All-in-One: Transferring Vision Foundation Models into Stereo Matching

Jingyi Zhou[1*], Haoyu Zhang[1*], Jiakang Yuan[1*], Peng Ye[1,3,4†], Tao Chen[1†], Hao Jiang[2], Meiya Chen[2], Yangyang Zhang[2]

[1]School of Information Science and Technology, Fudan University, China
[2]Xiaomi Inc., Beijing, China
[3]Shanghai AI Laboratory, Shanghai, China
[4]The Chinese University of Hong Kong

Abstract

As a fundamental vision task, stereo matching has made remarkable progress. While recent iterative optimization-based methods have achieved promising performance, their feature extraction capabilities still have room for improvement. Inspired by the ability of vision foundation models (VFMs) to extract general representations, in this work, we propose AIO-Stereo which can flexibly select and transfer knowledge from multiple heterogeneous VFMs to a single stereo matching model. To better reconcile features between heterogeneous VFMs and the stereo matching model and fully exploit prior knowledge from VFMs, we proposed a dual-level feature utilization mechanism that aligns heterogeneous features and transfers multi-level knowledge. Based on the mechanism, a dual-level selective knowledge transfer module is designed to selectively transfer knowledge and integrate the advantages of multiple VFMs. Experimental results show that AIO-Stereo achieves start-of-the-art performance on multiple datasets and ranks 1st on the Middlebury dataset and outperforms all the published work on the ETH3D benchmark.

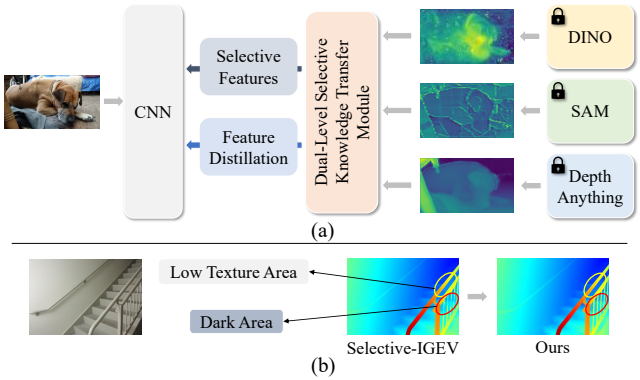

Figure 1: (a) The overview of AIO-Stereo which transfers selected knowledge from multiple VFMs to a single stereo matching model. (b) Comparisons between Selective-IGEV and our AIO-Stereo in dark and low texture areas.

Introduction

With the development of 3D vision tasks and their applications such as robotics and autonomous driving, stereo matching (Li et al. 2022; Chang and Chen 2018a) becomes a fundamental vision task since it can provide depth information in the real 3D world. Stereo matching models typically predict the pixel-wise displacement (*i.e.*, disparity) between a pair of rectified images and further decode the depth information with the camera calibration.

Benefiting from the success of deep learning, some works begin to explore learning-based methods (Kendall et al. 2017; Xu and Zhang 2020). As a milestone, PSMNet (Chang and Chen 2018a) utilizes 3D convolution to regularize a 4D cost volume and boost the performance. However, such learning-based methods need large computational costs. Recently, iterative optimization-based methods (Li et al. 2022; Xu et al. 2023a; Zhao et al. 2023) have shown great potential on stereo matching tasks by progressively updating the disparity map. Selective-Stereo (Wang et al. 2024) proposes selective recurrent unit and contextual spatial attention module to further improve the ability to predict detailed areas.

Despite the great improvement in performance, the general feature extraction capabilities of the existing models are relatively weak which can be attributed as follows: (1) Recent state-of-the-art (SOTA) works (*e.g.*, Selective-Stereo (Wang et al. 2024)) mainly focus on refining iterative update mechanisms while partially ignoring the quality of encoded features. Meanwhile, the task-oriented optimization objectives also make it difficult for the encoder to learn global and contextual information. (2) The amount of stereo matching data is relatively small and most of them are synthetic data. It is difficult for models to learn general representations from insufficient data. For example, as shown in Fig. 1 (b), previous methods (*e.g.*, Selective-IGEV (Wang et al. 2024)) fail to predict the depth information in dark areas with vague texture features since features in these areas are highly similar. As a result, disparity prediction based on feature matching between pixels is prone to have large mismatches. Recently, vision foundation models (VFMs) have emerged and shown promising performance on various tasks. These VFMs are trained on large-scale datasets and can extract general representations which motivates us to consider injecting the general feature extraction capabili-

*These authors contributed equally.
†Corresponding authors.

ties of multiple VFMs into the stereo matching models.

However, directly transferring knowledge from multiple VFMs to the single stereo matching model is not easy which is mainly caused by the following two reasons. (1) Most of the existing VFMs are based on Transformer-architecture while the stereo matching models are often based on CNN. The heterogeneity of model architectures will lead to feature mismatch when simply merging or distilling the intermediate features. (2) Different VFMs have different attention to feature representations due to the various training data, methods, and tasks. For example, as shown in 1 (a), DINO (Caron et al. 2021; Oquab et al. 2023) which are pretrained in a self-supervised manner, tend to extract global semantic information. In contrast, large segmentation models (Kirillov et al. 2023; Wang et al. 2023) represented by SAM (Kirillov et al. 2023) pay more attention to capturing the semantic information of small objects and edges. As a result, directly using the features of multiple vision foundation models without selecting will cause feature conflicts.

Based on the observation and analysis, we claim that the quality of encoded features is equally crucial for the stereo matching task, as they constitute the main sources of information for the iterative update modules, directly influencing every step of the iteration process. To this end, we propose an efficient knowledge transfer framework, named AIO-Stereo, that sifts and learns advantageous knowledge from multiple VFMs to obtain sufficiently effective and informative features. To transfer the knowledge from heterogeneous VFMs effectively and take full advantage of different VFMs, we develop a dual-level knowledge utilization module to bridge the gap between misaligned features and transfer multi-level knowledge. Furthermore, considering that the features derived from multiple VFMs are vastly divergent and potentially conflicting, a dual-level selective knowledge transfer module is proposed to selectively transfer knowledge and fully leverage the strengths of each VFM. Our contribution can be summarized as follows:

- To enhance the general understanding of stereo networks, we first propose to leverage the diverse and general knowledge of multiple vision foundation models for stereo matching.
- We proposed a flexible knowledge transfer framework, named AIO-stereo, which consists of dual-level knowledge utilization and a selective knowledge transfer module that can effectively and efficiently transfer the multi-level knowledge from multiple heterogeneous vision foundation models to a single stereo matching model.
- Experimental results show that the proposed AIO-Stereo ranks 1^{st} on the Middlebury dataset and outperforms the published methods on the ETH3D benchmark.

Related Work

Stereo Matching

As a difficult pixel-level 3D task, stereo matching has been studied for a long time and early works primarily utilize traditional matching algorithms (Boykov, Veksler, and Zabih 2001; Klaus, Sormann, and Karner 2006; Sun, Zheng, and Shum 2003; Yang et al. 2008; Hirschmüller, Innocent, and Garibaldi 2002; Van Meerbergen et al. 2002; Hirschmuller 2005). Since Zbontar and LeCun (Zbontar and LeCun 2015) first introduced CNN to calculate the matching cost, traditional matching algorithms have gradually been replaced by learning-based methods (Kendall et al. 2017; Xu and Zhang 2020; Mayer et al. 2016). PSMNet (Chang and Chen 2018b) incorporates contextual information with 3D convolution and used feature concatenation to construct 4D cost volume. HITNet (Tankovich et al. 2021a) proposes a fast multi-resolution initialization step, differentiable 2D geometric propagation, and warping mechanisms that take both speed and accuracy into account. More recently, iterative optimization-based methods (Lipson, Teed, and Deng 2021; Xu et al. 2023a) have shown great potential in stereo matching tasks. Inspired by (Teed and Deng 2020), RAFT-Stereo (Lipson, Teed, and Deng 2021) first explores the iteration of multi-scale update blocks to generate the final disparity map from coarse to fine. IGEV-Stereo (Xu et al. 2023b) applies additional Geometry Encoding Volume to supplement the missing non-local geometry knowledge. CREStereo (Li et al. 2022) designs an adaptive group correlation layer and releases a new large-scale, high-quality synthetic dataset. Selective-Stereo (Wang et al. 2024) proposes a selective recurrent unit and a contextual spatial attention module that can better capture details. However, current works mainly focus on designing the iterative process and relatively ignore the feature extraction ability of encoders, which is also important in the stereo matching task.

Vision Foundation Models

In recent years, thanks to the improvement of hardware performance and the construction of large-scale datasets, vision foundation models (VFMs) with extremely high performance have emerged. These VFMs can process and understand image or video information effectively and facilitate the development of other vision tasks. In image-level tasks, for example, CLIP (Radford et al. 2021) which is trained on image-text pairs shows a strong zero-shot classification capability. In pixel-level tasks, Depth Anything family (Yang et al. 2024a,b) demonstrates a strong generalization ability in different depth estimation scenarios. SAM (Kirillov et al. 2023) explores a semi-supervised pipeline and achieves promising object category-agnostic segmentation capabilities. Besides, lots of studies (Oquab et al. 2023; Darcet et al. 2023; Liu et al. 2021) introduce more general backbone networks through pre-training. As a representative, DINO family (Caron et al. 2021; Oquab et al. 2023) explores self-supervised learning on vision transformer and boosts the performance on various downstream tasks. Although these VFMs have shown great generalization and zero-shot ability, how to effectively utilize them to improve the performance of stereo matching still remains unexplored.

Knowledge Distillation

Knowledge distillation (KD) was first proposed by Hinton (Hinton, Vinyals, and Dean 2015) and has been widely used in model compression (Kim, Park, and Kwak 2018; Bai et al. 2020) and knowledge transfer (Komodakis and Zagoruyko 2017). Recently, with the expansion of model

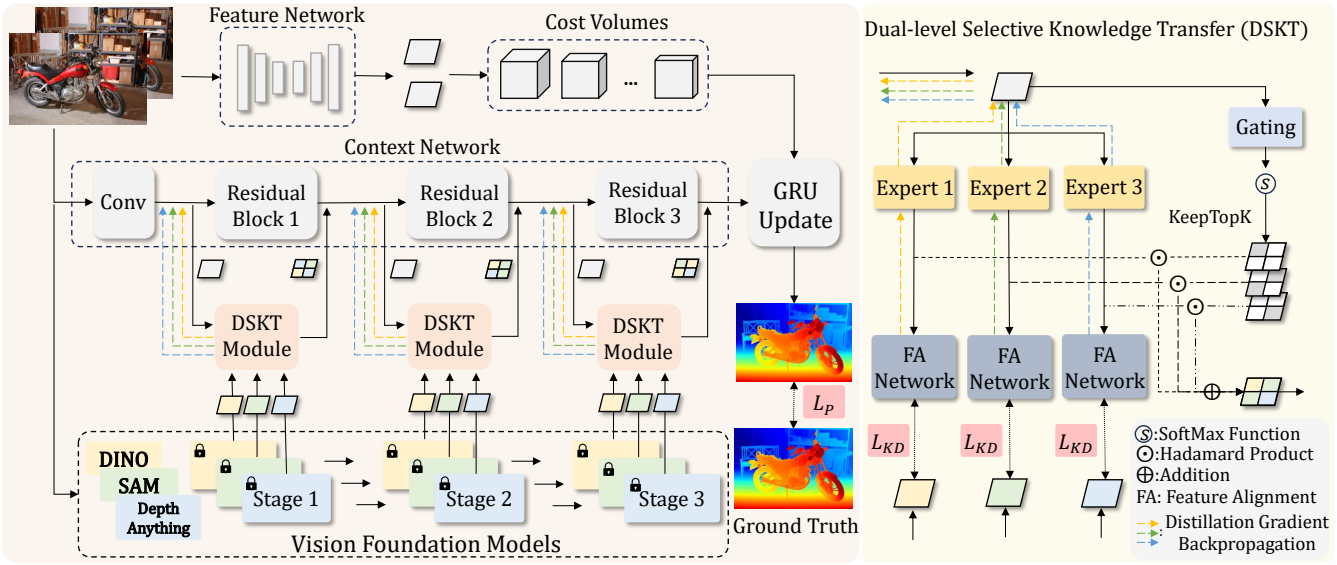

Figure 2: The Overall framework of AIO-Stereo. Left: AIO-Stereo selectively learns knowledge from SAM, DINO and Depth Anything by the proposed dual-level selective knowledge transfer module. Right: The detailed structure of our proposed dual-level selective knowledge transfer module.

zoos and the development of large models, learning knowledge from multiple teachers and heterogeneous teachers has attracted increasing attention. Following the trend, researchers began to study multi-teacher distillation (Mehak and Balasubramanian 2018; Fukuda et al. 2017) and distillation from heterogeneous architectures (Shen et al. 2019; Touvron et al. 2021; Hao et al. 2024). To take advantage of multiple teachers, FEED (Park and Kwak 2019) and Knowledge Flow (Liu, Peng, and Schwing 2019) add non-linear transformations to align the features between the student model and multiple teacher models. Besides, since models of different architecture (*i.e.*, CNN, ViT, MLP) have their own distinct inductive bias, directly distilling knowledge between heterogeneous models will result in significant degradation of the performance. To handle such a problem, OFA-KD (Hao et al. 2024) transfers the mismatched feature representations into the aligned logits which contain less architecture-aware information. In this paper, we explore the knowledge transfer problem from multiple VFMs to a single stereo matching model for the first time.

Method

In this section, we first introduce the VFMs we utilized. Then, we analyze the challenges in transferring abundant knowledge from multiple heterogeneous VFMs to single stereo matching model. Finally, we detail our AIO-Stereo, which can flexibly transfer the required knowledge from multiple VFMs for the stereo matching task.

Preliminaries: VFMs

DINO (Caron et al. 2021; Oquab et al. 2023) is built upon the Vision Transformer (ViT) (Dosovitskiy et al. 2020) architecture. In the training process, DINO aligns the feature representations of the same image with different augmentation methods in a self-supervised learning manner. As a result, DINO is encouraged to develop invariant and robust feature representations, especially the foreground areas and salient regions of the input image.

SAM (Kirillov et al. 2023) is an image segmentation model that includes a ViT-based image encoder, a prompt encoder, and a lightweight mask decoder. The model is trained on a large amount of semi-supervised data in which the unlabeled data is automatically annotated by the model. Thus, features in SAM demonstrate strong zero-shot generalization to unfamiliar objects and images. Besides, as a segmentation model, SAM can extract abundant representations of diverse objects and edges and generate the segmentation mask in any location of the image according to the prompt.

Depth Anything family (Yang et al. 2024a,b) is a series of VFMs for monocular depth estimation, leveraging the Transformer architecture. The model is trained on a wide variety of supervised and extensive unlabeled data, which captures depth information at multiple scales. Therefore, Depth Anything exhibits impressive generalization abilities across diverse data. It adeptly captures subtle visual cues to differentiate depth variations between objects and their surroundings, especially in dark and low-texture areas.

Challenges of Transferring Knowledge from Multiple Heterogeneous Vision Foundation Models

Heterogeneity between Stereo Matching Models and various vision foundation models Based on the characteristics of the stereo matching task, as well as the considerations of computation cost and inference speed, most of the stereo matching models are built upon CNN structures, while vision foundation models predominantly adopt the

Transformer architecture. As revealed in (Hao et al. 2024), features from heterogeneous models reside in different latent spaces. For example, features in CNN have a strong inductive bias of locality and spatial invariance, while features in Transformer represent more global and contextual information with the self-attention module. Consequently, simply merging or distilling features between vision foundation models and stereo matching models is unsuitable and could potentially prevent the feature learning of the stereo matching network. In the experiment process, it has a negative impact (*i.e.*, -0.71 in terms of EPE) on our results and causes unstable performance.

Knowledge Discrepancies and Conflicts Among Vision Foundation Models Despite the strong generalization and semantic understanding capability of VFMs, they still have their independent characteristics due to different training methodologies and optimization objectives. As mentioned above, DINO pays more attention to the foreground areas of objects. SAM is better at extracting features of small objects and edge areas by training on the segmentation tasks. Depth Anything, the large monocular depth estimation model trained on various datasets, achieves superior generalization across data from different sources. Thus, features in the model excel at discerning subtle visual cues to infer the relative depth changes on the surfaces of objects.

Therefore, each VFM pays attention to different areas of the image and there exist differences and conflicts between features and knowledge from different VFMs as shown in Fig. 1. As a result, indiscriminate acceptance of knowledge from multiple VFMs can lead not only to potential interferences among the disparate sources of knowledge but might also impede the model's learning process and convergence.

AIO-Stereo

In this section, we detail AIO-Stereo, a simple but effective method that can transfer knowledge from multiple VFMs flexibly and selectively. To take advantage of VFMs, a dual-level knowledge utilization is designed to effectively transfer knowledge between heterogeneous models, and a selective knowledge transfer module is proposed to integrate knowledge from multiple VFMs into a single stereo matching model which will be introduced later.

Overview of AIO-Stereo As shown in Fig. 2, the feature extractor can be divided into a feature network to calculate cost volumes and a context network to generate context features for GRU refinement. Specifically, the context network contains three residual blocks with each block consisting of a series of residual and downsampling layers. Our dual-level selective knowledge transfer (DLSKT) module is adapted to transfer the rich knowledge of the vision foundation models to the context network. In detail, given the left images or the right images $I^{l(r)} \in \mathbb{R}^{3\times H\times W}$ as inputs, local features $c^{l(r)}$ are extracted by the feature network and the pixel-wise correlation volume can be calculated by:

$$Corr(x, y, z) = \langle c^l(x, y), c^r(x - z, y)\rangle, \tag{1}$$

where $\langle\cdot,\cdot\rangle$ is the inner product. Meanwhile, context features can be obtained by the context network taking the left images I^l as the input. In this process, a single layer convolution with a kernel size of 7 is first applied to the image to get the original features $f_0 \in \mathbb{R}^{C_0\times H\times W}$, where C_0 is the number of channels. Then, the original features are fed into the residual blocks, with knowledge gradually learned from the VFMs by the DLSKT module. Specifically, a series of intermediate features $f_i \in \mathbb{R}^{C_i\times \frac{H}{2^{i-1}}\times \frac{W}{2^{i-1}}} (i = 1, 2, 3)$ can be obtained, where C_i is the feature channels. For the i^{th} block, f_{i-1} is taken as the input, and f_i is the output feature. Following RAFT-Stereo (Lipson, Teed, and Deng 2021), correlation volumes at four resolutions are obtained by average pooling. Since the context features of the left image and the right image are semantically similar, the correlation volumes and the context features of the left image are injected into every step of the GRU updating operator to refine the disparity map step by step. With a series of intermediate predictions $\{\mathbf{p}_i\}_{i=1}^N$, the prediction loss can be calculated following (Lipson, Teed, and Deng 2021):

$$L_P = \sum_{i=1}^{N} \gamma_P^{N-i} ||\mathbf{p}_i - \mathbf{p}_{GT}||_1, \tag{2}$$

where γ_P is the decay factor and $\mathbf{p}_{GT}$ is the ground truth.

Dual-Level Knowledge Utilization To better integrate knowledge from the heterogeneous VFM into our stereo matching model, we propose dual-level knowledge utilization which utilizes the knowledge in both distillation and fusion levels. For simplicity and better understanding, we take a single large model, SAM, as an example to show our knowledge utilization method. As mentioned above, since VFMs and our conv-based backbone are heterogeneous, direct feature distillation or feature fusion is not appropriate since the features are in different latent spaces and will result in misalignment. To bridge the architectural gap between SAM and the CNN backbone, we integrate an expert network and a feature alignment network to align the features between heterogeneous models and transfer general representations from the VFM to the stereo matching model. Specifically, a lightweight expert network E_i^s is designed and takes the intermediate features f_i as the input to get the expert features e_i^s which can be calculated as follows:

$$e_i^s = E_i^s(f_i \mid \varphi_i^s), \tag{3}$$

where φ_i^s denotes the parameter of the expert network. Then, the expert features e_i^s subsequently pass through a heavier feature alignment network A_s^i which reconciles the feature space of the stereo network with that of SAM. A distillation loss is finally designed to make the stereo matching model learn from the VFM in the aligned feature space which can be written as follows:

$$L_{KD,i}^s = \mathbf{MSE}(A_i^s(e_i^s \mid \theta_i^s), \mathcal{F}^s(s_i)), \tag{4}$$

where s_i represents the features extracted from the i^{th} stage of SAM, $\mathcal{F}^s$ denotes the interpolate function to align the resolutions, and θ_i^s is the parameter of feature alignment network. During the distillation process, the knowledge derived from SAM is propagated to the expert and the backbone

network through the feature alignment network. A heavier feature alignment network can reduce the misalignment among features, but may also lead to greater knowledge attrition. Thus, we initialize the feature alignment network with a higher learning rate to rapidly learn the mapping relationships within the feature space and a larger learning rate decay factor to prevent excessive retention of knowledge within the network. Through the distillation process, knowledge is conveyed to the preceding blocks by backpropagation, while the forward propagation of the current block may attenuate the knowledge to some extent. To enhance the knowledge, we add the output of the expert network back into the output of the of the current residual block:

$$f_{i+1} = B_i(f_i \mid \zeta_i) + e_i^s, \tag{5}$$

where B_i is the residual block with ζ_i to be its parameter. This dual-level design ensures a more effective knowledge transfer from the heterogeneous vision foundation model.

Selective Knowledge Transfer To facilitate effective knowledge transfer from multiple distinct VFMs while preventing knowledge interference between different models, we design a dual-level selective knowledge transfer module. Inspired by the mechanism of the mixture of experts (MoE) (Jacobs et al. 1991), we employ a selective knowledge transfer mechanism with a trainable gating network to select the experts for the features of each pixel. Specifically, for the i^{th} block, we extract the features from the i^{th} stage of DINO, SAM, and Depth Anything, represented as d_i, s_i, and a_i respectively. As shown in Fig. 2, each VFM is associated with an expert network that learns the knowledge from the corresponding VFM through distillation. Taking f_i as inputs, e_i^d, e_i^s, e_i^a can be obtained by the expert networks E_i^d, E_i^s, and E_i^a respectively. The distillation loss for multiple VFMs can be calculated as follows:

$$\begin{aligned} L_{KD,i} &= \sum_{x \in \{d,s,a\}} L_{KD,i}^x \\ &= \sum_{x \in \{d,s,a\}} \mathbf{MSE}(A_i^x(e_i^x), \mathcal{F}^x(x_i)). \end{aligned} \tag{6}$$

$$g_i = \text{KeepTopK}(\text{Softmax}(G_i(f_i) \mid \psi_i), k, dim = 0), \tag{7}$$

where k is the number of experts to be retained at each pixel, and $\text{KeepTopK}(\cdot, k, \cdot)$ is a function that keeps the top-k values with the highest weight at the specific dimension. Subsequently, we use selection weights to determine the importance of the features from three experts, selectively fusing and discarding certain features. Finally, the selected features are then added to the output of the current residual block as the input of the next block.

$$f_{i+1} = B_i(f_i \mid \zeta_i) + \sum_{x \in \{d,s,a\}} e_i^x \odot g_i(x) \tag{8}$$

By selecting the features from experts, we indirectly choose the knowledge from multiple VFMs at different regions of the image to transfer the most beneficial and effective knowledge from each large model.

Loss Function The overall loss function consists of the prediction loss L_P and the distillation loss L_{KD}. The L1 Loss is calculated between the predicted disparity maps and the ground truth. And the MSE Loss is applied for knowledge distillation. The final loss is defined as:

$$L_{AIO} = L_P + L_{KD} = L_P + \sum_{j=1}^{3} \gamma_{KD}^{4-j} L_{KD,j}, \tag{9}$$

where γ_{KD} is the decay factor and L_P, $L_{KD,j}$ are defined in Eq. (2) and Eq. (4), respectively.

Experiments

Datasets

Following Selective-Stereo (Wang et al. 2024), we verify the effectiveness of AIO-Stereo on four widely used datasets including Scene Flow (Mayer et al. 2016), Middlebury 2014 (Scharstein et al. 2014), KITTI-2015 (Menze and Geiger 2015) and ETH3D (Schops et al. 2017). Scene Flow (Mayer et al. 2016) contains more than 39000 synthetic stereo frames which are divided into training and testing set. Middlebury 2014 (Scharstein et al. 2014) provides a training set with images of 23 indoor scenes and a testing set with images of 10 indoor scenes, and both sets have three resolutions to use. KITTI-2015 (Menze and Geiger 2015) contains 200 training pairs and 200 testing pairs with sparse disparity maps which were collected in real-world driving scenes. ETH3D (Schops et al. 2017) provides gray-scale image pairs covering both indoor and outdoor scenes.

Implementation Details

We implement our AIO-Stereo with Pytorch framework and perform our experiments using NVIDIA A100 GPUs while using the AdamW optimizer. For pre-training, we trained our model on the augmented Scene Flow training set (*i.e.*, both cleanpass and finalpass) for 200k steps with a batch size of 8, and we use a random crop size of 320 $\times$ 720. We use a one-cycle learning rate schedule with warm up strategy and the learning rate gradually increases to 0.0002 in the first 1% of steps and gradually decreases thereafter. And for finetune, the learning rate linearly decays from 0.0003 to 0.

Ablation Study

Ablation for Each VFM AIO-Stereo leverages the knowledge from three VFMs (*i.e.*, DINOv2, SAM, and Depth Anything v2) to enhance feature representation and improve overall accuracy. To verify our method can effectively integrate the advantage of multiple VFMs, we conduct experiments on using different numbers of VFMs (*i.e.*, only DINO, DINO and SAM, DINO, SAM and Depth Anything). As shown in the upper part of Tab. 1, performance improves when using VFMs which verifies that AIO-Stereo can effectively learn from VFMs. Besides, as the number of used VFMs increases, the performance is consistently improved. This is because each VFM has its unique advantages and knowledge from different VFMs can be integrated effectively by our AIO-Stereo. Moreover, the results highlight the

Method	Distillation	Forward	Selection	DINOv2	SAM	Depth Anything v2	EPE (px)	>2px (%)
Baseline							0.74	4.68
w/o Selection	✓	✓		✓	✓	✓	0.68	3.57
w/o Distillation		✓	✓	✓	✓	✓	0.72	3.87
w/o Forward Fusion	✓			✓	✓	✓	0.67	3.52
Only DINO	✓	✓	✓	✓			0.66	3.64
DINO+SAM	✓	✓	✓	✓	✓		0.68	3.61
full*	✓	✓	✓	✓	✓	✓	**0.66**	**3.48**

Table 1: Ablation study of proposed networks on the MiddEval v3 training set in full resolution and all metrics are on all pixels. The baseline is the Selevtive-IGEV (Wang et al. 2024).* means the final version of our method.

Method	Middlebury				ETH3D				KITTI 2015		
	bad1.0	bad2.0	avgerr	A90	bad0.5	bad1.0	bad2.0	avgerr	D1-bg	D1-fg	D1-all
PSMNet (2018a)	63.9	42.1	6.68	17.0	-	-	-	-	1.86	4.62	2.32
HITNet (2021b)	13.3	6.46	1.71	2.32	7.83	2.79	0.80	0.20	1.54	2.72	1.74
RAFT-Stereo (2021)	9.37	4.74	1.27	1.10	7.04	2.44	0.44	0.18	1.75	2.89	1.96
LEAStereo (2020)	20.8	7.15	1.43	1.68	-	-	-	-	1.74	3.20	1.98
CREStereo (2022)	8.25	3.71	1.15	0.92	3.58	0.98	0.22	0.13	1.45	2.86	1.69
GMStereo (2023c)	23.6	7.14	1.31	1.64	5.94	1.83	0.25	0.19	1.49	3.14	1.77
IGEV-Stereo (2023a)	9.41	4.83	2.89	4.87	3.52	1.12	**0.21**	0.14	1.38	2.67	1.59
DLNR (2023)	6.82	3.20	1.06	0.85	-	-	-	-	1.60	2.59	1.76
Selective-IGEV (2024)	6.53	2.51	0.91	0.79	3.06	1.23	0.22	**0.12**	**1.33**	2.61	1.55
AIO-Stereo (Ours)	**6.08**	**2.36**	**0.85**	**0.76**	**2.91**	**0.94**	**0.21**	0.13	1.35	**2.46**	**1.54**

Table 2: Quantitative evaluation on Middlebury(Scharstein et al. 2014), ETH3D (Schops et al. 2017), and KITTI 2015 (Menze and Geiger 2015). **Bold**: Best. Underline: Second best.

inherent flexibility of our AIO-Stereo, which is not dependent on a single foundation model but is designed to effectively orchestrate multiple models, leveraging their strengths to serve our stereo matching task. It suggests that our AIO-Stereo can flexibly take advantage of various VFMs.

Effectiveness of Dual-level Knowledge Utilization To effectively transfer the knowledge between heterogeneous models, we use the knowledge in both distillation and fusion levels. In this section, we evaluate the effectiveness of our dual-level approach to knowledge utilization from VFMs. In detail, we first exclude the aligned knowledge distillation to eliminate the effect of extra parameters brought by the expert networks. It can be observed from Tab. 1 that the performance decreases largely (*i.e.* 3.48 to 3.87 on the 2 pixels error index) which is because the model will be unable to learn from the knowledge of VFMs without the distillation process. Besides, there is also a performance drop without the forward fusion. This is because the model is more prone to forgetting the knowledge it has acquired, leading to a diminished effect in knowledge transfer. By utilizing knowledge at both levels, our approach achieves a more comprehensive and rich transfer of visual knowledge.

Exploration of DLSKT In this section, we explore the expert selection mechanism of our DLSKT module. In particular, we apply a non-selective knowledge transfer which accepts all the knowledge from VFMs indiscriminately to compare with our selective knowledge transfer. The results in Tab. 1 (*i.e.*, w/o selection) indicate that the non-selective knowledge transfer underperforms our selective methods on both EPE and 2 pixels error index, attributed to the incompatible and sometimes contradictory knowledge among different VFMs.

Comparisons with State-of-the-art

To evaluate the effectiveness of our method, we compare AIO-Stereo with the current SOTA methods on the Middlebury, ETH3D, and KITTI 2015 datasets as shown in Tab. 2. Note that AIO-Stereo ranks 1^{st} on the Middlebury leaderboard and achieves SOTA on multiple datasets.

Middlebury. For the Middlebury dataset, following (Wang et al. 2024), we first finetune our pre-trained model on the mixed Tartan Air (Wang et al. 2020), CREStereo Dataset (Li et al. 2022), Scene Flow, Falling things (Tremblay, To, and Birchfield 2018), InStereo2k (Bao et al. 2020), CARLA HR-VS (Yang et al. 2019), and Middlebury datasets 200k steps using a crop size of 384 $\times$ 512 with a batch size of 8. Then we finstune it on the mixed CREStereo Dataset, Falling Things, InStereo2k, CARLA HR-VS, and Middlebury datasets using a crop size of 384 $\times$ 768 with a batch size of 8 for another 100k steps. As shown

Method	F		H	
	EPE	D1	EPE	D1
PSMNet (2018a)	40.51	57.93	9.79	32.19
RAFT-Stereo (2021)	**3.84**	15.64	1.44	11.21
IGEV-Stereo (2023a)	5.87	11.85	1.36	7.21
GMStereo (2023c)	4.10	29.15	1.92	15.69
EAI-Stereo (2022)	6.16	18.25	2.15	11.74
DLNR (2023)	6.57	14.46	1.45	9.46
Selective-IGEV (2024)	5.28	12.07	1.35	7.31
AIO-Stereo (Ours)	4.16	**11.67**	**0.89**	**6.48**

Table 3: Zero-shot evaluation on Middlebury. **Bold**: Best. Underline: Second best.

in Tab. 2, our method achieves SOTA performance on the Middlebury test set. Specifically, our method surpasses Selective-IGEV (Wang et al. 2024) and DLNR (Zhao et al. 2023) by 5.98% and 26.25% on the bad 2 pixels error respectively without extra design to the refinement process, demonstrating the effectiveness of our designs.

ETH3D. For the ETH3D dataset, following (Wang et al. 2024), we finetune the pre-trained model on the mixed Tartan Air, CREStereo Dataset, Scene Flow, Sintel Stereo (Butler et al. 2012), InStereo2k, and ETH3D datasets for 300k steps. Then we fintune it on the mixed CREStereo Dataset, InStereo2k, and ETH3D datasets for another 90k steps. Our method achieves the best performance among all published methods for most metrics, and outperforms Selective-IGEV (Wang et al. 2024) by 23.58% on bad 1.0 metric. quantitative results are shown in Tab 2.

KITTI-2015. For the KITTI-2015 dataset, following (Wang et al. 2024), we finetune the pretrained model on the mixed dataset of KITTI-2012 (Geiger, Lenz, and Urtasun 2012) and KITTI-2015 with a batch size of 8 for 50k steps. As shown in Tab. 2, our method achieves comparable results and surpasses Selective-IGEV by 5.75% on D1-fg metric.

Zero-Shot Generalization

To evaluate the generalization capabilities of our proposed method, we pre-train our model on the synthetic Scene Flow dataset and directly test it on the Middlebury dataset, an unseen real-world dataset with challenging indoor scenes. As shown in Tab. 3, AIO-Stereo achieves state-of-the-art performance at most of the metrics. Attributed to the knowledge transferred from the vision foundation models, our method performs well on unseen environments.

Visualization

Visual Comparisons on Middlebury Further, we compare the visualization results with other works (*i.e.*, RAFT-Stereo (Lipson, Teed, and Deng 2021), CREStereo (Li et al. 2022), and IGEV-Stereo (Xu et al. 2023a)). It can be seen from Fig. 3 that AIO-Stereo can achieve better visualization quality, especially in texture and dark areas. This is because our method can take advantage of multiple VFMs and learns general representations from them.

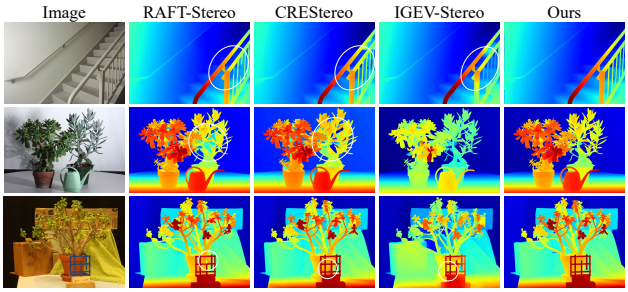

Figure 3: Visual comparison on the Middlebury dataset.

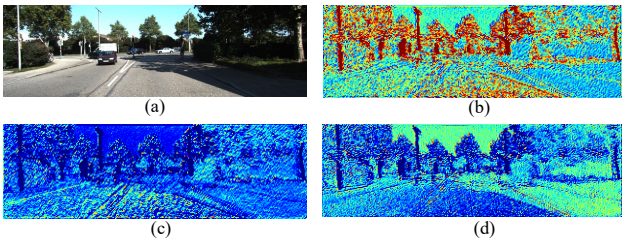

Figure 4: Visualization of the selection weights for each VFM. (a) Reference image. (b-d) Selection weights of DINO, SAM, and Depth Anything respectively.

Visualization of Selection Weights for Each Expert Our method indirectly selects different VFMs by selecting different experts. We visualize the selection weight for each VFM in Fig. 4 to show our strategy of selecting from different regions based on the independent strengths of various models. The visualization indicates a clear preference for foreground regions when integrating features from DINO, attributable to its enhanced focus and robustness within these regions. For SAM, features are mainly selected on the edges, aligning with SAM's capability for identifying differences between objects. As for Depth Anything, our model learns features of dark and low-texture areas, where Depth Anything performs well. The visualization results further verify that AIO-Stereo can combine the advantage of different VFMs.

Conclusion

In this paper, for the first time, we explore leveraging the knowledge of VFMs to improve the performance of stereo matching. Specifically, we propose AIO-Stereo which can transfer knowledge from multiple VFMs into a single stereo matching model. Our AIO-Stereo combines the advantages of multiple VFMs by selective knowledge transfer module and effectively adapts the knowledge from heterogeneous VFMs to our stereo matching model by dual-level knowledge utilization module. Experimental results show that our AIO-Stereo achieves SOTA performance on multiple datasets and rank 1^{st} on the Middlebury dataset. Further, with the knowledge of VFMs, our method shows strong generalization capability on ETH3D dataset.

Acknowledgments

This work is supported by Shanghai Natural Science Foundation (No. 23ZR1402900), National Key Research and Development Program of China (No. 2022ZD0160101), Shanghai Municipal Science and Technology Major Project (No.2021SHZDZX0103). The computations in this research were performed using the CFFF platform of Fudan University.

References

Bai, H.; Wu, J.; King, I.; and Lyu, M. 2020. Few shot network compression via cross distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 3203–3210.

Bao, W.; Wang, W.; Xu, Y.; Guo, Y.; Hong, S.; and Zhang, X. 2020. Instereo2k: a large real dataset for stereo matching in indoor scenes. *Science China Information Sciences*, 63: 1–11.

Boykov, Y.; Veksler, O.; and Zabih, R. 2001. Fast approximate energy minimization via graph cuts. *IEEE Transactions on pattern analysis and machine intelligence*, 23(11): 1222–1239.

Butler, D. J.; Wulff, J.; Stanley, G. B.; and Black, M. J. 2012. A naturalistic open source movie for optical flow evaluation. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI 12*, 611–625. Springer.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging Properties in Self-Supervised Vision Transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

Chang, J.-R.; and Chen, Y.-S. 2018a. Pyramid stereo matching network. In *CVPR*, 5410–5418.

Chang, J.-R.; and Chen, Y.-S. 2018b. Pyramid stereo matching network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5410–5418.

Cheng, X.; Zhong, Y.; Harandi, M.; Dai, Y.; Chang, X.; Li, H.; Drummond, T.; and Ge, Z. 2020. Hierarchical neural architecture search for deep stereo matching. *NeurIPS*, 33: 22158–22169.

Darcet, T.; Oquab, M.; Mairal, J.; and Bojanowski, P. 2023. Vision Transformers Need Registers.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fukuda, T.; Suzuki, M.; Kurata, G.; Thomas, S.; Cui, J.; and Ramabhadran, B. 2017. Efficient Knowledge Distillation from an Ensemble of Teachers. In *Interspeech*, 3697–3701.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 3354–3361. IEEE.

Hao, Z.; Guo, J.; Han, K.; Tang, Y.; Hu, H.; Wang, Y.; and Xu, C. 2024. One-for-All: Bridge the Gap Between Heterogeneous Architectures in Knowledge Distillation. *Advances in Neural Information Processing Systems*, 36.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Hirschmuller, H. 2005. Accurate and efficient stereo processing by semi-global matching and mutual information. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, 807–814. IEEE.

Hirschmüller, H.; Innocent, P. R.; and Garibaldi, J. 2002. Real-time correlation-based stereo vision with reduced border errors. *International Journal of Computer Vision*, 47: 229–246.

Jacobs, R. A.; Jordan, M. I.; Nowlan, S. J.; and Hinton, G. E. 1991. Adaptive Mixtures of Local Experts. *Neural Computation*, 3(1): 79–87.

Kendall, A.; Martirosyan, H.; Dasgupta, S.; Henry, P.; Kennedy, R.; Bachrach, A.; and Bry, A. 2017. End-to-end learning of geometry and context for deep stereo regression. In *ICCV*, 66–75.

Kim, J.; Park, S.; and Kwak, N. 2018. Paraphrasing complex network: Network compression via factor transfer. *Advances in neural information processing systems*, 31.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643*.

Klaus, A.; Sormann, M.; and Karner, K. 2006. Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 3, 15–18. IEEE.

Komodakis, N.; and Zagoruyko, S. 2017. Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. In *ICLR*.

Li, J.; Wang, P.; Xiong, P.; Cai, T.; Yan, Z.; Yang, L.; Liu, J.; Fan, H.; and Liu, S. 2022. Practical stereo matching via cascaded recurrent network with adaptive correlation. In *CVPR*, 16263–16272.

Lipson, L.; Teed, Z.; and Deng, J. 2021. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *3DV*, 218–227. IEEE.

Liu, I.-J.; Peng, J.; and Schwing, A. G. 2019. Knowledge flow: Improve upon your teachers. *arXiv preprint arXiv:1904.05878*.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Mayer, N.; Ilg, E.; Hausser, P.; Fischer, P.; Cremers, D.; Dosovitskiy, A.; and Brox, T. 2016. 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*, 4040–4048.

Mehak, M.; and Balasubramanian, V. N. 2018. *Knowledge distillation from multiple teachers using visual explanations*. Ph.D. thesis, Indian Institute of Technology Hyderabad.

Menze, M.; and Geiger, A. 2015. Object scene flow for autonomous vehicles. In *CVPR*, 3061–3070.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H. V.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; Howes, R.; Huang, P.-Y.; Xu, H.; Sharma, V.; Li, S.-W.; Galuba, W.; Rabbat, M.; Assran, M.; Ballas, N.; Synnaeve, G.; Misra, I.; Jegou, H.; Mairal, J.; Labatut, P.; Joulin, A.; and Bojanowski, P. 2023. DINOv2: Learning Robust Visual Features without Supervision.

Park, S.; and Kwak, N. 2019. Feed: Feature-level ensemble for knowledge distillation. *arXiv preprint arXiv:1909.10754*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Scharstein, D.; Hirschmüller, H.; Kitajima, Y.; Krathwohl, G.; Nešić, N.; Wang, X.; and Westling, P. 2014. High-resolution stereo datasets with subpixel-accurate ground truth. In *GCPR*, 31–42. Springer.

Schops, T.; Schonberger, J. L.; Galliani, S.; Sattler, T.; Schindler, K.; Pollefeys, M.; and Geiger, A. 2017. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *CVPR*, 3260–3269.

Shen, C.; Xue, M.; Wang, X.; Song, J.; Sun, L.; and Song, M. 2019. Customizing student networks from heterogeneous teachers via adaptive knowledge amalgamation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3504–3513.

Sun, J.; Zheng, N.-N.; and Shum, H.-Y. 2003. Stereo matching using belief propagation. *IEEE Transactions on pattern analysis and machine intelligence*, 25(7): 787–800.

Tankovich, V.; Hane, C.; Zhang, Y.; Kowdle, A.; Fanello, S.; and Bouaziz, S. 2021a. Hitnet: Hierarchical iterative tile refinement network for real-time stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14362–14372.

Tankovich, V.; Hane, C.; Zhang, Y.; Kowdle, A.; Fanello, S.; and Bouaziz, S. 2021b. Hitnet: Hierarchical iterative tile refinement network for real-time stereo matching. In *CVPR*, 14362–14372.

Teed, Z.; and Deng, J. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 402–419. Springer.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, 10347–10357. PMLR.

Tremblay, J.; To, T.; and Birchfield, S. 2018. Falling things: A synthetic dataset for 3d object detection and pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2038–2041.

Van Meerbergen, G.; Vergauwen, M.; Pollefeys, M.; and Van Gool, L. 2002. A hierarchical symmetric stereo algorithm using dynamic programming. *International Journal of Computer Vision*, 47: 275–285.

Wang, W.; Zhu, D.; Wang, X.; Hu, Y.; Qiu, Y.; Wang, C.; Hu, Y.; Kapoor, A.; and Scherer, S. 2020. Tartanair: A dataset to push the limits of visual slam. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 4909–4916. IEEE.

Wang, X.; Xu, G.; Jia, H.; and Yang, X. 2024. Selective-stereo: Adaptive frequency information selection for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19701–19710.

Wang, X.; Zhang, X.; Cao, Y.; Wang, W.; Shen, C.; and Huang, T. 2023. Seggpt: Segmenting everything in context. *arXiv preprint arXiv:2304.03284*.

Xu, G.; Wang, X.; Ding, X.; and Yang, X. 2023a. Iterative Geometry Encoding Volume for Stereo Matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21919–21928.

Xu, G.; Wang, X.; Ding, X.; and Yang, X. 2023b. Iterative Geometry Encoding Volume for Stereo Matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21919–21928.

Xu, H.; and Zhang, J. 2020. Aanet: Adaptive aggregation network for efficient stereo matching. In *CVPR*, 1959–1968.

Xu, H.; Zhang, J.; Cai, J.; Rezatofighi, H.; Yu, F.; Tao, D.; and Geiger, A. 2023c. Unifying flow, stereo and depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yang, G.; Manela, J.; Happold, M.; and Ramanan, D. 2019. Hierarchical deep stereo matching on high-resolution images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5515–5524.

Yang, L.; Kang, B.; Huang, Z.; Xu, X.; Feng, J.; and Zhao, H. 2024a. Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data. In *CVPR*.

Yang, L.; Kang, B.; Huang, Z.; Zhao, Z.; Xu, X.; Feng, J.; and Zhao, H. 2024b. Depth Anything V2. *arXiv:2406.09414*.

Yang, Q.; Wang, L.; Yang, R.; Stewénius, H.; and Nistér, D. 2008. Stereo matching with color-weighted correlation, hierarchical belief propagation, and occlusion handling. *IEEE transactions on pattern analysis and machine intelligence*, 31(3): 492–504.

Zbontar, J.; and LeCun, Y. 2015. Computing the stereo matching cost with a convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1592–1599.

Zhao, H.; Zhou, H.; Zhang, Y.; Chen, J.; Yang, Y.; and Zhao, Y. 2023. High-Frequency Stereo Matching Network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1327–1336.

Zhao, H.; Zhou, H.; Zhang, Y.; Zhao, Y.; Yang, Y.; and Ouyang, T. 2022. EAI-stereo: Error aware iterative network for stereo matching. In *Proceedings of the Asian Conference on Computer Vision*, 315–332.