

Learning Robust Stereo Matching in the Wild with Selective Mixture-of-Experts

Yun Wang¹, Longguang Wang², Chenghao Zhang³, Yongjian Zhang²,
Zhanjie Zhang⁴, Ao Ma⁵, Chenyou Fan⁶, Tin Lun Lam^{7,‡}, Junjie Hu^{7,‡}

¹City University of Hong Kong, ²Shenzhen Campus, Sun Yat-sen University,

³Chinese Academy of Sciences, ⁴Zhejiang University, ⁵JD.com,

⁶South China Normal University, ⁷The Chinese University of Hong Kong, Shenzhen

ywang3875-c@my.cityu.edu.hk, wanglg9@mail.sysu.edu.cn, zhangchenghao18@mails.ucas.ac.cn,

zhangyj85@mail2.sysu.edu.cn, cszzj@zju.edu.cn, maaaoama@126.com,

fanchenyou@gmail.com, {tllam, hujunjie}@cuhk.edu.cn

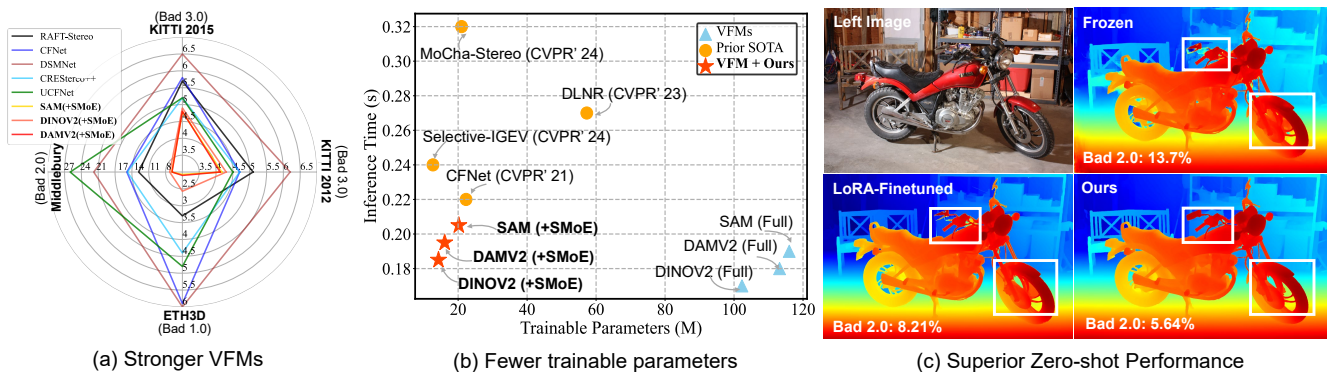


Figure 1. Our SMoEStereo largely outperforms previous SoTA domain generalized methods (a). SMoEStereo leverages Visual Foundation Models (VFMs) through a parameter- and computationally-efficient fine-tuning framework, enabling robust stereo matching (b). The proposed framework integrated into VFMs (e.g., DAMV2 [74]) demonstrates superior generalization versus other fine-tuning methods (c).

Abstract

Recently, learning-based stereo matching networks have advanced significantly. However, they often lack robustness and struggle to achieve impressive cross-domain performance due to domain shifts and imbalanced disparity distributions among diverse datasets. Leveraging Vision Foundation Models (VFMs) can intuitively enhance the model’s robustness, but integrating such a model into stereo matching cost-effectively to fully realize their robustness remains a key challenge. To address this, we propose SMoEStereo, a novel framework that adapts VFMs for stereo matching through a tailored, scene-specific fusion of Low-Rank Adaptation (LoRA) and Mixture-of-Experts (MoE) modules. SMoEStereo introduces MoE-LoRA with adaptive ranks and MoE-Adapter with adaptive kernel sizes. The former dynamically selects optimal experts within MoE to adapt varying scenes across domains, while the latter injects inductive bias into frozen VFMs to improve geometric feature extraction. Importantly, to mitigate computational overhead,

we further propose a lightweight decision network that selectively activates MoE modules based on input complexity, balancing efficiency with accuracy. Extensive experiments demonstrate that our method exhibits state-of-the-art cross-domain and joint generalization across multiple benchmarks without dataset-specific adaptation. The code is available at <https://github.com/cocowyl/SMoE-Stereo>.

1. Introduction

Stereo matching is a critical vision task focused on identifying pixel-wise correspondences between rectified stereo images. It has significant applications in autonomous driving [5], robot navigation [40], and augmented reality [58]. While recent learning-based stereo matching methods have shown impressive performance on standard benchmarks, their generalizability across diverse datasets remains limited. This is primarily due to significant scene differences and unbalanced disparity distributions across diverse datasets in the wild, potentially leading to noisy and distorted feature maps [22, 77], compromising the robustness

[‡] Corresponding author.

of current stereo matching models.

To enhance the robustness of feature representations against domain shifts in stereo matching, we aim to leverage the recent advancements in Vision Foundation Models (VFMs). These models, such as DepthAnythingV2 [74] for monocular depth estimation and SegmentAnything [24] for segmentation, are built on Vision Transformers (ViTs) pre-trained on large-scale diverse datasets. Such VFMs have demonstrated great effectiveness in providing robust, general-purpose deep features in many vision tasks. However, fully unleashing their potential for robust cross-domain stereo matching remains limited, and thus two intuitive limitations emerged: First, directly applying VFMs for stereo matching exhibits limited zero-shot performance. Although VFMs are highly effective at extracting semantic information from tasks such as single-image segmentation/classification/regression, they struggle to generate discriminative features necessary for precise similarity measurements in dense cross-view feature matching [34, 67, 79], as illustrated in Fig. 1 (c). Second, existing fine-tuning methods such as Low-Rank Adaptation [18] or small decoder [79] struggle to address the varying complexity of real-world in-the-wild stereo scenes, as they tend to employ a uniform low-rank subspace or a fixed CNN decoder across all inputs, treating vastly heterogeneous domains with rigid, one-size-fits-all feature refinement. This inflexibility limits their capacity to dynamically adjust to scene-specific characteristics, resulting in suboptimal generalization for in-the-wild scenarios [7, 53, 66].

To tackle the above problems, we propose incorporating Low-Rank Adaptation (LoRA) with varying ranks into a Mixture-of-Experts (MoE) design to adapt VFMs for robust stereo matching with minimal costs. Rather than naively using LoRA with a fixed rank, which limits its adaptability to diverse scenes for robust stereo matching, we extend conventional LoRA by developing a scene-conditioned selection mechanism that dynamically selects optimal low-rank subspaces from a spectrum of predefined rank values, enabling adaptive feature refinement tailored to in-the-wild scene characteristics. We also observe that frozen VFMs, even with learnable LoRA layers, lack intrinsic Inductive Bias, essential for modeling local visual structures for stereo matching [14, 30]. To bridge this gap, we introduce Inductive Bias by integrating Convolutional Neural Networks (CNNs) into each ViT block. Similarly, we embed MoE modules with several CNN adapters of varying receptive fields in each ViT block to incorporate Inductive Bias, while enhancing the model’s ability to capture local geometric structures. This hybrid design achieves complementary feature learning: CNN branches emphasize fine-grained local geometric details, while LoRA pathways model long-range interactions. As a result, it reduces stereo matching D1 error up to 30% versus vanilla VFM-LoRA

baselines, significantly improving model robustness.

Besides, integrating MoE into all ViT blocks introduces redundancy and extra costs, posing a critical bottleneck for deployment efficiency, a paramount concern in stereo matching applications. To this end, we introduce a lightweight decision network into each MoE layer. This network predicts binary decisions to activate MoE modules, saving computational costs by discarding redundant modules for simple samples and using more modules for complex ones. The decision network is jointly optimized with our MoE modules, incorporating a usage loss to measure computational costs and encourage policies that reduce redundancy while maintaining accuracy. The hyperparameter $\gamma \in (0, 1]$ can regulate the overall computational budget by scaling the activated MoE modules’ computational load proportionally, enabling flexible adaptation to varying resource constraints, critical for mobile devices with diverse computation demands [61].

Since both MoE modules and the experts within each MoE can be *selectively* activated to adapt to different input characteristics, we refer to our method as SMOEStereo. This method fully unleashes the potential of existing VFMs for stereo matching, as illustrated in Fig. 1. Extensive experiments show that SMOEStereo exhibits strong robustness, achieving state-of-the-art cross-domain generalization performance on the KITTI, Middlebury, ETH3D, and DrivingStereo datasets. Additionally, it achieves state-of-the-art joint generalization on the ETH3D, KITTI, and Middlebury benchmarks using the same trained model without any adaptation. The key contributions are as follows:

- We present SMOEStereo, an efficient yet powerful approach that leverages pre-trained Vision Foundation Models for robust stereo matching with minimal cost.
- We propose integrating MoE LoRA with varying ranks and MoE Adapter layers with varying kernel sizes into VFMs. These tailored designs facilitate scene-specific adaptation, enabling robust stereo matching across diverse real-world scenarios.
- We design a lightweight decision network integrated within each MoE module, which dynamically selects the relevant modules and deactivates less critical ones. This mechanism balances model accuracy and efficiency, enabling flexible adaptation to diverse resource constraints.
- Our approach exhibits strong cross-domain generalization and excels across multiple benchmarks using the same fixed model without further adaptation, significantly outperforming previous robust counterparts.

2. Related Work

2.1. Robust Stereo Matching

Recently, following the success of RAFT [57] in optical flow tasks, iterative approaches such as RAFTStereo [32],

IGEVStereo [27], and Selective-IGEV [60] have set new benchmarks by iteratively updating the disparity field through correlation volume sampling. Despite advancements in model design, achieving robust performance across varied scenarios remains challenging.

To tackle this, robust stereo matching methods have been increasingly studied, broadly divided into two categories: 1) Cross-domain Generalization: This category focuses on the network’s generalization to unseen scenes. Previous works try to tackle the problem by introducing domain normalization [76], leveraging pre-trained features on ImageNet [33], or developing diverse training strategies to learn domain-invariant features [4, 9, 78]. 2) Joint Generalization: This category aims to push the network to perform consistently well on various datasets without retraining. CFNet [53] and its improved UCFNet [55] introduce a cascaded and fused cost volume-based network to handle domain differences. CREStereo++ [22] introduces an uncertainty-guided adaptive warping module to enhance the robustness of the recurrent network in different scenarios. LoS [29] integrates structure information to improve model performance in challenging areas. However, these robust methods typically rely on classic feature extractor backbones such as ResNet [16], UNet [48], and Feature Pyramid Network [31], which suffer from limited receptive fields. Consequently, the potential of more powerful Vision Foundation Models (VFMs) based on Vision Transformers (ViTs) in robust stereo matching remains unexplored.

2.2. Parameter-Efficient Fine-tuning (PEFT)

Visual Foundation Models (VFMs) are defined as base vision models trained on large-scale data through self/semi-supervised learning, designed for adaptation to downstream vision tasks [1]. Recently, they [24, 41, 47, 59, 74] have emerged as a solution to enhance model discriminability and robustness. Given the substantial computational costs of fully fine-tuning VFMs, Parameter Efficient Fine-Tuning (PEFT) has emerged as a promising alternative. Notable examples include Visual Prompt Tuning (VPT) [20], which augments inputs with extra learnable tokens; AdaptFormer [6], proposing AdaptMLP to replace the MLP block; Adapter-tuning [34, 79], adding lightweight decoder modules; and Low-Rank Adaptation (LoRA) [17], injecting trainable rank decomposition matrices into transformer layers. Among these, the work of Zhang et al. [79] is most closely related to ours. They adapt VFMs by proposing a feature adapter to obtain robust features. However, our SMOE differs in two key aspects: (1) Instead of a fixed CNN decoder, SMOE uses rank-adaptive LoRA and kernel-adaptive CNNs to encode local and global cues, enabling scene-specific adaptation; (2) Unlike their computationally rigid pipeline requiring sequential processing, SMOE dynamically skips non-critical MoE modules for simple

scenes, boosting efficiency and real-world applicability.

2.3. Mixture of Experts (MoE)

Mixture-of-Experts (MoE) [19, 52] is designed to expand model capacity while introducing small computational overhead. Sparse MoE [52] introduces a router to select a subset of experts, using the gating network to regulate sparsity for computational savings. Feed Forward Networks (FFN) are commonly employed as the default choice for experts [2, 12, 46, 52, 83]. Recently, several studies combine MoE with multiple uniform LoRAs into large language/vision models by developing diverse routing mechanisms [11, 26, 35, 36, 75], enabling multi-task learning.

While previous works on MoE aim to expand model capacity or enable multi-task learning, our SMOE focuses on dynamically selecting optimal experts to adapt to diverse in-the-wild scenes, thereby enhancing robust stereo matching. Overall, our work differs from prior studies in three aspects: 1) Unlike homogeneous MoE experts, our SMOE employs heterogeneous experts with varying-rank LoRA or varying-kernel CNN configurations to handle diverse in-the-wild scenarios. 2) While MoE is mostly employed during pre-training, we integrate it as a parameter-efficient tuning mechanism for stereo matching task. 3) We novelly introduce a decision network to selectively activate the most suitable MoE modules, enhancing efficiency and applicability for stereo matching.

3. Methodology

3.1. Overview

RAFT-Stereo [32] is used as our backbone. We replace its feature extractor with VFMs while the remaining structures are unchanged. Specifically, the network first extracts stereo features using the well-designed VFM from stereo pairs. We modify the ViT blocks by integrating the designed MoE LoRA and Adapter layer with the vanilla blocks. A shallow CNN block¹ is then used to compress the feature dimension and enhance the locality. Then, a correlation volume pyramid is generated by computing the inner product along the epipolar lines. A multi-level GRU network is subsequently introduced to perform cost aggregation by recurrently updating the disparity field from the correlation volume pyramid. A sketch of the proposed pipeline is shown in Fig. 2.

3.2. Selective Mixture of Expert (SMoE)

As described in Sec. 1, the proposed Selective MoE integrates MoE LoRA and MoE Adapter layers within each ViT block. Besides, a layer decision network is inserted before each of the MoE layers, and it is trained to produce reasonable usage policies for MoE layer selection. In this way, our SMOEStereo learns to adaptively choose 1) which

¹The shallow CNN block comprises several CNN residual layers.

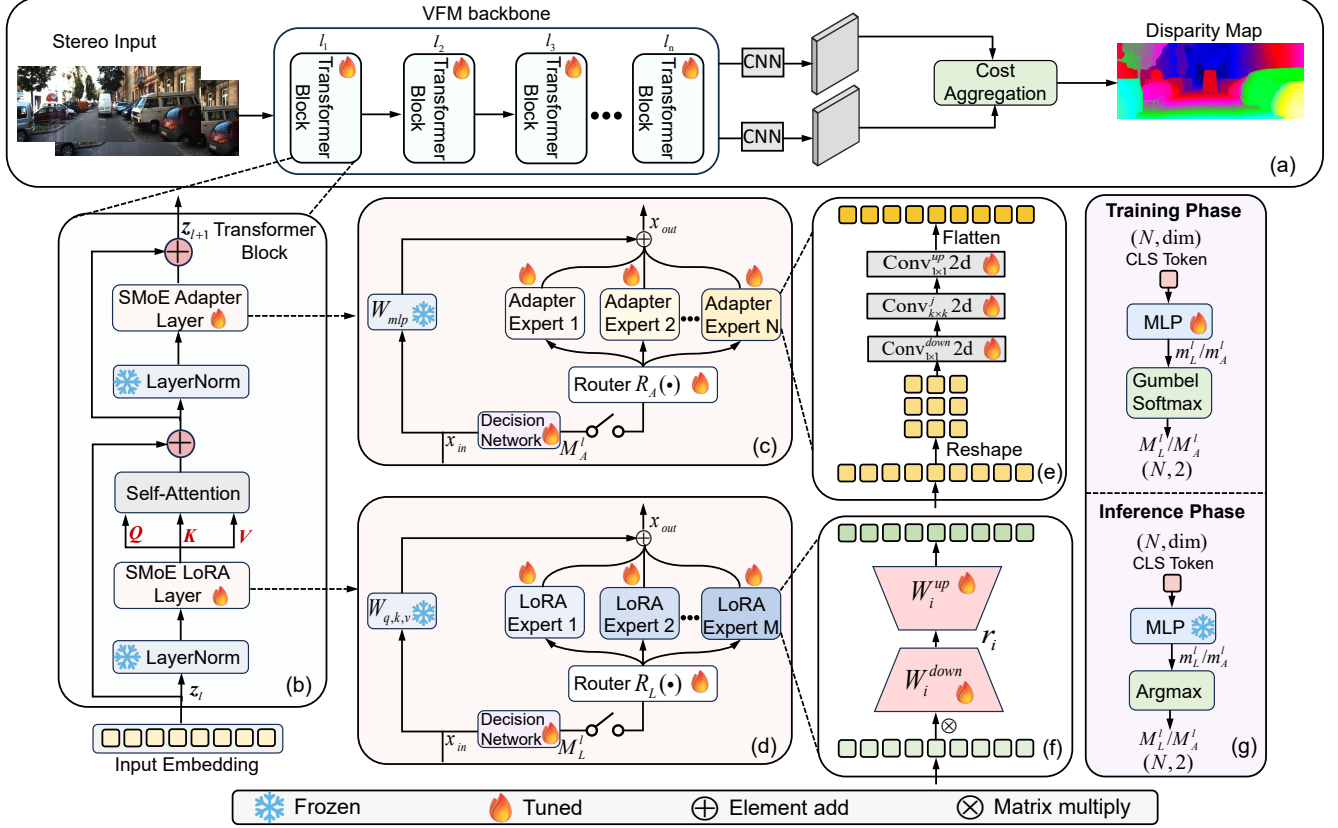


Figure 2. Overview of the designed SMoEStereo framework. (a) Overall model structure; (b) Overall the transformer block; (c) Details of the designed SMoE Adapter layer; (d) Details of the designed SMoE LoRA layer; (e) Details of each Adapter expert; (f) Details of each LoRA expert; (g) Details of the decision network. Note that our SMoE can serve as a plug-and-play module for most stereo networks.

LoRA and Adapter experts to use and 2) which MoE layer to skip based on per input to improve the model efficiency.

MoE LoRA Layer. Vanilla LoRA [17] typically pre-define a matrix rank to update the learnable weights ΔW while freezing the pre-trained weights W . However, in practical deployment environments, it is challenging to pre-define an ideal rank for different test samples. To address this, we introduce a MoE LoRA layer architecture that employs various LoRA layers with different matrix ranks as experts, thereby identifying the optimal LoRA expert for each input query. As shown in Fig. 2 (d), the MoE LoRA layer contains M experts, denoted as $\{E_L^i\}_{i=1}^M$. Each LoRA expert $E_L^i(\cdot)$ corresponds to a specific rank r_i . We then design a router network $R_L(\cdot)$ to dynamically select the Top- k optimal experts (default: $k=1$). Specifically, given input $x_{in} \in \mathbb{R}^{N \times dim}$ is fed into the MoE LoRA layer, where N is the token count and dim is the feature dimension, the forward propagation of the MoE-LoRA layer is defined as:

$$x_{out} = W_{q,k,v} x_{in} + \sum_{i=1}^M R_L(x_{in}) \cdot E_L^i(x_{in}), \quad (1)$$

where the pre-trained weights $W_{q,k,v}$ is frozen during training and $W_{q,k,v} \in \{W_q, W_k, W_v\}$ could be either query, key,

or value matrices. Each LoRA expert is formulated by:

$$E_L^i(x_{in}) = \Delta W x_{in} = W_i^{up} W_i^{down} x_{in}, \quad (2)$$

where $W_i^{up} \in \mathbb{R}^{d \times dim}$ and $W_i^{down} \in \mathbb{R}^{r_i \times r_i}$ are two trainable matrices, and $r_i \ll \{D, dim\}$. The router mechanism is presented as follows:

$$R(x_{in}) = \text{Topk}\left(\frac{\exp(W^{router} x_{in}/\tau)}{\sum_{k=1}^M \exp(W^{router} x_{in}/\tau)}, k\right), \quad (3)$$

where W^{router} represents the slim trainable weights, and τ denotes the temperature, empirically set to 5. In addition to the MoE LoRA layer, a router mechanism is also employed for the subsequent MoE Adapter layer.

MoE Adapter Layer. Existing ViTs struggle to learn from small datasets due to a lack of intrinsic inductive bias in modeling local visual structures [42, 43, 70, 80, 81]. To address this, we propose an MoE Adapter layer that captures local spatial priors with different receptive fields before injecting them into the plain ViT blocks. As shown in Fig. 2 (c), the designed MoE Adapter layer consists of several CNN Adapter experts $\{E_A^j\}_{j=1}^N$ with different kernel sizes k to embed local geometry into tokens. Similar

Table 1. Domain generalization evaluation on real-world datasets. ‡ denotes that VFMs use the ViT-large model capacity. (The best results in **bold** and the sub-optimal best results in **blue**).

Method	KIT 2012 Bad 3.0	KIT 2015 Bad 3.0	Middle Bad 2.0	ETH3D Bad 1.0
CFNet [53]	4.7	5.8	15.3	5.8
RAFT-Stereo [32]	5.1	5.7	12.6	3.3
UCFNet_pretrain [55]	4.5	5.2	26.0	4.8
CREStereo++ [22]	4.7	5.2	14.8	4.4
DLNR [82]	9.1	16.1	9.8	23.0
LoS [28]	4.4	5.5	19.6	3.1
Selective-IGEV [60]	5.7	6.1	13.3	6.1
MochaStereo [8]	4.9	5.9	11.5	3.9
Former-PSMNet [‡] (SAM) [79]	4.3	5.0	9.5	6.4
Former-CFNet [‡] (DINOv2) [79]	4.0	5.0	8.8	4.5
Former-RAFT [‡] (DAM) [79]	3.9	5.1	8.1	3.3
SMoEStereo (DINOv2 [41])	4.39	4.97	7.56	2.57
SMoEStereo (SAM [24])	4.27	4.89	7.10	2.07
SMoEStereo (DAM [73])	4.44	5.05	7.57	2.56
SMoEStereo (DAMV2 [74])	4.22	4.86	7.05	2.10

to the router mechanism in the MoE LoRA layer, we introduce a router network $R_A(\cdot)$ to select the optimal experts dynamically. The output can be formularized as:

$$x_{out} = W_{mlp} x_{in} + \sum_{j=1}^N R_A(x_{in}) \cdot E_A^j(x_{in}), \quad (4)$$

where W_{mlp} is frozen as the pre-trained weights during training. For each CNN Adapter expert $E_A(\cdot)$, it can be calculated as (we omit the activation layer):

$$E_A^j(x_{in}) = \text{Conv}_{1 \times 1}^{up}(\text{Conv}_{k \times k}^j(\text{Conv}_{1 \times 1}^{down}(x_{in}))), \quad (5)$$

where $\text{Conv}_{1 \times 1}^{down}$ and $\text{Conv}_{1 \times 1}^{up}$ are two 1×1 convolution layers, which reduce and restore the number of channels, respectively. $\text{Conv}_{k \times k}^j$ indicates the specific convolution layer with kernel size k in different experts.

Meanwhile, to prevent the gating network from assigning large weights to the same few experts, we apply a soft constraint [52] on the batch-wise average of each MoE LoRA and MoE Adapter router $R(\cdot)$. As such, for a given batch of data X , the MoE balance loss \mathcal{L}_{blc} is defined as:

$$\mathcal{L}_{blc} = \frac{\sigma^2(Q(X))}{\mu(Q(X))}; \quad Q(X) = \sum_{x \in X} R(x), \quad (6)$$

where $\sigma^2(Q(x))$ and $\mu(Q(X))$ represent the variance and mean of $Q(x)$, respectively, and we use $Q(x)$ to define the importance of experts with the MoE layer. Consequently, the balance loss \mathcal{L}_{blc} ensures equal importance among all experts, fostering their diverse utilization.

Decision Network. Previous studies [10, 25] demonstrated that different layers across the ViTs have varying contributions to the overall performance, depending on the input data distributions. Therefore, the goal of the decision policy is to identify ‘important’ MoE layers within each ViT block (i.e., 0 for ‘unimportant’ modules and 1 for ‘important’

Table 2. Zero-shot performance on DrivingStereo under different weathers, evaluated with official weights and the D1 metric.

Method	Sunny	Cloudy	Rainy	Foggy	Avg.
CFNet [53]	5.4	5.8	12.0	6.0	7.3
PCWNet [54]	5.6	5.9	11.8	6.2	7.4
DLNR [82]	27.1	28.3	34.5	29.0	29.8
IGEVStereo [68]	5.3	6.3	21.6	8.0	10.3
Selective-IGEV [60]	7.0	8.0	18.4	12.9	11.1
MochaStereo [8]	12.8	27.4	24.6	22.8	21.9
Former-CFNet [‡] [79]	3.8	2.7	8.3	5.2	5.0
SMoEStereo	3.3	3.0	6.0	4.7	4.3

Table 3. Inference efficiency comparison with VFM-based generalized methods. † indicates evaluation using the author’s codes. * denotes extra activated parameters within VFMs during inference.

Method	Capacity	Memory (GB)	Time (s)	Params. * (M)
Former-PSMNet (SAM) [†] [79]	ViT-Large	6.0	0.52	6.9
Former-RAFT (DAM) [†] [79]	ViT-Large	4.1	0.47	6.9
SMoEStereo (SAM)	ViT-Base	2.0	0.20	4.06
SMoEStereo (DAM)	ViT-Base	1.9	0.18	2.86

ones). To achieve this, a decision network with few learnable parameters is introduced to reduce redundancy by selectively keeping or dropping MoE layers based on the input samples. Specifically, the class token input $x_{cls} \in \mathbb{R}^{N \times dim}$ is first fed to MLP to produce the corresponding probability vectors m_L^l for MoE LoRA layer or m_A^l for MoE Adapter Layer, respectively. Then, to make the whole framework end-to-end trainable, the Gumbel softmax trick [37] is used to obtain the softened binary masks at l -th layer as follows:

$$M_i^l = \frac{\exp((m_i^l + G_i)/\tau)}{\sum_{j=1}^K \exp((m_j^l + G_j)/\tau)} \text{ for } i = 1, \dots, K, \quad (7)$$

where K represents the total number of categories ($K = 2$ for binary decisions in our case), G_i is a Gumbel noise tensor with all elements sampled from $\text{Gumbel}(0, 1)$, and τ controls the smoothness of M_i^l . During training, we use the first element, M_1^l , while binary values are obtained via Argmax during inference, as shown in Fig. 2 (g). For simplicity, we denote the usage policies for the MoE LoRA and MoE Adapter layers as M_L^l and M_A^l , respectively. Thus, Eq. (1) & (4) at the l -th layer can be reformulated as:

$$x_{out} = W_{q,k,v} x_{in} + M_L^l \left(\sum_{i=1}^M R_L(x_{in}) \cdot E_L^i(x_{in}) \right) \quad (8)$$

$$x_{out} = W_{mlp} x_{in} + M_A^l \left(\sum_{j=1}^N R_A(x_{in}) \cdot E_A^j(x_{in}) \right),$$

In summary, the decision network generates usage policies for each MoE layer within transformer blocks based on the input. The input is then processed through the block according to these policies. Meanwhile, to encourage reducing the overall computational cost, we devise the usage loss as:

$$\mathcal{L}_{usage} = \left(\frac{1}{L} \sum_{l=1}^L M_L^l - \gamma \right)^2 + \left(\frac{1}{L} \sum_{l=1}^L M_A^l - \gamma \right)^2, \quad (9)$$

Table 4. Performance comparison of the proposed SMoE against other finetuning methods. Params (M) Train/Test refers to the number of learnable and additional activated parameters within the VFM backbone for the training and inference phases, respectively.

Backbone	Fine-tune Method	Params (M) Train/Test	KIT 2012	KIT 2015	Middle	ETH3D
			Bad 3.0	Bad 3.0	Bad 2.0	Bad 1.0
DAMV2 [74] (ViT-base)	Frozen	0/0	11.7	15.4	24.6	16.2
	Full-Finetuning	97.5/0	10.1	14.3	22.3	15.7
	VPT [20]	4.72/4.72	4.65	5.18	8.44	2.71
	AdapterFormer [6]	4.14/4.14	4.59	5.11	8.39	2.77
	AdapterTuning [79]	4.20/4.20	4.40	5.09	8.03	2.61
	LoRA [17]	4.72/4.72	4.47	5.03	7.67	2.83
	SMoE (Ours)	6.81/2.86	4.22	4.86	7.05	2.10
SAM [24] (ViT-base)	Frozen	0/0	10.3	14.8	22.8	16.4
	Full-Finetuning	93.7/0	8.91	10.5	18.7	13.6
	VPT [20]	4.72/4.72	4.60	5.19	8.21	2.39
	AdapterFormer [6]	4.14/4.14	4.53	4.95	8.85	2.99
	AdapterTuning [79]	6.91/6.91	4.58	5.11	7.82	2.53
	LoRA [17]	4.72/4.72	4.49	5.17	7.76	2.78
	SMoE (Ours)	6.81/4.06	4.27	4.89	7.10	2.07

here, L denotes the total number of blocks of the ViT backbone. The hyper-parameters $\gamma \in (0, 1]$ indicate target computation budgets regarding the percentage of blocks to keep. **Total Loss Function.** Following [32], we supervise the $L1$ distance between the sequence of disparity predictions $\{\hat{D}_1, \dots, \hat{D}_N\}$ and the Ground Truth Disparity D_{gt} :

$$\mathcal{L}_{disp} = \sum_{i=1}^N \beta^{N-i} \|(D_{gt} - \hat{D}_i)\|_1, \quad (10)$$

where $N = 16$ during training and the exponentially weight β is set to 0.9. The total loss is formulated as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{disp} + \lambda_1 \mathcal{L}_{blc} + \lambda_2 \mathcal{L}_{usage}. \quad (11)$$

Where we set the hyper-parameters to $\lambda_1 = \lambda_2 = 1$.

4. Experiments

4.1. Experimental Settings

Datasets & Metrics. SceneFlow [38] is a large-scale synthesis dataset containing 35454 training samples and 4370 validation samples. KITTI 2015 [39] and KITTI 2012 [13] are two real-world datasets with 200 and 194 outdoor driving scenes stereo images, respectively. Middlebury [49] consists of 28 training samples and 15 high-resolution evaluation images of indoor scenes. ETH3D [50] is a low-resolution dataset with 27 gray images. We also qualitatively evaluate the robustness of our network on the DrivingStereo dataset [72] under challenging weather conditions. Except for Middlebury (half-resolution), we use full resolution for these datasets. For cross-domain generalization, we use the SceneFlow dataset to pre-train our model and report its cross-domain generalization ability on realistic datasets. For joint generalization evaluation, we strictly follow previous Robust Vision Challenging (RVC) settings [22, 53, 55], where we adopt KITTI 2015 & 2012,

Table 5. Domain generalization evaluation (peak results) on four target training sets. Note that, the inference time is evaluated on a KITTI stereo pair using a single Nvidia 5000 Ada GPU.

VFM	Model Capacity	Time (s)	KIT 2012 Bad 3.0	KIT 2015 Bad 3.0	Middle Bad 2.0	ETH3D Bad 1.0
DAMV2	ViT-Small [74]	0.16	4.57	5.22	7.78	2.43
DAMV2	ViT-Base [74]	0.20	4.22	4.86	7.05	2.10
DAMV2	ViT-Large [74]	0.32	3.94	4.59	6.71	1.87

	KITTI 12				KITTI 15				Middlebury				ETH3D															
	Expert 0 (rank=4)	Expert 1 (rank=8)	Expert 2 (rank=16)	Expert 3 (rank=32)	Expert 0 (rank=4)	Expert 1 (rank=8)	Expert 2 (rank=16)	Expert 3 (rank=32)	Expert 0 (kernel=9)	Expert 1 (kernel=7)	Expert 2 (kernel=9)	Expert 3 (kernel=9)	Expert 0 (kernel=9)	Expert 1 (kernel=7)	Expert 2 (kernel=9)	Expert 3 (kernel=9)												
LORA	18.1%	17.6%	21.9%	20.2%	0.20%	0.19%	0.02%	0.93%	0.26%	0.16%	3.61%	4.01%	17.6%	18.1%	16.1%	17.7%	16.4%	16.6%	13.4%	13.1%	19.6%	19.4%	17.1%	19.5%	8.14%	8.43%	8.05%	6.79%

Figure 3. Dynamic expert selection on four real-world training datasets. Lighter colors represent lower activation rates.

Middlebury, and ETH3D training sets to finetune our pre-trained model jointly and evaluate on three real-world public benchmarks (KITTI 2015, Middlebury, and ETH3D) using a single fixed model without adaptation. Following previous works [15, 62–64], EPE and t px (the percentage of outliers with an absolute error greater than t pixels) are used to evaluate the performance.

Implementation Details. DAMV2 [74] (ViT-base) and RAFT-Stereo [32] are used as final backbones to present SMoEStereo in the experiments. We also adopt DAM [73], SAM [24], and DINOv2 [41] as VFM variants. During pre-training, we pre-train our model on SceneFlow with 20K iterations and a batch size of 32. For joint generalization, we follow previous RVC settings [55], augmenting the Middlebury and ETH3D training sets to match KITTI 2015/2012 in size, preventing small datasets from being overwhelmed by large datasets. We then fine-tune the model on mixed datasets with 20K iterations. All experiments are implemented in Pytorch using 8 RTX 5000 Ada GPUs. We use the AdamW [23] optimizer and a one-cycle learning rate schedule [56] with a maximum learning rate of 2×10^{-4} . For data augmentation, we employ asymmetric chromatic augmentations [27] and asymmetric occlusion [65, 71] to the right image. The image pairs are randomly cropped to 320×832 for training. We perform 16 iterations during training and report the results of 24 inference iterations.

4.2. Robustness Evaluation

In stereo matching, cross-domain generalization [77, 79] and joint generalization [53, 55] are key for evaluating the robustness of stereo matching methods. Therefore, we conduct these two types of robustness experiments.

Table 6. Robustness comparison among ETH3D, Middlebury, and KITTI2015 testsets with existing SOTA methods in RVC. All methods are tested on three datasets with a single fixed model. We evaluate Middlebury and ETH3D using the ‘‘All’’ metrics. The overall rank is obtained by Schulze Proportional Ranking [51] to combine multiple rankings into one. Our approach achieves the best overall performance.

Methods	Middlebury				KITTI 2015				ETH3D				Overall Rank
	bad 2.0 (%)	bad 4.0 (%)	AvgErr	Rank	D1-bg (%)	D1-fg (%)	D1-All (%)	Rank	bad 1.0 (%)	bad 2.0 (%)	AvgErr	Rank	
AANet_RVC [69]	31.8	25.8	12.8	11	2.23	4.89	2.67	9	5.41	1.95	0.33	9	12
GANet_RVC [76]	24.9	16.3	15.8	10	1.88	4.58	2.33	8	6.97	1.25	0.45	11	11
HSMNet_RVC [71]	16.5	9.68	3.44	5	2.74	8.73	3.74	10	4.40	1.51	0.28	8	10
MaskLacGwcNet_RVC [45]	15.8	10.3	13.5	8	1.65	3.68	1.99	6	6.42	1.88	0.38	10	9
NLCANetV2_RVC [44]	16.4	10.3	5.60	8	1.51	3.97	1.92	4	4.11	1.20	0.29	7	8
CroCo_RVC [67]	19.7	12.2	5.14	9	2.04	3.75	2.33	8	1.54	0.50	0.21	3	7
CFNet_RVC [53]	16.1	11.3	5.07	7	1.65	3.53	1.96	5	3.70	0.97	0.26	6	6
UCFNet_RVC [21]	16.7	10.9	5.96	4	1.57	3.33	1.86	3	3.37	0.78	0.25	5	5
iRaftStereo_RVC [21]	13.3	8.02	2.90	4	1.88	3.03	2.07	7	1.88	0.55	0.17	4	4
CREStereo++_RVC [22]	9.46	6.25	2.20	3	1.55	3.53	1.88	4	1.70	0.37	0.16	2	3
LoS_RVC [28]	9.30	6.03	2.36	2	1.58	3.03	1.83	2	1.47	0.25	0.14	1	2
SMoEStereo_RVC (Ours)	9.74	5.41	1.94	1	1.50	2.88	1.73	1	1.13	0.26	0.14	1	1

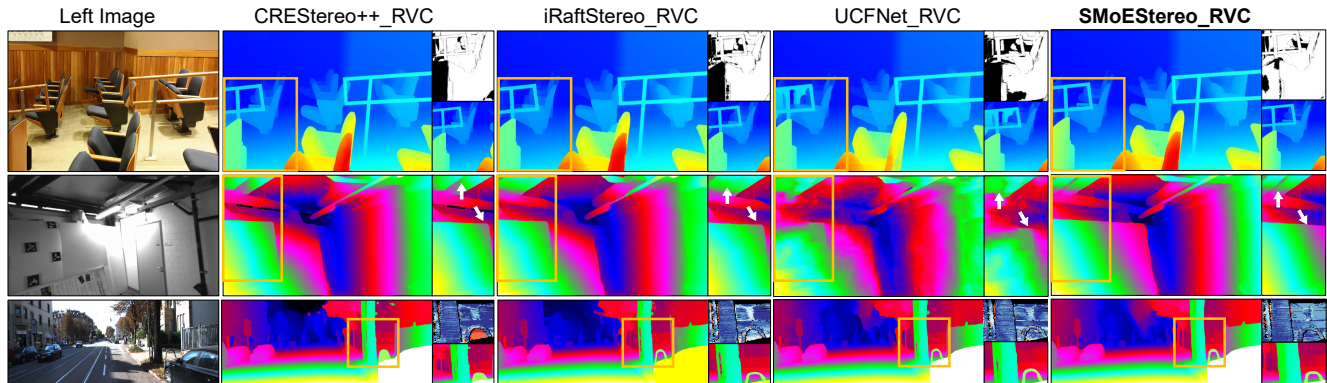


Figure 4. Qualitative results on the test set of Middlebury, ETH3D, and KITTI2015. Since ETH3D benchmark do not provide the error map, we only show the zoomed highlight parts. In the error map, lighter colors represent smaller errors, and vice versa. Best zoomed in.

4.2.1. Cross-domain Generalization

Our SMoEStereo first demonstrates strong robustness in zero-shot settings, as evidenced in Table 1 and Table 2.

Comparison with Domain Generalized Approaches. Our SMoEStereo significantly outperforms most stereo matching methods tailored for domain generalization. Notably, HVT-RAFT [4], Former-PSMNet, Former-CFNet, and Former-RAFT [79] use specialized modules and losses for domain generalization, while our method does not require. We also compare the robustness of SMoEStereo under challenging outdoor conditions such as cloud, fog, and rain, as illustrated in Table 2. Additionally, while Former-RAFT [79] uses the ViT-Large of DAM, we employ the smaller ViT-Base version. As shown in Table 3, our method has fewer parameters and faster inference time.

More VFMs. We extend our experiments by integrating SMoE with other VFMs, such as DAM [73], SAM [24], DINOv2 [41] to highlight the versatility of our SMoE. As shown in Table 1, our findings reveal that SMoE exhibits remarkable performance with diverse VFM backbones.

Comparing SMoE with PEFT Methods. We conduct a comprehensive performance comparison of SMoE against existing PEFT methods for domain generation, as detailed in Table 4. In addition to ‘Frozen’ and ‘Full-finetuning’, we

developed four network variants by replacing SMoE modules with other PEFT methods: VPT [20], Adapter [6], LoRA [17], and Adapter-Tuning [79]. Leveraging the robust feature extraction of VFMs, these PEFT methods have shown notable advancements in generalization ability. Using identical VFM backbones, SMoE outperforms previous domain generalization and other PEFT methods. Additionally, SMoE enjoys fewer parameters during the inference phase than other PEFT methods.

VFM Capacity. Using DAMV2 [74] as an example, we develop variants with different capacities to assess the impact on zero-shot performance. Table 5 shows that larger capacities enhance zero-shot performance, as larger models provide stronger representation abilities and robust priors essential for generalization.

Expert Selection Distribution. Different stereo datasets exhibit significant domain gaps. From Fig. 3, we observe distinct LoRA and Adapter expert selection distributions among these four datasets. Our SMoE framework dynamically activates the optimal combination of LoRA and Adapter experts for each dataset, which empirically validates SMoE’s flexible adaptability - a critical advantage for in-the-wild deployment, where robust cross-domain generalization hinges on dynamic expert selection.

Table 7. Ablation of the proposed method on Middlebury, KITTI2015, and ETH3D training datasets. DAMV2 [74] (ViT-Base) is used as the frozen VFM. Our MoE LORA and MoE Adapter selectively activate the optimal expert rather than all experts. Params. denote the extra activated parameters within the DAMV2 backbone. **Apricot color** represents the baseline results, while **blue color** represents the final model results. The network component is evaluated individually in the table. Inference time is measured on KITTI by A5000 Ada GPU.

ID	MoE LORA	MoE Adapter	Decision Network	Random Policy	KITTI 2015		Middlebury		ETH3D		Number of MoE	Params. (M)	Runtime (s)
					EPE	D1.All	EPE	Bad 2.0	EPE	Bad 1.0			
1	-	-	-	-	0.72	2.01	0.97	7.67	0.25	1.79	0	0	0.17
2	✓	-	-	-	0.63	1.65	0.81	4.80	0.17	0.92	12	2.29	0.19
3	-	✓	-	-	0.64	1.71	0.77	4.68	0.18	1.01	12	4.49	0.19
4	✓	✓	-	-	0.61	1.57	0.72	4.24	0.14	0.56	24	6.77	0.23
5	✓	✓	-	✓	0.64	1.69	0.75	4.66	0.17	0.84	~14	2.86	0.20
6	✓	✓	✓	-	0.60	1.51	0.71	4.12	0.15	0.63	~14	2.86	0.20

Table 8. Impacts of the matrix rank of MoE LoRA layers.

Setting	KITTI 2015		Middlebury		ETH3D	
	EPE	D1.All	EPE	Bad 2.0	EPE	Bad 1.0
rank = 4	0.65	1.77	0.85	4.75	0.16	0.73
rank = 8	0.64	1.73	0.78	4.64	0.17	0.80
rank = 16	0.63	1.61	0.74	4.29	0.18	0.84
rank = 32	0.62	1.57	0.76	4.42	0.16	0.77
SMoE	0.60	1.51	0.71	4.12	0.15	0.63

4.2.2. Joint Generalization Evaluation

Table 6 presents the results of our method and existing SOTA methods in stereo matching for the Robust Vision Challenge (RVC). In RVC, all methods should be jointly trained on real-world datasets (KITTI 2012 & 2015, Middlebury, and ETH3D) and evaluated on three public benchmarks using a single fixed model without adaptation. As shown, SMoEStereo_RVC achieves 1st on these public benchmarks among all methods, clearly demonstrating the advantage of our approach. Note that, LoS_RVC [29] use additional large-scale datasets such as Instereo2K [3] and CRE [27] to achieve impressive results. The larger data capacity improves the overall performance (See supplementary). The visual comparison results are shown in Fig. 4.

4.3. Ablation Studies

We conduct ablation studies to demonstrate the efficacy of each network component in RVC settings. Cross-domain generalization ablations are detailed in the Supplementary.

Ablation of Main Components. To evaluate the effectiveness of the designed PEFT modules, we report cross-dataset accuracy results in Table 7. Utilizing MoE LoRA and MoE adapters enhances disparity estimation performance compared to the vanilla VFM baseline (ID = 1) by capturing long-range interactions and local geometry cues. This improvement is attributed to our MoE design, which effectively learns robust features in the target domain while preserving knowledge from dense prediction tasks. However, with more MoE modules, computational costs inevitably increase (ID = 4). The proposed decision network mitigates this by removing redundancy while maintaining high performance (ID = 6). Additionally, replacing learned usage policies with randomly generated ones of similar computational cost (ID = 5) results in a notable drop in accuracy,

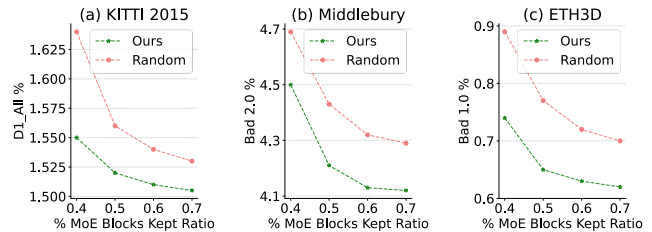


Figure 5. Quantitative results on real-world datasets between the proposed layer selection strategy and their Random counterparts.

confirming the effectiveness of the learned usage policy.

Study on Rank r . In Table 8, we compare the cross-dataset performance of SMoEStereo with LoRA experts of varying ranks. Different datasets exhibit varying optimal LoRA ranks. Our heterogeneous MoE dynamically selects the optimal LoRA rank for each input query, consistently outperforming the homogeneous MoE LoRA experts on all three datasets. Similarly, the impact of different adapters and LoRA with more ranks is investigated in the supplementary. This fully demonstrates the effectiveness of our scene-conditioned selection mechanism in stereo matching.

Different Computational Budgets. SMoEStereo is designed to accommodate the needs of different computational budgets flexibly by varying the hyperparameter γ (MoE kept ratio). As demonstrated in Fig. 5, our method can cover a wide range of trade-offs between efficiency and accuracy and outperforms Random baselines significantly.

5. Conclusion

In this paper, we introduce SMoEStereo, a general framework designed to effortlessly leverage VFMs for robust stereo matching in the wild. By dynamically selecting suitable MoE and experts based on inputs, SMoEStereo exhibits strong robustness with few learnable parameters, significantly outperforming previous robust methods. Experimental results show that our approach performs well on various datasets and has generic applicability.

6. Acknowledgements

This work is partially supported by the Shenzhen Science and Technology Program under Grant RCBS20231211090736065, Guangdong Basic and Applied Basic Research Foundation under Grant

2023A151511, Guangdong Natural Science Fund under Grant 2024A1515010252, Shenzhen Science and Technology Program under Grant JCYJ20241202124010014, the Science and Technology Research Projects of the Education Office of Jilin Province under Grant JJKH20251951KJ, and the National Natural Science Foundation of China under Grant 62301601.

References

- [1] Muhammad Awais, Muzammal Naseer, Salman Khan, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Foundational models defining a new era in vision: A survey and outlook. *arXiv preprint arXiv:2307.13721*, 2023. 3
- [2] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems (NIPS)*, 35:32897–32912, 2022. 3
- [3] Wei Bao, Wei Wang, Yuhua Xu, Yulan Guo, Siyu Hong, and Xiaohu Zhang. Instereo2k: a large real dataset for stereo matching in indoor scenes. *Science China Information Sciences*, 63:1–11, 2020. 8
- [4] Tianyu Chang, Xun Yang, Tianzhu Zhang, and Meng Wang. Domain generalized stereo matching via hierarchical visual transformation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9559–9568, 2023. 3, 7
- [5] Chenyi Chen, Ari Seff, Alain Kornhauser, and Jianxiong Xiao. DeepDriving: Learning affordance for direct perception in autonomous driving. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2722–2730, 2015. 1
- [6] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems (NIPS)*, 35:16664–16678, 2022. 3, 6, 7
- [7] Shuo Chen, Jindong Gu, Zhen Han, Yunpu Ma, Philip Torr, and Volker Tresp. Benchmarking robustness of adaptation methods on pre-trained vision-language models. *Advances in Neural Information Processing Systems (NIPS)*, 36:51758–51777, 2023. 2
- [8] Ziyang Chen, Wei Long, He Yao, Yongjun Zhang, Bingshu Wang, Yongbin Qin, and Jia Wu. Mocha-stereo: Motif channel attention network for stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27768–27777, 2024. 5
- [9] WeiQin Chuah, Ruwan Tennakoon, Reza Hoseinnezhad, Alireza Bab-Hadiashar, and David Suter. Itsa: An information-theoretic approach to automatic shortcut avoidance and domain generalization in stereo matching networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13022–13032, 2022. 3
- [10] Alessio Devoto, Federico Alvetreti, Jary Pomponi, Paolo Di Lorenzo, Pasquale Minervini, and Simone Scardapane. Adaptive layer selection for efficient vision transformer fine-tuning. *arXiv preprint arXiv:2408.08670*, 2024. 5
- [11] Shihan Dou, Enyu Zhou, Yan Liu, Songyang Gao, Jun Zhao, Wei Shen, Yuhao Zhou, Zhiheng Xi, Xiao Wang, Xiaoran Fan, Shiliang Pu, Jiang Zhu, Rui Zheng, Tao Gui, Qi Zhang, and Xuanjing Huang. Loramoe: Revolutionizing mixture of experts for maintaining world knowledge in language model alignment, 2023. 3
- [12] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research (JLMR)*, 23(120):1–39, 2022. 3
- [13] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3361, 2012. 6
- [14] Weiyu Guo, Zhaoshuo Li, Yongkui Yang, Zheng Wang, Russell H Taylor, Mathias Unberath, Alan Yuille, and Yingwei Li. Context-enhanced stereo transformer. In *European Conference on Computer Vision (ECCV)*, pages 263–279. Springer, 2022. 2
- [15] Yulan Guo, Yun Wang, Longguang Wang, Zi Wang, and Chen Cheng. Cvcnet: Learning cost volume compression for efficient stereo matching. *IEEE Transactions on Multimedia (TMM)*, 25:7786–7799, 2022. 6
- [16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 3
- [17] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 3, 4, 6, 7
- [18] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 2
- [19] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991. 3
- [20] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision (ECCV)*, pages 709–727. Springer, 2022. 3, 6, 7
- [21] Hualie Jiang, Rui Xu, and Wenjie Jiang. An improved raft-stereo trained with a mixed dataset for the robust vision challenge 2022. *arXiv preprint arXiv:2210.12785*, 2022. 7
- [22] Junpeng Jing, Jiankun Li, Pengfei Xiong, Jiangyu Liu, Shuaicheng Liu, Yichen Guo, Xin Deng, Mai Xu, Lai Jiang, and Leonid Sigal. Uncertainty guided adaptive warping for robust and efficient stereo matching. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3318–3327, 2023. 1, 3, 5, 6, 7

- [23] Diederik P Kingma and Jimmy Ba. ADAM: A method for stochastic optimization. In *Proceedings of the IEEE International Conference on Learning Representations (ICLR Poster)*, 2015. 6
- [24] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023. 2, 3, 5, 6, 7
- [25] Yoonho Lee, Annie S Chen, Fahim Tajwar, Ananya Kumar, Huaxiu Yao, Percy Liang, and Chelsea Finn. Surgical fine-tuning improves adaptation to distribution shifts. *International Conference on Learning Representations (ICLR)*, 2023. 5
- [26] Dengchun Li, Yingzi Ma, Naizheng Wang, Zhiyuan Cheng, Lei Duan, Jie Zuo, Cal Yang, and Mingjie Tang. Mixlora: Enhancing large language models fine-tuning with lora based mixture of experts. *arXiv preprint arXiv:2404.15159*, 2024. 3
- [27] Jiankun Li, Peisen Wang, Pengfei Xiong, Tao Cai, Ziwei Yan, Lei Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. Practical stereo matching via cascaded recurrent network with adaptive correlation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16263–16272, 2022. 3, 6, 8
- [28] Kunhong Li, Longguang Wang, Ye Zhang, Kaiwen Xue, Shunbo Zhou, and Yulan Guo. Los: Local structure-guided stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19746–19756, 2024. 5, 7
- [29] Kunhong Li, Longguang Wang, Ye Zhang, Kaiwen Xue, Shunbo Zhou, and Yulan Guo. Los: Local structure guided stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3, 8
- [30] Zhaoshuo Li, Xingtong Liu, Nathan Drenkow, Andy Ding, Francis X Creighton, Russell H Taylor, and Mathias Unberath. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 6197–6206, 2021. 2
- [31] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2117–2125, 2017. 3
- [32] Lahav Lipson, Zachary Teed, and Jia Deng. RAFT-Stereo: Multilevel recurrent field transforms for stereo matching. *2021 International Conference on 3D Vision (3DV)*, pages 218–227, 2021. 2, 3, 5, 6
- [33] Biyang Liu, Huimin Yu, and Guodong Qi. Graftnet: Towards domain generalized stereo matching with a broad-spectrum and task-oriented feature. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13012–13021, 2022. 3
- [34] Chuang-Wei Liu, Qijun Chen, and Rui Fan. Playing to vision foundation model’s strengths in stereo matching. *arXiv preprint arXiv:2404.06261*, 2024. 2, 3
- [35] Zefang Liu and Jiahua Luo. Adamole: Fine-tuning large language models with adaptive mixture of low-rank adaptation experts. *arXiv preprint arXiv:2405.00361*, 2024. 3
- [36] Tongxu Luo, Jiahe Lei, Fangyu Lei, Weihao Liu, Shizhu He, Jun Zhao, and Kang Liu. Moelora: Contrastive learning guided mixture of experts on parameter-efficient fine-tuning for large language models. *arXiv preprint arXiv:2402.12851*, 2024. 3
- [37] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016. 5
- [38] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. 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)*, pages 4040–4048, 2016. 6
- [39] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3061–3070, 2015. 6
- [40] Don Murray and James J Little. Using real-time stereo vision for mobile robot navigation. *autonomous robots*, 8:161–171, 2000. 1
- [41] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2024. 3, 5, 6, 7
- [42] Namuk Park and Songkuk Kim. How do vision transformers work? *arXiv preprint arXiv:2202.06709*, 2022. 4
- [43] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in neural information processing systems (NIPS)*, 34:12116–12128, 2021. 4
- [44] Zhibo Rao, Mingyi He, Yuchao Dai, Zhidong Zhu, Bo Li, and Renjie He. NLCA-Net: A non-local context attention network for stereo matching. *APSIPA Transactions on Signal and Information Processing*, 9, 2020. 7
- [45] Zhibo Rao, Bangshu Xiong, Mingyi He, Yuchao Dai, Renjie He, Zhelun Shen, and Xing Li. Masked representation learning for domain generalized stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5435–5444, 2023. 7
- [46] Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems (NIPS)*, 34:8583–8595, 2021. 3

- [47] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 3
- [48] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 3
- [49] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *German conference on pattern recognition (GCPR)*, pages 31–42. Springer, 2014. 6
- [50] Thomas Schöps, Johannes L. Schönberger, S. Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. 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)*, pages 2538–2547, 2017. 6
- [51] Markus Schulze. A new monotonic, clone-independent, reversal symmetric, and condorcet-consistent single-winner election method. *Social choice and Welfare*, 36(2):267–303, 2011. 7
- [52] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *International Conference on Learning Representations (ICLR)*, 2017. 3, 5
- [53] Zhelun Shen, Yuchao Dai, and Zhibo Rao. CFNet: Cascade and fused cost volume for robust stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13906–13915, 2021. 2, 3, 5, 6, 7
- [54] Zhelun Shen, Yuchao Dai, Xibin Song, Zhibo Rao, Dingfu Zhou, and Liangjun Zhang. Pcw-net: Pyramid combination and warping cost volume for stereo matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 280–297. Springer, 2022. 5
- [55] Zhelun Shen, Xibin Song, Yuchao Dai, Dingfu Zhou, Zhibo Rao, and Liangjun Zhang. Digging into uncertainty-based pseudo-label for robust stereo matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 30(2):1–18, 2023. 3, 5, 6
- [56] Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, pages 369–386. SPIE, 2019. 6
- [57] Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, pages 402–419, 2020. 2
- [58] Jiali Wang, Daniel Scharstein, Akash Bapat, Kevin Blackburn-Matzen, Matthew Yu, Jonathan Lehman, Suhil Alsisan, Yanghan Wang, Sam Tsai, Jan-Michael Frahm, et al. A practical stereo depth system for smart glasses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21498–21507, 2023. 1
- [59] Shuzhe Wang, Vincent Leroy, Yann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3
- [60] Xianqi Wang, Gangwei Xu, Hao Jia, and Xin Yang. Selective-stereo: Adaptive frequency information selection for stereo matching. *arXiv preprint arXiv:2403.00486*, 2024. 3, 5
- [61] Yan Wang, Zihang Lai, Gao Huang, Brian H Wang, Laurens Van Der Maaten, Mark Campbell, and Kilian Q Weinberger. Anytime stereo image depth estimation on mobile devices. In *2019 international conference on robotics and automation (ICRA)*, pages 5893–5900. IEEE, 2019. 2
- [62] Yun Wang, Longguang Wang, Hanyun Wang, and Yulan Guo. SPNet: Learning stereo matching with slanted plane aggregation. *IEEE Robotics and Automation Letters*, 2022. 6
- [63] Yun Wang, Longguang Wang, Kunhong Li, Yongjian Zhang, Dapeng Oliver Wu, and Yulan Guo. Cost volume aggregation in stereo matching revisited: A disparity classification perspective. *IEEE Transactions on Image Processing (TIP)*, 2024.
- [64] Yun Wang, Kunhong Li, Longguang Wang, Junjie Hu, Dapeng Oliver Wu, and Yulan Guo. Adstereo: Efficient stereo matching with adaptive downsampling and disparity alignment. *IEEE Transactions on Image Processing (TIP)*, 2025. 6
- [65] Yun Wang, Jiahao Zheng, Chenghao Zhang, Zhanjie Zhang, Kunhong Li, Yongjian Zhang, and Junjie Hu. Dualnet: Robust self-supervised stereo matching with pseudo-label supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 8178–8186, 2025. 6
- [66] Zhixiang Wei, Lin Chen, Yi Jin, Xiaoxiao Ma, Tianle Liu, Pengyang Ling, Ben Wang, Huaian Chen, and Jinjin Zheng. Stronger fewer & superior: Harnessing vision foundation models for domain generalized semantic segmentation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28619–28630, 2024. 2
- [67] Philippe Weinzaepfel, Thomas Lucas, Vincent Leroy, Yann Cabon, Vaibhav Arora, Romain Brégier, Gabriela Csurka, Leonid Antsfeld, Boris Chidlovskii, and Jérôme Revaud. Croco v2: Improved cross-view completion pre-training for stereo matching and optical flow. In *Proceedings of the IEEE International Conference on Computer Vision (CVPR)*, pages 17969–17980, 2023. 2, 7
- [68] Gangwei Xu, Xianqi Wang, Xiaohuan Ding, and Xin Yang. Iterative geometry encoding volume for stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21919–21928, 2023. 5
- [69] Haofei Xu and Juyong Zhang. AANet: Adaptive aggregation network for efficient stereo matching. In *Proceedings of the*

- IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1959–1968, 2020. [7](#)
- [70] Yufei Xu, Qiming Zhang, Jing Zhang, and Dacheng Tao. Vi-tae: Vision transformer advanced by exploring intrinsic inductive bias. *Advances in Neural Information Processing Systems (NIPS)*, 34:28522–28535, 2021. [4](#)
- [71] Gengshan Yang, Joshua Manela, Michael Happold, and Deva Ramanan. Hierarchical deep stereo matching on high-resolution images. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5515–5524, 2019. [6](#), [7](#)
- [72] Guorun Yang, Xiao Song, Chaoqin Huang, Zhidong Deng, Jianping Shi, and Bolei Zhou. Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 899–908, 2019. [6](#)
- [73] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. [5](#), [6](#), [7](#)
- [74] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024. [1](#), [2](#), [3](#), [5](#), [6](#), [7](#), [8](#)
- [75] Yuqi Yang, Peng-Tao Jiang, Qibin Hou, Hao Zhang, Jinwei Chen, and Bo Li. Multi-task dense prediction via mixture of low-rank experts. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27927–27937, 2024. [3](#)
- [76] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip HS Torr. GA-Net: Guided aggregation net for end-to-end stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 185–194, 2019. [3](#), [7](#)
- [77] Feihu Zhang, Xiaojuan Qi, Ruigang Yang, Victor Prisacariu, Benjamin Wah, and Philip Torr. Domain-invariant stereo matching networks. In *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, pages 420–439. Springer, 2020. [1](#), [6](#)
- [78] Jiawei Zhang, Xiang Wang, Xiao Bai, Chen Wang, Lei Huang, Yimin Chen, Lin Gu, Jun Zhou, Tatsuya Harada, and Edwin R Hancock. Revisiting domain generalized stereo matching networks from a feature consistency perspective. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13001–13011, 2022. [3](#)
- [79] Yongjian Zhang, Longguang Wang, Kunhong Li, Yun Wang, and Yulan Guo. Learning representations from foundation models for domain generalized stereo matching. In *European Conference on Computer Vision (ECCV)*, pages 146–162. Springer, 2025. [2](#), [3](#), [5](#), [6](#), [7](#)
- [80] Zhanjie Zhang, Yuxiang Li, Ruichen Xia, Mengyuan Yang, Yun Wang, Lei Zhao, and Wei Xing. Lgast: Towards high-quality arbitrary style transfer with local–global style learning. *Neurocomputing*, 623:129434, 2025. [4](#)
- [81] Zhanjie Zhang, Quanwei Zhang, Junsheng Luan, Mengyuan Yang, Yun Wang, and Lei Zhao. Spast: Arbitrary style transfer with style priors via pre-trained large-scale model. *Neural Networks*, page 107556, 2025. [4](#)
- [82] Haoliang Zhao, Huizhou Zhou, Yongjun Zhang, Jie Chen, Yitong Yang, and Yong Zhao. High-frequency stereo matching network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1327–1336, 2023. [5](#)
- [83] Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew M Dai, Quoc V Le, James Laudon, et al. Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems (NIPS)*, 35:7103–7114, 2022. [3](#)