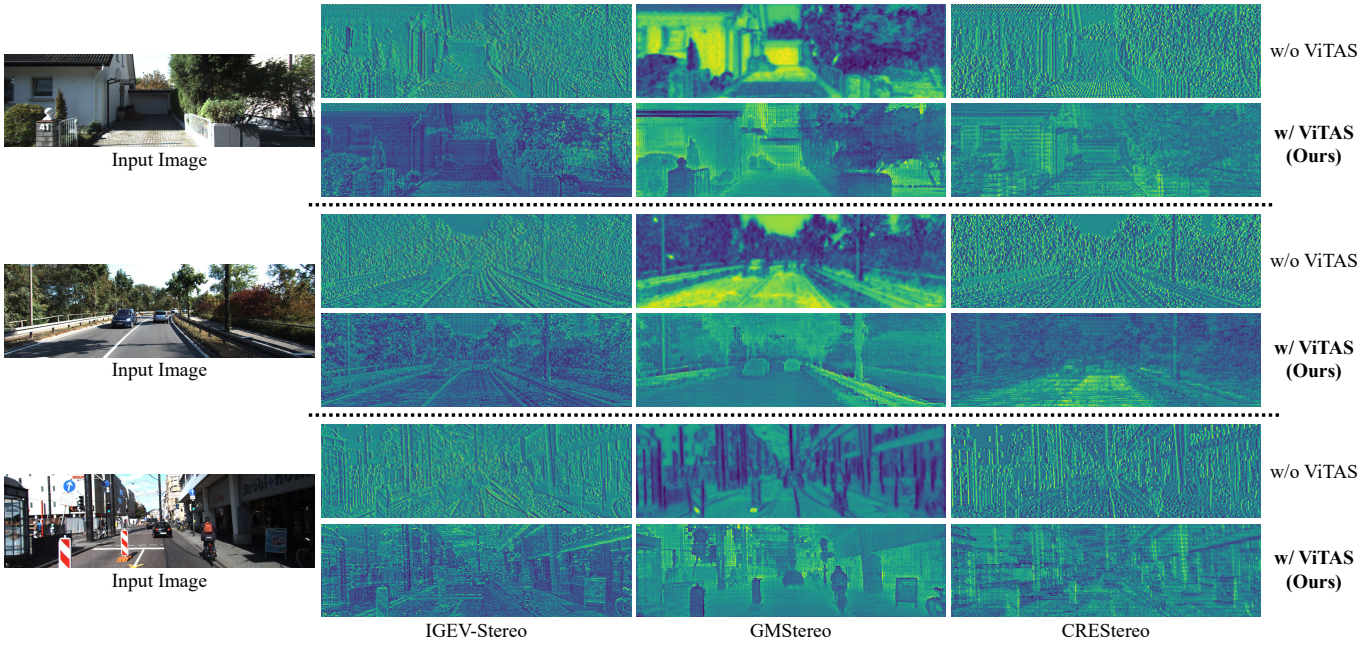

Fig. 1. Visualizations of extracted features at 1/4 of full image resolution without and with our proposed ViTAS applied.

TABLE I
QUANTITATIVE EXPERIMENTAL RESULTS OF CROCO-STEREO [1].

Network	Dataset for Model Fine-tuning	KITTI Eval		Midd Eval		ETH3D	
		EPE (px)	D1-all (%)	EPE (px)	D1-all (%)	EPE (px)	D1-all (%)
Original CroCo-Stereo [1]	KITTI Train	1.16	6.17	8.80	25.1	51.5	97.4
	Midd Train	6.70	35.0	2.04	9.90	50.7	88.0
Modified CroCo-Stereo	KITTI Train	0.86	3.26	8.15	25.4	50.7	89.1
	Midd Train	4.05	44.9	1.87	8.82	62.9	85.2

SUPPLEMENTARY MATERIAL

A. Feature Visualization

To fully verify the improvements of our ViTAS in deep features compared with the conventional feature extraction DCNNs, we visualize the feature maps extracted by the three SoTA stereo matching networks [2]–[4] without and with our proposed ViTAS applied using principal component analysis (PCA) [5]. As shown in Fig. 1, feature maps extracted by ViTAS contain richer details, such as object boundaries, compared to the traditional backbone networks. Moreover, feature maps extracted by ViTAS result in larger PCA values at informative areas such as disparity discontinuities and small-scale objects, while yielding smaller PCA values at tractable areas with minimal texture and disparity variations. We further visualize the feature pyramids extracted by the feature encoder of IGEV-Stereo [2] and our proposed ViTAStereo using PCA. The results presented in Fig. 2 suggest that these feature improvements are also evident in feature maps at deep layers. In general, these feature improvements at object boundaries and small-scale objects further demonstrate the superiority of our ViTAS compared to traditional backbone networks.

B. Generalizability Evaluation on CroCo-Stereo

To answer the question of whether cost volumes are becoming less critical or even expendable in SoTA stereo matching networks, we conduct an additional experiment with the recently published cost volume-free network, CroCo-Stereo, as detailed in Table I and Fig. 3. In this experiment, we compare the performance of the original CroCo-Stereo against a modified version, where its ViT backbone is replaced with our utilized vision foundation model (VFM). It can be observed in Table I that both the original and modified CroCo-Stereo achieve accurate disparity estimation results on the KITTI Eval and Midd Eval datasets after being fine-tuned on the KITTI Train and Midd Train datasets, respectively. However, their disparity estimation accuracy notably decreases on the other two evaluation datasets, with significantly higher EPE and D1-all observed on the ETH3D dataset. The results presented in Fig. 3 emphasize a significant issue with both networks: scale ambiguity. This is evident in the relatively smaller disparity estimation results on the Midd Eval dataset after fine-tuning on the KITTI Train dataset, as well as the relatively larger disparity estimation results on the ETH3D dataset. Hence, we argue that cost volumes remain essential, particularly when

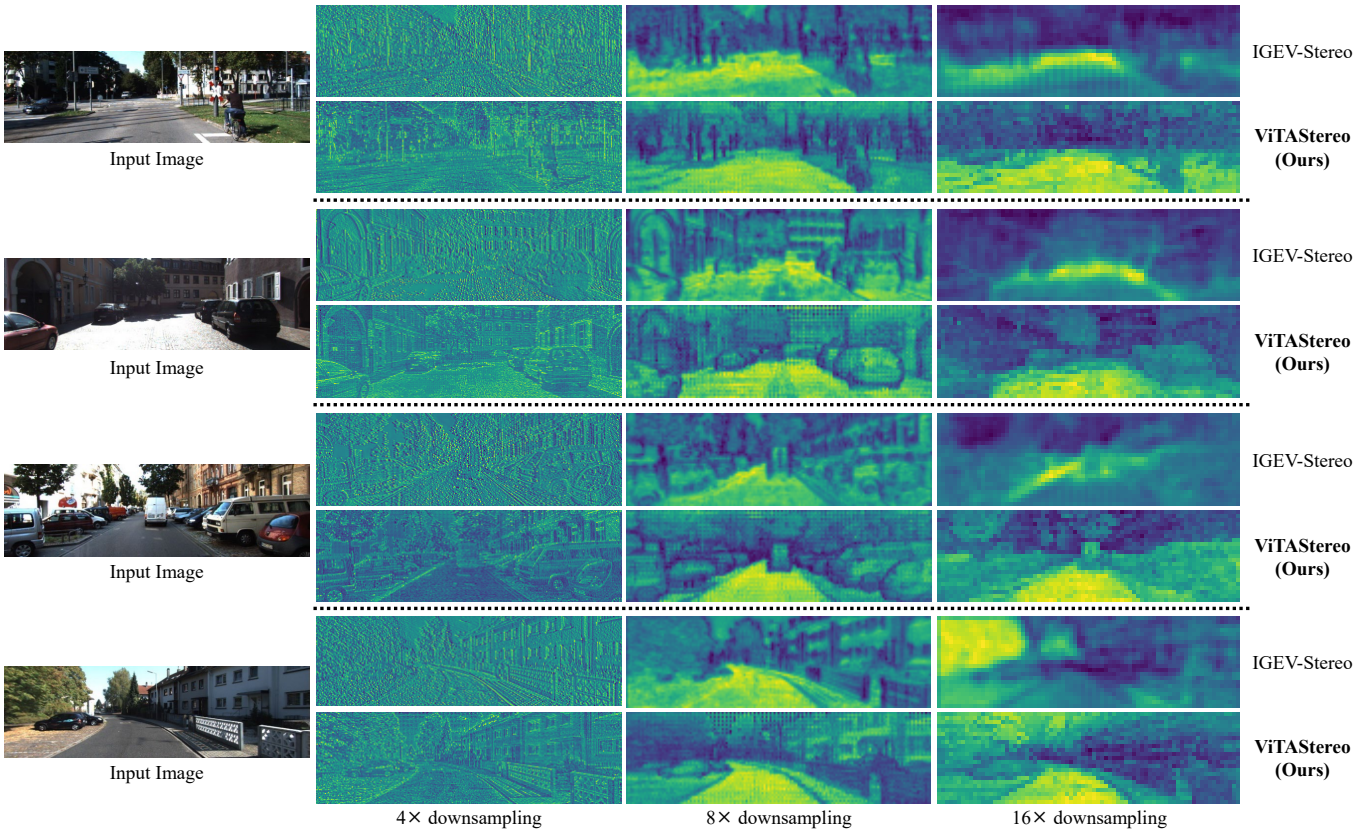

Fig. 2. Visualizations of the feature pyramids extracted by the feature encoders of IGEV-Stereo [2] and ViTAStereo.

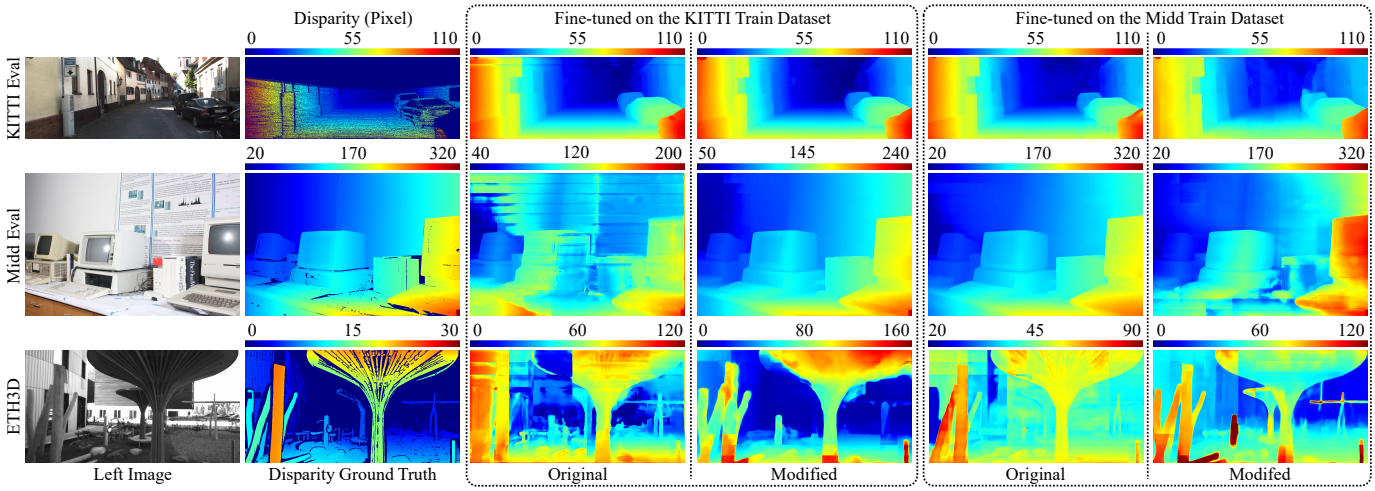

Fig. 3. Qualitative experimental results of CroCo-Stereo [1].

aiming for generalizable stereo matching networks.

References

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

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

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

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

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