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

Supplementary Material

Table 1. Runtime efficiency comparison on KITTI 2015.

Method	Feature Extractor	Iterations	Time (*s*)	Res.	GPU
RAFTStereo [9]	ResNet-like	32	0.17	1242 x 375	5000Ada
DLNR [22]	Transformer	32	0.27	1242 x 375	5000Ada
Selective-IGEV [16]	ResNet-like	32	0.21	1242 x 375	5000Ada
MochaStereo [3]	ResNet-like	32	0.28	1242 x 375	5000Ada
Former-RAFT [21]	ViT-Large	32	0.47	1242 x 375	5000Ada
Ours	ViT-Base	24	0.20	1242 x 375	5000Ada

1. Overview

In this supplementary material, we provide additional details of our method and experiments, including:

- Analysis on runtime efficiency [21].
- The detailed experimental settings.
- More ablation studies.
- The data capacity of SMoEStereo when more synthetic samples are used for training.
- More zero-shot visualization results.

2. Analysis on Runtime Efficiency

Our SMoEStereo achieves better inference efficiency than most RAFT-based methods, as shown in Tab. 1. Although VFMs slightly increase feature extraction time, their robust features significantly reduce the need for iterative disparity refinements. Specifically, our GRU performs only 24 iterations *vs.* 32 iterations in [9, 16, 21, 22, 22], thus improving overall efficiency. Notably, SMoE offers a flexible selection mechanism to control the number of MoE modules, reducing inference time for various real-world applications. We benchmark inference runtime on KITTI 2015 (1242×375) using an RTX 5000 Ada GPU.

3. More Details about Experiments

3.1. More Details about VFMs.

SAM. Aligning with the methodology described in the foundational paper [5], we employ the ViT-Large architecture as our image encoder, making use of pre-trained weights that were trained on SA-1B [5] for a promptable segmentation task. The patch size of this model is set to 16×16, and each layer is designed to output features with a dimensionality of 768, summing up to a total of 12 layers. The positional embeddings of the model are upscaled to a length of 1024 via bicubic interpolation. From this model, we extract features from the 2nd, 5th, 7th, and 11th layers layers and feed them into the decoder.

DAM& DAMV2. DAM and DAMV2 designed a data engine to automatically generate depth annotations for unlabeled images, enabling data scaling up to an arbitrary scale. It collects 62M diverse and informative images from eight public large-scale datasets, e.g., SA-1B [5], Open Images [6], and BDD100K [19] for training an initial MDE mode in a self-training manner [7]. Similar to SAM, we also employ the ViT-Base capacity as our image encoder. The patch size of this model is set to 16×16, and each layer is designed to output features with a dimensionality of 128, summing up to 12 layers. The positional embeddings of the model are upscaled to a length of 1024 via bicubic interpolation. From this model, we extract features from the 2nd, 5th, 8th, and 11th layers and feed them into the subsequent cost aggregators.

DINOv2. Our choice of backbone for this study is DINOv2-Base, which has been distilled from DINOv2-Large. As noted in the original documentation, DINOv2-Base occasionally surpasses the performance of DINOv2-Large [11]. we apply equivalent processing to both the positional embeddings and patch embed layer of DINOv2-Base. The features extracted from the 2nd, 5th, 8th, and 11th layers are subsequently fed into the decode head. DINOv2 is originally pretrained in a self-supervised fashion on the LVD-142M [11] dataset, following the procedures outlined in its respective paper.

3.2. More Details about PEFT Methods.

1. VPT, Adapter-Tuning, AdaptFormer, and LoRA. Based on extensive experimentation, we have optimized the implementation of PEFT methods for DAMV2 and SAM, utilizing PEFT configurations that enhance robust performance. These methods include: 1) VPT: We use the VPT-Deep configuration and incorporate 256 learnable tokens within each ViT layer (12 layers for ViT-Base).

2) LoRA: Applied to the query and value components for self-attention, default configured with a rank of 128.

3) AdaptFormer: Similar to LoRA, equipped with MLP layers and employs a bottleneck design with a default ank of 32, initialized using LoRA, and notably omits layer normalization.

4) AdapterTuning: We adopt a DPT decoder strategy [12] with multi-scale fusion, utilizing input channels of 128 dimensions.

2. Frozen, Full Finetuning, and LoRA. We define the differences between these fine-tuning methods.

Frozen: The VFM backbone is frozen and SMoE modules are removed. Only the shallow CNN (Fig.2a) and subsequent GRU modules are trainable.

Table 2. Zero-shot Non-Lambertian Generalization. Comparison with state-of-the-art models. Networks trained on SceneFlow. We use the officially provided weights.

Model	>2 px (%)	>4 px (%)	>6 px (%)	>8 px (%)	Avg. (px)
PSMNet [2]	34.5	24.8	20.5	17.8	7.30
RAFTStereo [9]	17.8	13.1	10.8	9.24	3.60
Selective-RAFT [16]	20.0	15.1	12.5	10.9	4.12
Selective-IGEV [16]	18.5	14.2	12.1	10.8	4.38
DLNR [22]	18.6	14.6	12.6	11.2	3.97
SMoEStereo (Ours)	**11.3**	**7.16**	**6.47**	**5.13**	**2.09**

Full Finetuning: All components are trainable.

LoRA: The single LoRA layer (fixed rank=128) is used within each ViT block. Besides, Tab. 5 explores varying ranks to demonstrate SMoE's dynamic design effectiveness.

4. More Ablation Studies

In this section, we systematically evaluate the components of our framework through four key analyses. First, we conduct cross-domain generalization ablations on SceneFlow, dissecting the contribution of each SMoE component to domain-shift robustness. Next, we investigate architectural compatibility, validating SMoE's integration with diverse robust training frameworks (i.e., DKT framework [20]) and classical stereo architectures (e.g., IGEVStereo [17], PSMNet [2]). Additionally, We further analyze the computational efficiency of MoE LoRA and MoE Adapters by profiling their dynamic token allocation patterns across network layers, quantifying parameter savings under varying scene complexities. Concurrently, we evaluate the effectiveness of MoE balance losses in ensuring equitable expert utilization, demonstrating their role in stabilizing training while maintaining task performance. Finally, we assess data scalability by measuring performance gains from incremental synthetic data integration.

Ablation of Main Components. Table 3 presents the cross-domain generalization accuracy results, demonstrating that the integration of MoE LoRA and MoE Adapter layers significantly enhances disparity estimation performance compared to the vanilla VFM baseline (ID = 1). This improvement can be attributed to the inherent capability of our MoE architecture to adaptively learn domain-invariant features while preserving the transferable knowledge acquired from dense prediction tasks, thereby fostering robust generalization across diverse scenarios. However, the increased computational overhead associated with additional MoE modules (ID = 4) highlights a trade-off between performance and efficiency. To address this, our proposed decision network effectively reduces redundancy by selectively activating the most relevant MoE components, achieving a balance between computational efficiency and high performance (ID = 6). Notably, replacing the learned usage policy with a randomly generated one of comparable computational cost (ID = 5) results in a significant decline in accuracy, underscoring the critical role of the learned policy in optimizing expert selection and ensuring superior cross-domain adaptability.

Dynamic Expert Selection Mechanism. We conduct a comprehensive performance evaluation of SMoE across varying ranks of Low-Rank Adaptation (LoRA) for domain generalization, as detailed in Table 4. The zero-shot performance exhibits notable variability across different target domains, highlighting the critical influence of LoRA rank selection on model adaptability and generalization capabilities. This variability underscores the necessity of dynamically optimizing rank configurations to tailor the model's representational capacity to the unique characteristics of each domain, thereby maximizing cross-domain performance and ensuring robust generalization in diverse scenarios. Overall, this fully demonstrates the effectiveness of dynamic expert selection mechanisms with varying ranks.

Comparing SMoE with Different Rank of LoRA. We conduct a comprehensive performance comparison of SMoE against Low-Rank Adaptation (LoRA) models with varying ranks (e.g., r=4, 8, 32, 64, 128) for cross-domain generation tasks, as detailed in Table 5. Our analysis reveals stark variability in zero-shot generalization across distinct target domains (e.g., indoor and outdoor scenes) For various target domains, the zero-shot performance demonstrates significant variability, influenced by the specific ranks of Low-Rank Adaptation (LoRA) models. This variability underscores the importance of selecting appropriate ranks to optimize performance across different domains. Naively adopting the single uniform LoRA is inadequate for robust stereo matching.

DKT framework [20] on our SMoEStereo. DKT [20] is a robust training framework that was trained on real-world target datasets and demonstrates strong generalization capabilities across diverse datasets. In this section, we provide a comprehensive experimental analysis of the DKT framework, highlighting its strengths and limitations. Specifically, under identical DKT settings, our method achieves significant performance improvements over DKT-RAFT, as evidenced in Tab. 6. These results underscore the effectiveness of our proposed Sparse Mixture of Experts (SMoE) design, demonstrating its ability to enhance robustness and adaptability in challenging scenarios.

Compatibility of Different Baselines. Our SMoE design enhances the generalization performance of different baselines. The zero-shot results of all baselines are consistently improved within our SMoE framework, as detailed in Tab. 8. For example, the Bad-error rate of PSMNet [2] on each dataset decreases by 32%, 22%, 37%, 47%, and 48%, respectively. A significant improvement is achieved in Middlebury since there is much abundant semantic infor-

Table 3. Cross-domain performance ablation study trained on SceneFlow. DAMV2 (ViT-Base) used.

ID	MoE LORA	MoE Adapter	Decision Network	Random Decision	KITTI 2012 D1_All	KITTI 2015 D1_All	Middlebury Bad 2.0	ETH3D Bad 1.0	Number of MoE	Used Params. (M)
1	-	-	-	-	11.7	15.4	24.6	16.2	0	0
2	✓	-	-	-	4.31	4.93	7.60	2.56	12	2.29
3	-	✓	-	-	4.45	5.03	7.43	2.41	12	4.49
4	✓	✓	-	-	**4.19**	**4.79**	7.14	**1.99**	24	6.77
5	✓	✓	-	✓	4.51	5.19	8.32	3.04	14	2.86
6	✓	✓	✓	-	4.22	4.86	**7.05**	2.10	14	2.86

Table 4. Ablations of the dynamic selection of MoE LoRA layers.

Setting	VFM	KIT 2012 Bad 3.0	KIT 2015 Bad 3.0	Middle Bad 2.0	ETH3D Bad 1.0
MoE LoRA (Rank = 4)	DAMV2	4.55	5.01	7.93	3.08
MoE LoRA (Rank = 8)	DAMV2	4.60	5.11	8.01	2.93
MoE LoRA (Rank = 16)	DAMV2	4.74	5.17	7.82	3.39
MoE LoRA (Rank = 32)	DAMV2	4.37	4.97	8.10	2.87
SMoE w.o/ MoE Adapter	DAMV2	**4.31**	**4.93**	**7.60**	**2.56**
SMoE	DAMV2	**4.22**	**4.86**	**7.05**	**2.10**

Table 5. Zero-shot performance comparison of the proposed SMoE against other finetuning methods. Params (M) Train/Test refers to the 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 Bad 3.0	KIT 2015 Bad 3.0	Middle Bad 2.0	ETH3D Bad 1.0
DAMV2 [18] (ViT-base)	LoRA [4] (rank=4)	0.15/0.15	4.62	5.31	7.92	3.12
	LoRA [4] (rank=8)	0.30/0.30	**4.44**	5.07	8.22	3.08
	LoRA [4] (rank=16)	0.59/0.59	4.81	5.43	**7.64**	2.87
	LoRA [4] (rank=32)	1.18/1.18	4.56	5.21	8.43	**2.79**
	LoRA [4] (rank=64)	2.36/2.36	4.51	5.19	7.92	2.94
	LoRA [4] (rank=128)	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 [5] (ViT-base)	LoRA [4] (rank=4)	0.15/0.15	**4.41**	5.04	7.85	2.99
	LoRA [4] (rank=8)	0.30/0.30	4.63	5.30	**7.68**	3.02
	LoRA [4] (rank=16)	0.59/0.59	4.55	5.19	7.98	2.81
	LoRA [4] (rank=32)	1.18/1.18	4.48	**4.98**	8.35	2.76
	LoRA [4] (rank=64)	2.36/2.36	4.60	5.23	8.02	**2.66**
	LoRA [4] (rank=128)	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**

mation for reasoning. A similar improvement can be observed even when the robust CFNet [15] is utilized as the baseline, with the Bad-error rate being decreased by 15%, 17%, %, 52%, and 45%, respectively. Furthermore, we observe that integrating our SMoE into the IGEV [17] baseline results in a significant improvement in cross-domain performance. Notably, our approach does not require any specialized losses or additional modules to enhance domain generalization performance, making it broadly applicable to most learning-based stereo frameworks.

The Effectiveness of Kernel Sizes of Adapter Layers. As introduced in Section 3, the designed CNN adapters with different receptive fields incorporate local geometry priors of input samples into the ViT block. From Table 7, we present a comparative analysis of the cross-dataset performance of the proposed SMoEStereo, utilizing individual Adapter experts with varying kernel sizes. Notably, different datasets demonstrate distinct optimal Adapter experts.

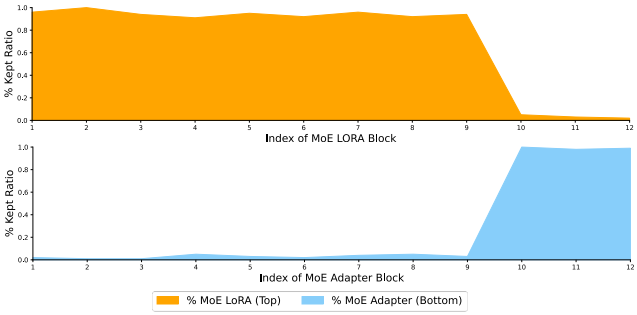

Figure 1. Computational cost throughout the network. The kept/activated ratio of MoE LoRA modules (top) and MoE Adapter modules (bottom) throughout the backbone are reported.

Our SMoE method adeptly identifies the most suitable local feature extractor for each input, achieving superior results compared to employing a single fixed Adapter. It is important to note that the SMoE approach introduces only negligible additional activated parameters, underscoring that the performance improvements are attributed to the proposed MoE learning scheme rather than scaling up the model. In summary, these findings comprehensively validate the effectiveness of our SMoE design.

Computational Saving throughout the Network. SMoEStereo leverages computational redundancy to enhance the efficiency of VFMs. We collect usage policies on MoE LoRA and MoE Adapter selection predicted by our method across four real-world datasets, illustrating the distribution of computational cost (i.e., percentage of MoE LoRA/Adapter retained) throughout the backbone. As shown in Fig. 1, SMoEStereo strategically allocates more computation to MoE LoRA layers in the earlier stages of the network while reserving MoE Adapter layers for the latter stages. This allocation suggests an optimization where MoE LoRA layers, which require less computational effort, handle initial processing, and MoE Adapter layers, functioning as decoders, manage more complex tasks in later stages of the ViT layers. This approach not only balances the computational load but also ensures efficient processing and superior performance across different scenarios, highlighting the robustness and adaptability of SMoEStereo.

The Effectiveness of MoE Balance Loss. The router network often assigns disproportionately large weights to a

Table 6. Comparison with DKT-RAFT trained on KITTI.

Method	KIT 2012		KIT 2015		DrivingStereo					MIDDLE	ETH3D
	Noc	All	Bg	All	Sunny	Cloudy	Rainy	Foggy	Avg.	(Bad 2.0)	(Bad 1.0)
DKT-RAFT [20]	1.43	1.85	1.65	1.88	1.85	1.46	1.32	5.44	2.52	7.51	2.28
DKT-SMoE	1.17	1.56	1.47	1.63	1.46	1.21	1.12	4.38	2.04	7.13	1.99

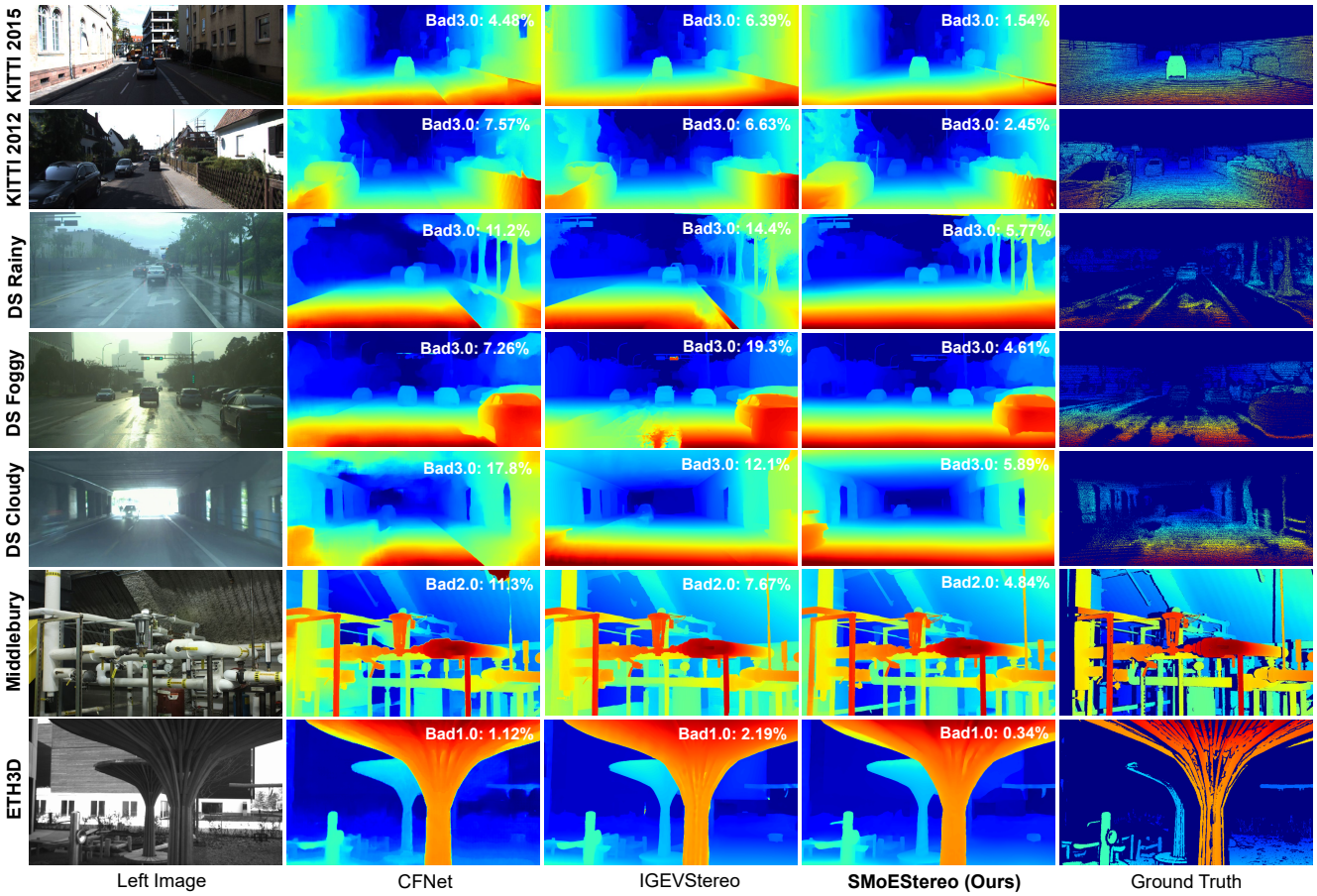

Figure 2. Zero-shot performance of CFNet [15], IGEVStereo [17], and SMoEStereo on diverse scenes. Note that, all models are only trained on the SceneFlow dataset. In the middle row, DS denotes the DrivingStereo dataset.

Table 7. Impacts of the kernel sizes of Adapter layers and the effectiveness of SMoE.

Setting	KITTI 2015		Middlebury		ETH3D	
	EPE	D1_All	EPE	Bad 2.0	EPE	Bad 1.0
Kernel Size = 3	**0.62**	**1.54**	**0.74**	**4.26**	0.16	0.69
Kernel Size = 5	0.66	1.62	0.77	4.61	**0.15**	0.68
Kernel Size = 7	0.68	1.74	0.81	4.66	0.17	0.75
Kernel Size = 9	0.64	1.62	0.76	4.41	**0.15**	**0.66**
SMoE	**0.60**	**1.51**	**0.71**	**4.12**	**0.15**	**0.63**

few experts [14], leading to overfitting issues. To counteract this, we introduce an MoE balance loss component to ensure equal importance among all experts, preventing the model from getting trapped in local optima. In this subsection, we evaluate the impacts of the proposed $\mathcal{L}_{blc}$. Table 9

Table 8. Zero shot performance comparisons on different baselines.

Models	VFM Capacity	KIT 2012 Bad 3.0	KIT 2015 Bad 3.0	Middle Bad 2.0	ETH3D Bad 1.0
PSMNet [2]	-	6.0	6.3	15.8	10.2
PSMNet-SMoE	ViT-Base [18]	**4.1**	**4.9**	**8.3**	**5.4**
CFNet [15]	-	4.7	5.8	15.3	5.8
CFNet-SMoE	ViT-base [18]	**4.0**	**4.8**	**7.4**	**3.2**
IGEV [17]	-	5.1	5.6	7.1	3.6
IGEV-SMoE	ViT-base [18]	**4.1**	**4.7**	**6.8**	**2.0**

presents the cross-dataset evaluation performance with different values of the $\mathcal{L}_{blc}$ loss weight λ_1 in Eq. (11), where $\lambda_1 = 0$ represents no MoE loss applied. The application of $\mathcal{L}_{blc}$ effectively mitigates router overfitting and consis-

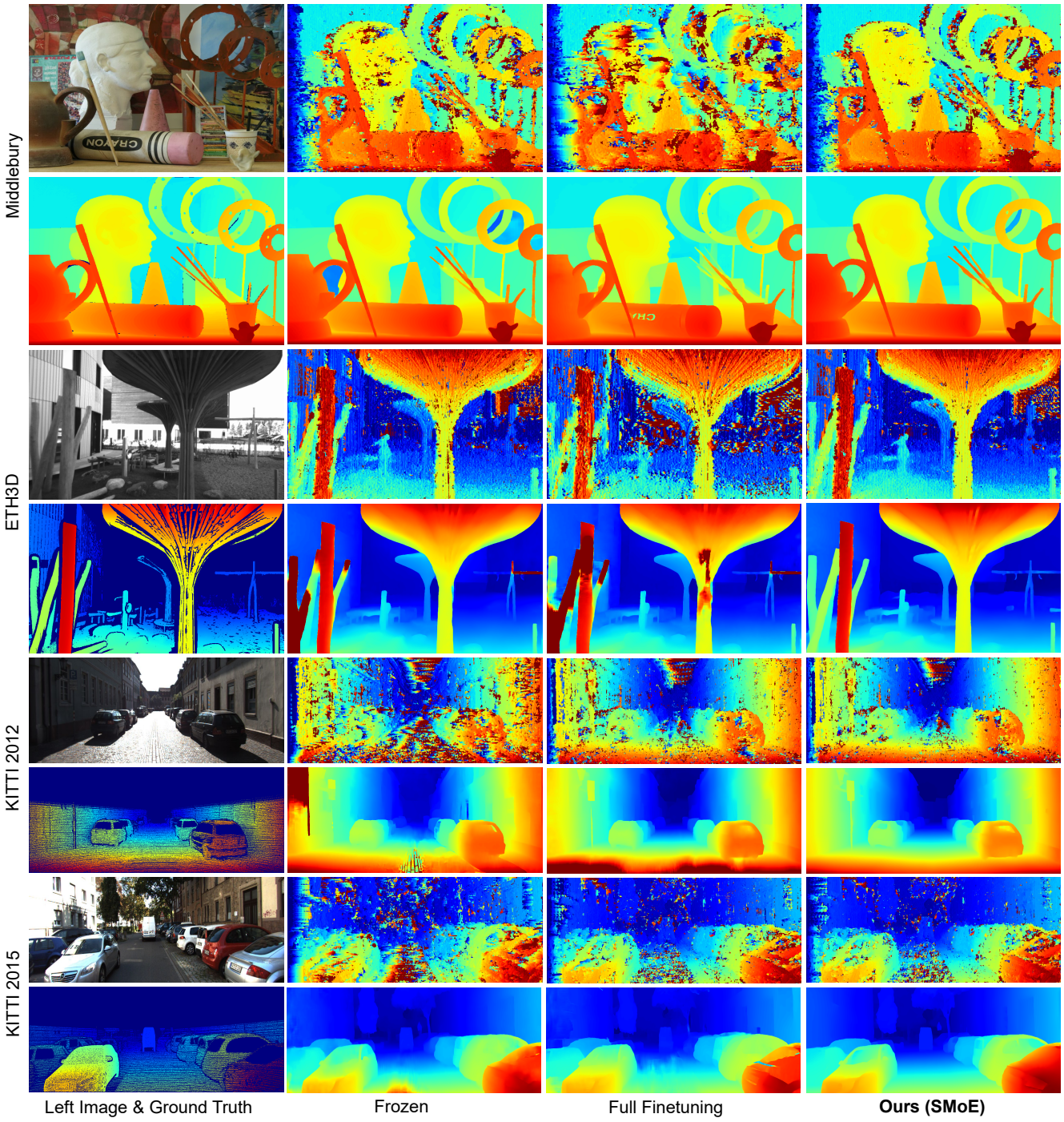

Figure 3. Zero-shot visual comparison of Frozen, Full-Finetuning, and our SMoE on diverse scenes. Note that, all models are only trained on the SceneFlow dataset. The rows (1 & 3 & 5 & 7) indicate Winner-Take-All (WTA) disparity from the feature correlation (1/4 scale) achieved by the dot products among left and right features before the subsequent cost aggregation network. WTA disparity, when enhanced with our SMoE-equipped pre-trained VFM, exhibits significantly less noise compared to other vanilla finetuning methods for VFMs.

tently enhances model generalizability. The model achieves optimal generalizability with λ_1 set at 1. These findings underscore the critical importance of balanced expert contributions within the MoE framework. The $\mathcal{L}_{blc}$ component ensures that the learned representations are more diverse and generalizable across different datasets. This is especially crucial in practical applications where data distribution can vary significantly. The optimal performance ob-

Table 9. Effectiveness of the proposed loss components.

λ	KITTI 2015		Middlebury		ETH3D	
	EPE	D1_All	EPE	Bad 2.0	EPE	Bad 1.0
$\lambda = 0$	0.69	1.68	0.79	4.74	0.18	0.82
$\lambda = 0.1$	0.65	1.60	0.75	4.42	0.17	0.74
$\lambda = 0.5$	**0.60**	**1.48**	**0.74**	**4.17**	0.15	0.67
$\lambda = 1$	**0.60**	**1.51**	**0.71**	**4.12**	**0.15**	**0.63**
$\lambda = 5$	0.63	1.57	0.74	4.25	**0.16**	**0.71**

Table 10. Gains from data capacity. SF, VK2, and CRE denote the SceneFlow, Virtual KITTI2, and CREStereo datasets.

SF	VK2	CRE	KITTI 2012 D1_All	KITTI 2015 D1_All	Middlebury Bad 2.0	ETH3D Bad 1.0
√	-	-	4.22	4.86	‘7.05	2.10
√	√	-	3.21	3.90	7.43	2.19
√	√	√	**3.15**	**3.78**	**6.79**	**1.90**

served at $\lambda_1 = 1$ suggests that balancing expert utilization and regularization is vital.

Gains from More Synthetic Data. In the main paper, we trained our model solely on the SceneFlow dataset [10]. This section investigates how increased data capacity affects SMoEStereo's performance with additional training samples. We consider two additional synthetic datasets: Virtual KITTI2 [1], a synthetic outdoor driving dataset with 20K samples, and CREStereo [8], a synthetic dataset with diverse delicate structures. Table 10 demonstrates that performance on KITTI 2012/2015 and ETH3D consistently improves with more datasets used for training. However, performance on Middlebury [13] deteriorates when VKITTI2 samples are included. We argue that this is due to distinct domain shifts in the additional data, weakening generalization on Middlebury. Conversely, using the CREStereo dataset for training improves generalization on Middlebury. In summary, these results suggest that SMoEStereo's generalization ability is enhanced with increased training data capacity.

5. SMoE *vs.* Multi Experts.

To demonstrate the effectiveness of SMoE's dynamic selection of optimal LoRA and Adapter experts, we compare it with Multi Experts (Multi-E), which aggregates outputs from all designed experts. As shown in Table 11, the SMoE mechanism selectively activates sparse experts, achieving a $1.25\times$ speedup in training and a $1.12\times$ speedup in inference. Despite these gains, SMoE outperforms Multi-E in cross-dataset performance, highlighting its effectiveness in selecting the optimal expert for robust stereo matching. This suggests that naively aggregating multiple experts may not be optimal, as it can suppress informative features while introducing noise.

Table 11. SMoE *vs.* Multi-E (All experts are involved in MoE).

Setting	Training Time (s)	Inference Time (s)	Number of MoE	KITTI 2015 EPE	KITTI 2015 D1_All	Middlebury EPE	Middlebury Bad 2.0	ETH3D EPE	ETH3D Bad 1.0
Multi-E	2.10 iter/s	4.68 iter/s	24	**0.59**	**1.46**	0.76	4.37	0.17	0.78
SMoE	**2.63 iter/s**	**5.26 iter/s**	14	0.60	1.51	**0.71**	**4.12**	**0.15**	**0.63**

6. Zero-shot Performance on Diverse Scenarios.

We illustrate the predicted disparity maps across various scenarios in Fig. 2 for a qualitative comparison. Compared to previous state-of-the-art methods [15, 17], our method shows exceptional generalization across diverse scenes, including outdoor, indoor, and challenging weather conditions. Additionally, our approach produces significantly less noisy disparity maps compared to vanilla finetuning methods for VFMs, showcasing enhanced robustness and effectiveness. This improvement highlights our method's ability to maintain finer details and generate smoother results, particularly in zero-shot settings, as illustrated in Figure 3. The robust features produced by our SMoE fully demonstrate the superiority of our approach in enhancing the overall zero-shot performance of stereo matching outcomes, ensuring fine detail preservation and smoothness even in complex and varied environments.

References

[1] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*, 2020. 6

[2] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5410–5418, 2018. 2, 4

[3] 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. 1

[4] 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

[5] 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. 1, 3

[6] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision (IJCV)*, 128(7):1956–1981, 2020. 1

[7] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural net-

works. In *Workshop on challenges in representation learning, ICML*, page 896. Atlanta, 2013. 1

[8] 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. 6

[9] 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. 1, 2

[10] 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

[11] 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. 1

[12] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *ArXiv preprint*, 2021. 1

[13] 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

[14] 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. 4

[15] 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. 3, 4, 6

[16] Xianqi Wang, Gangwei Xu, Hao Jia, and Xin Yang. Selective-stereo: Adaptive frequency information selection for stereo matching. *arXiv preprint arXiv:2403.00486*, 2024. 1, 2

[17] 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. 2, 3, 4, 6

[18] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024. 3, 4

[19] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2636–2645, 2020. 1

[20] Jiawei Zhang, Jiahe Li, Lei Huang, Xiaohan Yu, Lin Gu, Jin Zheng, and Xiao Bai. Robust synthetic-to-real transfer for stereo matching. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20247–20257, 2024. 2, 4

[21] 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. 1

[22] 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. 1, 2