditional random fields for stereo," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2007, pp. 1–8.

[54] H. Hirschmuller and D. Scharstein, "Evaluation of cost functions for stereo matching," in *2007 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2007, pp. 1–8.

[55] D. Scharstein *et al.*, "High-resolution stereo datasets with subpixel-accurate ground truth," in *Pattern Recognition: 36th German Conference (GCPR)*. Springer, 2014, pp. 31–42.

[56] T. Schops *et al.*, "A multi-view stereo benchmark with high-resolution images and multi-camera videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3260–3269.

[57] M. Feng *et al.*, "MC-Stereo: Multi-peak lookup and cascade search range for stereo matching," in *International Conference on 3D Vision (3DV)*. IEEE, 2024, pp. 344–353.

[58] T. Guan *et al.*, "Neural markov random field for stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 5459–5469.

[59] X. Cheng *et al.*, "Hierarchical neural architecture search for deep stereo matching," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 22 158–22 169, 2020.

[60] V. Tankovich *et al.*, "HITnet: Hierarchical iterative tile refinement network for real-time stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 14 362–14 372.

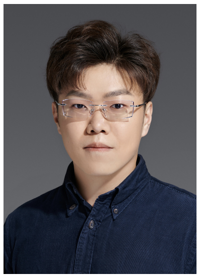
[61] G. Xu *et al.*, "Attention concatenation volume for accurate and efficient stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 12 981–12 990.

[62] Z. Shen *et al.*, "PCW-Net: Pyramid combination and warping cost volume for stereo matching," in *European Conference on Computer Vision (ECCV)*. Springer, 2022, pp. 280–297.

[63] Z. Shen *et al.*, "Digging into uncertainty-based pseudo-label for robust stereo matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 12, pp. 14 301–14 320, 2023.

[64] O.-H. Kwon and E. Zell, "Image-coupled volume propagation for stereo matching," in *2023 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2023, pp. 2510–2514.

[65] X. Guo *et al.*, "OpenStereo: A comprehensive benchmark for stereo matching and strong baseline," *arXiv preprint arXiv:2312.00343*, 2023.



Chuang-Wei Liu received his B.E. degree in automation from Tongji University in 2020. He is currently pursing his Ph.D. degree, supervised by Prof. Rui Fan, with the Machine Intelligence and Autonomous Systems (MIAS) Group in the Robotics and Artificial Intelligence Laboratory (RAIL) at Tongji University. His research interests include computer stereo vision, especially for unsupervised approaches, and long-term learning. He is currently a student member of IEEE.

Qijun Chen (Senior Member, IEEE) received the B.S. degree in automation from Huazhong University of Science and Technology, Wuhan, China, in 1987, the M.S. degree in information and control engineering from Xi'an Jiaotong University, Xi'an, China, in 1990, and the Ph.D. degree in control theory and control engineering from Tongji University, Shanghai, China, in 1999. He is currently a Full Professor in the College of Electronics and Information Engineering, Tongji University, Shanghai, China. His research interests include robotics control, environmental perception, and understanding of mobile robots and bioinspired control.

Rui Fan (Senior Member, IEEE) received the B.Eng. degree in Automation from the Harbin Institute of Technology in 2015 and the Ph.D. degree (supervisors: Prof. John G. Rarity and Prof. Naim Dahnoun) in Electrical and Electronic Engineering from the University of Bristol in 2018. He worked as a Research Associate (supervisor: Prof. Ming Liu) at the Hong Kong University of Science and Technology from 2018 to 2020 and a Postdoctoral Scholar-Employee (supervisors: Prof. Linda M. Zangwill and Prof. David J. Kriegman) at the University of California San Diego between 2020 and 2021. He began his faculty career as a Full Research Professor with the College of Electronics & Information Engineering at Tongji University in 2021, and was then promoted to a Full Professor in the same college, as well as at the Shanghai Research Institute for Intelligent Autonomous Systems in 2022.

Prof. Fan served as an associate editor for ICRA'23 and IROS'23/24, an area chair for ICIP'24, and a senior program committee member for AAAI'23/24/25. He is the general chair of the AVVision community and organized several impactful workshops and special sessions in conjunction with WACV'21, ICIP'21/22/23, ICCV'21, and ECCV'22. He was honored by being included in the Stanford University List of Top 2% Scientists Worldwide in both 2022 and 2023, recognized on the Forbes China List of 100 Outstanding Overseas Returnees in 2023, and acknowledged as one of Xiaomi Young Talents in 2023. His research interests include computer vision, deep learning, and robotics, with a specific focus on humanoid visual perception under the two-streams hypothesis.