DMS: Diffusion-Based Multi-Baseline Stereo Generation for Improving Self-Supervised Depth Estimation

Supplementary Material

In this supplementary material, we provide additional details and results to complement the main paper. Section 5.1 outlines the experimental setup, including dataset configurations and hyperparameters. Section 5.2 provides more experimental results both in the image quality of the DMS and subsequent training of the self-supervised stereo-matching networks. Finally, in Section 5.3 we illustrate more multi-baseline stereo image results across different datasets using the proposed DMS.

5.1. Additional Implementation Details.

5.1.1. Additional Implementation Details of Diffusion-Based Multi-Baseline Stereo Generator (DMS).

In this section, we detail the training process of the Diffusion-based Multi-baseline Stereo Generator (DMS) across various datasets to ensure reproducibility. We implement the DMS using Pytorch with the diffusers [67] as the code base and utilizing the Stable DiffusionV2 [55] as the initial parameter weight. We further report the computation resources that are needed in the inference stage in Table 8.

SceneFlow. For training on the SceneFlow dataset, we employ the widely-used FlyingThings3D test set [36, 45, 46, 71, 80, 81, 91] and a training subset of 19,984 images from the FlyingThings3D set, filtering out scenes where occlusion exceeds 80%. To conform to the input size requirements of the Stable Diffusion Model, which necessitates divisibility by 8, the original images are resized from 540×960 to 576×960 using top padding. We optimize memory usage by employing a batch size of 1 with a gradient accumulation equivalent to a batch size of 16. Optimization is performed using the Adam optimizer with a constant learning rate of $2e-5$ under half-precision (float16) settings. The training spans 20 epochs, with the resultant model used for both evaluation and generating new views. Inference utilizes a DDPM scheduler with a step of 50 denoising process for view synthesis. The efficacy of the generated views is quantitatively assessed in Table 2.

KITTI 2015 & 2012. For the limited view of the KITTI 2015 and 2012 datasets, we fine-tune the Diffusion model on the KITTI raw dataset [20] which compromises over 43,482 stereo images, where we split the 400 images and 394 images containing in KITTI 2015 and KITTI 2012 dataset and use the left views for training. We pad the original resolution of 375×1242 and 374×1238 into 284×1248 to meet the input size requirements of the Stable Diffusion Model. We use the same optimizer and learning rate that is adopted in training the ScenceFlow model and training for 10 epochs to get the final model. Inference utilizes a DDPM scheduler with a step of 32 denoising process for view synthesis. For further used in unsupervised stereo matching, we fine-tune the KITTI-raw pre-trained model on KITTI 2012 and KITTI 2015 datasets, respectively. Note that we both generate views on the KITTI raw dataset for improvement in the performance of monocular depth estimators as outlined in Section 4.4.

MPI-Sintel Dataset. For the MPI-Sintel dataset, we partition the dataset into a training set and an evaluation set using a $9:1$ ratio. The training utilizes the "final pass" images, and we adjust the original resolution from 436×1024 to 440×1024 to accommodate the model's input requirements. The optimization parameters, including the optimizer and learning rate, are consistent with those used for the SceneFlow model. The training duration is set to 50 epochs to finalize the model. During inference, view synthesis is performed using a DDPM scheduler with a 50-step denoising process.

CARLA Dataset. Existing stereo datasets typically contain only *left* and *right* views, making it challenging to evaluate the quality of extended multi-baseline images. To address this, we utilize the CARLA simulator [15] to generate a synthetic multi-baseline stereo dataset with 1000 image pairs (*left*, *center*, *right*, *left-left*, *right-right*) under 15 diverse weather conditions (e.g., *ClearNoon*, *WetNight*). The dataset is split into training and testing sets with a 9:1 ratio. For fine-tuning DMS, we adopt the same training protocol as the KITTI dataset, training for 50 epochs with KITTI pre-trained weights as initialization. During inference, we also utilize a DDPM scheduler with a step of 32 denoising processes for view synthesis. The performance of the generated multi-baseline images can be outlined in Table 10.

5.1.2. Additional Implementation Details of Training the Self-Supervised Depth Estimators.

In this section, we provide a detailed description of the implementation details used for training self-supervised depth networks with multi-baseline images generated by DMS. This includes ablation studies settings and experiments across different datasets on both self-supervised stereo matching and monocular depth estimation.

Implementation Details of the Ablation Studies. We leveraged PASMNet [71] as our baseline to evaluate the effectiveness of our multi-baseline stereo images in improving disparity estimation in self-supervised stereo matching settings. We conduct ablation studies on two NVIDIA 3090 GPUs with PyTorch. For the SceneFlow dataset, training

was performed on the FlyThings3D subset (as detailed in Section 5.1.1) and evaluated on the SceneFlow official test set. The model was trained with a batch size of 8, a disparity range of 192, and 100,000 steps. The initial learning rate was set to 1×10^{-4} and reduced using cosine decay. The checkpoint with the lowest End-Point Error (EPE) on the validation set was selected for final evaluation, the results of which are presented in Table 5. Fine-tuning on the KITTI dataset followed the protocol in [95], using 160 image pairs for training and 40 for validation, with weights pre-trained on SceneFlow. The training procedure was consistent with SceneFlow, incorporating data augmentation techniques from [81], such as random cropping and adjustments to brightness, saturation, and contrast. For the Sintel-MPI dataset, the model was trained from scratch using the same parameters as SceneFlow to ensure methodological consistency. This approach validates the robustness of our method across diverse datasets.

Implementation Details of Training the Self-Supervised Stereo Matching Networks. Besides the SceneFlow dataset, we further test the performance of the DMS integrated into existing stereo-matching networks to validate the 'plug-in-and-play' ability on the KITTI dataset. For the KITTI 2015 benchmark, we deployed the model which was initially pre-trained on the SceneFlow dataset and subsequently fine-tuned on a combined dataset of KITTI 2012 and 2015, encompassing 394 images. We selected the model with the optimal D1 value for submission to the official KITTI 2015 benchmark to obtain our final results. Considering the limited availability of open-source self-supervised stereo-matching methods, we extended the applicability of our proposed DMS by adapting supervised networks like RaftStereo [39] and IGEVStereo [80] to self-supervised settings using photometric warping loss, demonstrating the method's versatility and broad potential for adaptation.

Implementation Details of Training the Self-Supervised Monocular Depth Estimators. All models were trained on the full KITTI Eigen training set (45,200 stereo image pairs) and validated on a small set (4,424 images). We selected models with the lowest validation error and tested them on the KITTI Eigen test set (697 images). For SDFANet [97], marked as * in the main paper, we made small changes to its original loss computation to incorporate our proposed DMS properly during training. SDFANet predicts a disparity cost volume with shape $B \times C \times D^* \times H \times W$, where D^* is the number of disparity candidates. Disparity is estimated using Soft Argmin on the third dimension. The loss is computed by warping the left image with all disparity candidates, warping each sub-cost of one disparity candidate with shape $B \times C \times H \times W$ using the corresponding disparity, and calculating a weighted sum to synthesize the right image. The loss is then computed using the synthesized right image and the input right image. To match other methods and incorporate the proposed DMS, we used the same warping loss as the other three compared methods. We warped the right image to the left using the estimated disparity from SDFANet and computed the loss with the input left image. Similarly, by incorporating DMS, the disparity is used to warp the left-left image, right-right image, and center image to compute additional losses.

5.1.3. Evaluation Details of the Self-Supervised Depth Estimations.

Occlusion and Out-of-Frame Mask Generation for Evaluation. The MPI-Sintel dataset provides the ground truth occlusion and the out-of-frame mask for evaluation, but SceneFlow and KITTI did not provide such specific masks for evaluation. To address this issue, we use the same strategy used in [46] by using the left-right consistency to generate the occlusion mask and the ground disparity to calculate the out-of-frame mask. The process can be described as follows:

$$M_{occ} = \begin{cases} 1 & \text{if } D_{\Delta}(x,y) \geq 1, \\ 0 & \text{otherwise,} \end{cases} \tag{7}$$

$$M_{oof} = \begin{cases} 1 & \text{if } D_{shift}(x,y) < 0, \\ 0 & \text{otherwise,} \end{cases} \tag{8}$$

$$D_{\Delta}(x,y) = |d_l(x,y) - d_r(x + d_l(x,y), y)|, \tag{9}$$

$$D_{shift}(x,y) = x - d_l(x,y), \tag{10}$$

where M_{occ} and M_{off} represents the generated occlusion masks and the out-of-frame mask, respectively. And the d_l and d_r are the ground truth disparity map. For the reason that the KITTI dataset only provides the ground-truth sparse disparity maps for the left images, which makes it difficult to directly apply the left-right consistency check to generate the occlusion masks, following the strategy utilized in [46], we use a pre-trained model [11] to generate the pseudo-left and pseudo-left disparities, the left-right consistency check between the pseudo disparity maps for the left view and the right view is applied to generate a pseudo occlusion mask for performance evaluation.

Evaluation Metrics for Self-Supervised Stereo Matching. To showcase the effectiveness of our proposed DMS, especially on ill-conditioned regions, we report the End-Point-Error (EPE) and the $>$ 3px outliers(percentage of the error bigger than 3 pixels) on overall regions, the occluded regions, and out-of-frame regions, respectively. The definition of the EPE is as follows:

$$EPE(d, \hat{d}) = |d - \hat{d}|. \tag{11}$$

For the performance on the KITTI 2015 validation set, we report the $> 3px$ which describes the outliner ratio of the

Table 8. Computation resources for utilizing the DMS to generate multi-baseline stereo images across different datasets with different resolutions. Note the inference time and GPU Memory are tested on a single NVIDIA A6000 GPU.

Dataset	Denoising Steps	Inference Time Per Image	GPU Memory
SceneFlow [47]	50	5.34 s	7.61 G
KITTI [19]	32	4.12 s	6.82 G
MPI-Sintel [6]	50	6.04 s	6.72 G
CARLA [15]	32	4.30 s	7.05G

predicted disparity. where can be described as follows:

$$> 3px = \frac{N_{\Delta e > 3px}}{N_{total}}, \quad \Delta e = |d - \hat{d}|, \tag{12}$$

where N means the number of pixels.

For the KITTI 2015 testing benchmark, we follow the official evaluation protocol to report the D1-value as shown in Table 6.

Evaluation Metrics for Self-Supervised Monocular Depth Estimation. We evaluate each method using several metrics from prior work [16], which uses the predicted depth d^* and GT depth $\hat{d}^*$ in meters to compute the errors:

$$AbsRel = \frac{1}{|T|} \sum_{d^* \in T} \frac{|d^* - \hat{d}^*|}{\hat{d}^*}, \tag{13}$$

$$SqRel = \frac{1}{|T|} \sum_{d^* \in T} \frac{||d^* - \hat{d}^*||^2}{\hat{d}^*}, \tag{14}$$

$$RMSE = \sqrt{\frac{1}{|T|} \sum_{d^* \in T} ||d^* - \hat{d}^*||^2}, \tag{15}$$

$$RMSE(\log) = \sqrt{\frac{1}{|T|} \sum_{d^* \in T} ||\log d^* - \log \hat{d}^*||^2}, \tag{16}$$

$$A(thr) = \max(\frac{d_i^*}{\hat{d}_i^*}, \frac{\hat{d}_i^*}{d_i^*}) = \delta < thr, \tag{17}$$

where T denotes all the test pixels in all test image samples, and A1, A2, A3 denote the thr be set as 1.25, 1.25^2, and 1.25^3 respectively.

5.2. Addition Experimental Results.

In addition to the experimental results presented in the main paper, this section provides supplementary evaluations to thoroughly demonstrate the validity of our proposed Diffusion-Based Multi-Baseline Stereo Generation (DMS) and its impact on improving the performance of self-supervised depth estimation methods.

5.2.1. Diffusion-Based Multi-Baseline Stereo Generation

More Ablations on Rescale-Factor *X*. We select a rescale-factor of 2.0 in the paper as the *center* view provides the most effective representation of the intermediate view between left and right images. To further justify this choice, we extend the ablation studies in Table 9 with additional candidates (0.5, 1.5, 3.0), as shown in Table.1 below. These factors produce denser intermediate views, such as $\frac{1}{3}$ $l \rightarrow r$ and $\frac{2}{3}$ $l \rightarrow r$, improving performance over the baseline. However, the most effective configurations remain the default left-left (+*ll*), right-right (+*rr*), and center view (+*c*), highlighted in gray, which cover most out-of-view and occluded regions. Moreover, applying all new views together further improves overall EPE to 1.22, closely matching the 1.24 achieved with only +*ll*+*rr*+*c*. Considering both performance and efficiency, we chose a rescale-factor of 2.0 for intermediate view generation in the paper.

Table 9. Additional ablations on the KITTI2015 validation set with varying rescale factors. *Generate left-left and right-right images with a 0.5 scale.

Factor X	Views Used	EPE↓			D1↓		
		All	Occ	Oof	All	Occ	Oof
-	*l*,*r*	1.48	4.38	9.26	7.7	39.6	64.2
0.5	+ *ll* *+*rr**	1.37	3.97	7.90	6.8	36.1	48.0
1.0	+ *ll*+*rr*	1.34	3.83	7.82	6.5	34.4	42.8
1.5	+$\frac{2}{3}$ $l \rightarrow r$	1.40	4.22	8.17	7.1	38.1	51.7
2.0	+*c*	1.36	4.14	7.64	6.7	37.0	49.6
3.0	+$\frac{1}{3}$ $l \rightarrow r$	1.41	4.16	8.33	7.1	36.7	52.9
All Above	+ *all*	1.22	3.44	7.20	5.6	31.7	39.3

Computation Resources Analysis. We report the computation resources for utilizing the DMS to generate multi-baseline stereo images using an NVIDIA A6000 GPU and Intel i9-13900KF CPU. The image resolutions for inference are 540 × 960, 384 × 1280, 436 × 1024, and 540 × 960 for SceneFlow, KITTI, MPI-Sintel, and CARLA datasets, respectively. This demonstrates that our DMS can efficiently perform inference on a single GPU with less than 8GB of memory, highlighting its practical applicability.

Multi-Baseline Stereo Image Evaluation on CARLA. Stereo datasets typically provide only *left* and *right* views, limiting the evaluation of extended multi-baseline images. To address this, we generate a synthetic multi-baseline stereo dataset using the CARLA simulator [15], consisting of 1000 image pairs (*left*, *center*, *right*, *left-left*, *right-right*) across 15 weather conditions (e.g., *ClearNoon*, *WetNight*). During training, we also only used the left and right view to train the DMS and used the pre-trained DMS model to generate multi-baseline stereo images. As shown in Table 10, we report the PSNR and SSIM of the generated views including the novel view left-left, right-right, and center view, respectively. The check-marked annotations indicate the newly generated perspectives (left-left and right-right) obtained using our proposed inference method. These views exhibit the comparable performance of PSNR and SSIM to

Table 10. Novel view quality evaluations on synthesis dataset created by CARLA [15] simulator. We report the PSNR and SSIM for both the left view, right view, left-left view, right-right view, and center view, respectively.

Generated View	Input View	Direction Prompt	Upscaling	Novel View	PSNR	SSIM
Left	Right	*to left*	-		23.52	0.76
Right	Left	*to right*	-		23.06	0.75
Left-Left	Left	*to left*	-	✓	23.63	0.76
Right-Right	Right	*to right*	-	✓	22.71	0.72
Center	Left	*to right*	×2	✓	21.44	0.72

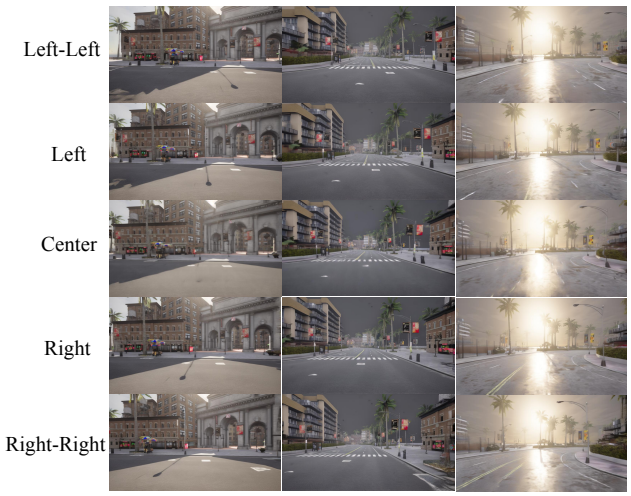

Figure 9. Multi-baseline stereo images generation using proposed DMS on the CARLA synthesis dataset.

the rendered left and right views, despite the absence of ground truth left-left and right-right views during training. While the generated intermediate views show a slight decrease in PSNR, their SSIM remains consistent with other views. This demonstrates that the multi-baseline images produced by DMS maintain geometric consistency, making them valuable for improving self-supervised depth estimation. Further visualization results are illustrated in Section 5.2.2 in this supplementary material.

5.2.2. Self-Supervised Stereo Matching

In addition to the ablation studies presented in the main paper, we further evaluated the impact of multi-baseline stereo images generated by our DMS on self-supervised stereo-matching performance using the KITTI 2012 dataset. Same as the main paper, we also use PASMNet [71] as the baseline for self-supervised stereo-matching training.

As detailed in Table 11, the results show that adding *ll* and *rr* significantly reduces both EPE and outlier ratios in all regions compared to the baseline, particularly improving occluded and out-of-frame areas. Further incorporating the center view (*c*) yields the best performance, achieving

Table 11. Ablation Studies On KITTI 2012 dataset for self-supervised streo matching.The terms *ll*, *rr*, and *c* refer to the left-left, right-right, and center views, respectively. Results include End-Point Error (EPE) and outlier ratios (errors $>$ 3px) across general, occluded, and out-of-frame regions."Occ" and "Oof" represent the occluded regions and the out-of-frame regions, respectively.

Method	KITTI 2012					
	EPE↓			>3px(%)↓		
	All	Occ	Oof	All	Occ	Oof
Baseline	1.44	5.01	15.58	7.5	40.0	66.3
+ *ll* + *rr*	1.24	4.58	10.65	6.0	37.7	53.4
+ *c*	1.41	4.96	14.83	6.8	40.1	65.6
+ *ll* + *rr* +*c*	1.16	4.39	9.77	5.81	36.1	51.4

Figure 10. Multi-baseline stereo images generation using proposed DMS on the KITTI 2015 dataset.

the lowest EPE and outlier percentages, demonstrating the effectiveness of multi-baseline integration in enhancing geometric consistency and depth estimation robustness.

5.3. Additional Visualization Results.

5.3.1. Multi-Baseline Stereo Image Generation Results Visualization.

Figure 10, Figure 11, Figure 12, and Figure 9 showcase additional visualizations of the proposed DMS model applied to the SceneFlow, KITTI, MPI-Sintel, and CARLA datasets. These figures illustrate the model's capability to synthesize novel views along the epipolar line, leveraging

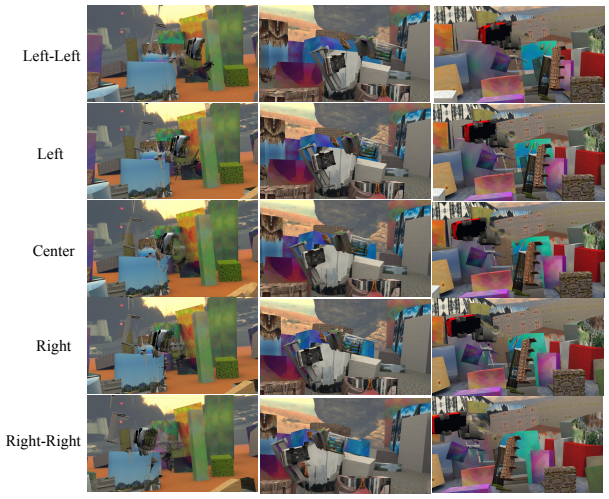

Figure 11. Multi-baseline stereo images generation using proposed DMS on the SceneFlow dataset.

Figure 12. Multi-baseline stereo images generation using proposed DMS on the MPI-Sintel dataset.

directional prompts to extend stereo baselines. The results highlight the versatility and robustness of DMS in handling diverse scenarios, from highly controlled synthetic datasets to complex real-world environments, while maintaining geometric consistency.

References

[1] Mohammadreza Armandpour, Huangjie Zheng, Ali Sadeghian, Amir Sadeghian, and Mingyuan Zhou. Re-imagine the negative prompt algorithm: Transform 2d diffusion into 3d, alleviate janus problem and beyond. *arXiv preprint arXiv:2304.04968*, 2023. 3

[2] Juan Luis Gonzalez Bello and Munchurl Kim. Deep 3d-zoom net: Unsupervised learning of photo-realistic 3d-zoom. *arXiv preprint arXiv:1909.09349*, 2019. 2

[3] Juan Luis Gonzalez Bello and Munchurl Kim. Deep 3d pan via local adaptive" t-shaped" convolutions with global and local adaptive dilations. In *International Conference on Learning Representations*, 2020. 2

[4] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 1

[5] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. *arXiv preprint arXiv:2410.02073*, 2024. 1

[6] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *Proc. Eur. Conf. Comp. Vis.*, pages 611–625. Springer-Verlag, 2012. 5, 6, 7, 3

[7] Eric R Chan, Koki Nagano, Matthew A Chan, Alexander W Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. Generative novel view synthesis with 3d-aware diffusion models. In *ICCV*, 2023. 3

[8] J. Chang and Y. Chen. Pyramid stereo matching network. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 5410–5418, 2018. 1

[9] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. *arXiv preprint arXiv:2303.13873*, 2023. 3

[10] Yiwen Chen, Chi Zhang, Xiaofeng Yang, Zhongang Cai, Gang Yu, Lei Yang, and Guosheng Lin. It3d: Improved text-to-3d generation with explicit view synthesis. *arXiv preprint arXiv:2308.11473*, 2023. 3

[11] Xuelian Cheng, Yiran Zhong, Mehrtash Harandi, Yuchao Dai, Xiaojun Chang, Hongdong Li, Tom Drummond, and Zongyuan Ge. Hierarchical neural architecture search for deep stereo matching. *Advances in neural information processing systems*, 33:22158–22169, 2020. 2

[12] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 3

[13] Congyue Deng, Chiyu Jiang, Charles R Qi, Xinchen Yan, Yin Zhou, Leonidas Guibas, Dragomir Anguelov, et al. Nerdi: Single-view nerf synthesis with language-guided diffusion as general image priors. In *CVPR*, 2023. 3

[14] Zijun Deng, Xiangteng He, Yuxin Peng, Xiongwei Zhu, and Lele Cheng. Mv-diffusion: Motion-aware video diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7255–7263, 2023. 3

[15] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017. 4, 6, 1, 3

[16] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014. 5, 3

[17] Xiule Fan, Soo Jeon, and Baris Fidan. Occlusion-aware self-supervised stereo matching with confidence guided raw disparity fusion. In *2022 19th Conference on Robots and Vision (CRV)*, pages 132–139, 2022. 1, 2

[18] Miaojie Feng, Junda Cheng, Hao Jia, Longliang Liu, Gangwei Xu, and Xin Yang. Mc-stereo: Multi-peak lookup and cascade search range for stereo matching. In *2024 International Conference on 3D Vision (3DV)*, pages 344–353. IEEE, 2024. 7, 8

[19] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 3354–3361. IEEE, 2012. 5, 6, 3

[20] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013. 1

[21] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 270–279, 2017. 2, 5, 8

[22] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 3828–3838, 2019. 2, 5, 8

[23] Juan Luis Gonzalez Bello and Munchurl Kim. Deep 3d pan via local adaptive" t-shaped" convolutions with global and local adaptive dilations. In *International Conference on Learning Representations, ICLR 2020*, pages 1–20. 2020 International Conference on Learning Representations, 2020. 2, 6

[24] Jiatao Gu, Alex Trevithick, Kai-En Lin, Joshua M Susskind, Christian Theobalt, Lingjie Liu, and Ravi Ramamoorthi. Nerfdiff: Single-image view synthesis with nerf-guided distillation from 3d-aware diffusion. In *ICML*, 2023. 3

[25] Tongfan Guan, Chen Wang, and Yun-Hui Liu. Neural markov random field for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5459–5469, 2024. 6

[26] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2485–2494, 2020. 2

[27] X. Guo, K. Yang, W. Yang, X. Wang, and H. Li. Group-wise correlation stereo network. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 3268–3277, 2019. 1

[28] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(2):328–341, 2007. 7, 8

[29] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Proc. Advances in Neural Inf. Process. Syst.*, 33:6840–6851, 2020. 3, 5

[30] Dominik Honegger, Torsten Sattler, and Marc Pollefeys. Embedded real-time multi-baseline stereo. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5245–5250, 2017. 2

[31] Yukun Huang, Jianan Wang, Yukai Shi, Xianbiao Qi, Zheng-Jun Zha, and Lei Zhang. Dreamtime: An improved optimization strategy for text-to-3d content creation. *arXiv preprint arXiv:2306.12422*, 2023. 3

[32] Saad Imran, Muhammad Umar Karim Khan, Sikander Mukaram, and Chong-Min Kyung. Unsupervised monocular depth estimation with multi-baseline stereo. In *Proc. British Machine Vis. Conf.*, 2020. 2

[33] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2024. 3

[34] Sameh Khamis, Sean Fanello, Christoph Rhemann, Adarsh Kowdle, Julien Valentin, and Shahram Izadi. Stereonet: Guided hierarchical refinement for real-time edge-aware depth prediction. In *Proc. Eur. Conf. Comp. Vis.*, pages 573–590, 2018. 7

[35] Jiabao Lei, Jiapeng Tang, and Kui Jia. Generative scene synthesis via incremental view inpainting using rgbd diffusion models. In *CVPR*, 2022. 3

[36] Ang Li and Zejian Yuan. Occlusion aware stereo matching via cooperative unsupervised learning. In *Proc. Asian Conf. Comp. Vis.*, pages 197–213. Springer, 2018. 1, 2, 7

[37] Peng Li, Yuan Liu, Xiaoxiao Long, Feihu Zhang, Cheng Lin, Mengfei Li, Xingqun Qi, Shanghang Zhang, Wenhan Luo, Ping Tan, et al. Era3d: High-resolution multiview diffusion using efficient row-wise attention. *arXiv preprint arXiv:2405.11616*, 2024. 3

[38] Chen-Hsuan Lin et al. Magic3d: High-resolution text-to-3d content creation. *arXiv preprint arXiv:2302.01335*, 2023. 3

[39] Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *2021 International Conference on 3D Vision (3DV)*, pages 218–227. IEEE, 2021. 7, 8, 2

[40] Miaomiao Liu, Xuming He, and Mathieu Salzmann. Geometry-aware deep network for single-image novel view synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4616–4624, 2018. 2

[41] Pengpeng Liu, Irwin King, Michael R Lyu, and Jia Xu. Flow2stereo: Effective self-supervised learning of optical flow and stereo matching. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 6648–6657, 2020. 2, 7

[42] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object, 2023. 3

[43] Xinhang Liu, Shiu-hong Kao, Jiaben Chen, Yu-Wing Tai, and Chi-Keung Tang. Deceptive-nerf: Enhancing nerf reconstruction using pseudo-observations from diffusion models. *arXiv preprint arXiv:2305.15171*, 2023. 3

[44] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 3

[45] Zihua Liu, Songyan Zhang, Zhicheng Wang, and Masatoshi Okutomi. Digging into normal incorporated stereo match-

ing. In *Proc. ACM Int. Conf. Multimedia*, pages 6050–6060, 2022. 1

[46] Zihua Liu, Yizhou Li, and Masatoshi Okutomi. Global occlusion-aware transformer for robust stereo matching. In *Proc. Winter Conf. on Appl. of Comp. Vis.*, pages 3535–3544, 2024. 1, 2

[47] 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 *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 4040–4048, 2016. 5, 6, 7, 3

[48] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 3061–3070, 2015. 5, 6, 7

[49] Don Murray and James J Little. Using real-time stereo vision for mobile robot navigation. *Autonomous Robots*, 8:161–171, 2000. 1

[50] Alex Nichol et al. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2

[51] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive separable convolution. In *Proceedings of the IEEE international conference on computer vision*, pages 261–270, 2017. 6

[52] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive separable convolution. In *Proceedings of the IEEE international conference on computer vision*, pages 261–270, 2017. 2

[53] Masatoshi Okutomi and Takeo Kanade. A multiple-baseline stereo. *IEEE Transactions on pattern analysis and machine intelligence*, 15(4):353–363, 1993. 2

[54] Ben Poole, Ajay Jain, Pieter Abbeel, et al. Dreamfusion: Text-to-3d using 2d diffusion models. *arXiv preprint arXiv:2209.14988*, 2022. 3

[55] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 10684–10695, 2022. 2, 3, 5, 1

[56] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 3

[57] Hoigi Seo, Hayeon Kim, Gwanghyun Kim, and Se Young Chun. Ditto-nerf: Diffusion-based iterative text to omni-directional 3d model. *arXiv preprint arXiv:2304.02827*, 2023. 3

[58] Junyoung Seo, Wooseok Jang, Min-Seop Kwak, Jaehoon Ko, Hyeonsu Kim, Junho Kim, Jin-Hwa Kim, Jiyoung Lee, and Seungryong Kim. Let 2d diffusion model know 3d-consistency for robust text-to-3d generation. *arXiv preprint arXiv:2303.07937*, 2023. 3

[59] Zhelun Shen, Yuchao Dai, and Zhibo Rao. Cfnet: Cascade and fused cost volume for robust stereo matching. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 13906–13915, 2021. 7

[60] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3

[61] Hideyuki Suenaga, Huy Hoang Tran, Hongen Liao, Ken Masamune, Takeyoshi Dohi, Kazuto Hoshi, and Tsuyoshi Takato. Vision-based markerless registration using stereo vision and an augmented reality surgical navigation system: a pilot study. *BMC Medical Imaging*, 15(1):1–11, 2015. 1

[62] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Viewset diffusion:(0-) image-conditioned 3d generative models from 2d data. *arXiv preprint arXiv:2306.07881*, 2023. 3

[63] Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. *arXiv preprint arXiv:2307.01097*, 2023.

[64] Ayush Tewari, Tianwei Yin, George Cazenavette, Semon Rezchikov, Joshua B Tenenbaum, Frédo Durand, William T Freeman, and Vincent Sitzmann. Diffusion with forward models: Solving stochastic inverse problems without direct supervision. *arXiv preprint arXiv:2306.11719*, 2023. 3

[65] Christina Tsalicoglou, Fabian Manhardt, Alessio Tonioni, Michael Niemeyer, and Federico Tombari. Textmesh: Generation of realistic 3d meshes from text prompts. *arXiv preprint arXiv:2304.12439*, 2023. 3

[66] Hung-Yu Tseng, Qinbo Li, Changil Kim, Suhib Alsisan, Jia-Bin Huang, and Johannes Kopf. Consistent view synthesis with pose-guided diffusion models. In *CVPR*, 2023. 3

[67] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022. 1

[68] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *CVPR*, 2023. 3

[69] Jiheng Wang, Abdul Rehman, Kai Zeng, Shiqi Wang, and Zhou Wang. Quality prediction of asymmetrically distorted stereoscopic 3d images. *IEEE Transactions on Image Processing*, 24(11):3400–3414, 2015. 5, 6

[70] Jiheng Wang, Shiqi Wang, Kede Ma, and Zhou Wang. Perceptual depth quality in distorted stereoscopic images. *IEEE Transactions on Image Processing*, 26(3):1202–1215, 2016. 5, 6

[71] Longguang Wang, Yulan Guo, Yingqian Wang, Zhengfa Liang, Zaiping Lin, Jungang Yang, and Wei An. Parallax attention for unsupervised stereo correspondence learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(4):2108–2125, 2020. 1, 2, 7, 4

[72] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4): 600–612, 2004. 5

[73] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and

diverse text-to-3d generation with variational score distillation. *arXiv preprint arXiv:2305.16213*, 2023. 3

[74] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628*, 2022. 3

[75] Jamie Watson, Michael Firman, Gabriel J Brostow, and Daniyar Turmukhambetov. Self-supervised monocular depth hints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2162–2171, 2019. 2

[76] Jinbo Wu, Xiaobo Gao, Xing Liu, Zhengyang Shen, Chen Zhao, Haocheng Feng, Jingtuo Liu, and Errui Ding. Hd-fusion: Detailed text-to-3d generation leveraging multiple noise estimation. *arXiv preprint arXiv:2307.16183*, 2023. 3

[77] Jianfeng Xiang, Jiaolong Yang, Binbin Huang, and Xin Tong. 3d-aware image generation using 2d diffusion models. *arXiv preprint arXiv:2303.17905*, 2023. 3

[78] Junyuan Xie, Ross Girshick, and Ali Farhadi. Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 842–857. Springer, 2016. 6

[79] Junyuan Xie, Ross Girshick, and Ali Farhadi. Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 842–857. Springer, 2016. 2

[80] Gangwei Xu, Xianqi Wang, Xiaohuan Ding, and Xin Yang. Iterative geometry encoding volume for stereo matching. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 21919–21928, 2023. 1, 6, 7, 8, 2

[81] Haofei Xu and Juyong Zhang. Aanet: Adaptive aggregation network for efficient stereo matching. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 1959–1968, 2020. 5, 1, 2

[82] Guorun Yang, Hengshuang Zhao, Jianping Shi, Zhidong Deng, and Jiaya Jia. Segstereo: Exploiting semantic information for disparity estimation. In *Proc. Eur. Conf. Comp. Vis.*, pages 636–651, 2018. 1, 2, 7

[83] 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 *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 899–908, 2019. 1

[84] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024. 1

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

[86] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9043–9053, 2023. 1

[87] Paul Yoo, Jiaxian Guo, Yutaka Matsuo, and Shixiang Shane Gu. Dreamsparse: Escaping from plato's cave with 2d frozen diffusion model given sparse views. *CoRR*, 2023. 3

[88] Chaohui Yu, Qiang Zhou, Jingliang Li, Zhe Zhang, Zhibin Wang, and Fan Wang. Points-to-3d: Bridging the gap between sparse points and shape-controllable text-to-3d generation. *arXiv preprint arXiv:2307.13908*, 2023. 3

[89] Jason J Yu, Adam W Harley, and Konstantinos G Derpanis. Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pages 3–10. Springer, 2016. 1

[90] Jason J. Yu, Fereshteh Forghani, Konstantinos G. Derpanis, and Marcus A. Brubaker. Long-term photometric consistent novel view synthesis with diffusion models. In *ICCV*, 2023. 3

[91] F. Zhang, V. Prisacariu, R. Yang, and P. H. S. Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2019. 1

[92] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 3

[93] Songyan Zhang, Zhicheng Wang, Qiang Wang, Jinshuo Zhang, Gang Wei, and Xiaowen Chu. Ednet: Efficient disparity estimation with cost volume combination and attention-based spatial residual. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 5429–5438, 2021. 1

[94] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for image restoration with neural networks. *IEEE Transactions on computational imaging*, 3(1):47–57, 2016. 5

[95] Chao Zhou, Hong Zhang, Xiaoyong Shen, and Jiaya Jia. Unsupervised learning of stereo matching. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 1567–1575, 2017. 1, 2, 6, 7

[96] Hang Zhou, David Greenwood, and Sarah Taylor. Self-supervised monocular depth estimation with internal feature fusion. *arXiv preprint arXiv:2110.09482*, 2021. 2, 8

[97] Zhengming Zhou and Qiulei Dong. Self-distilled feature aggregation for self-supervised monocular depth estimation. In *European Conference on Computer Vision*, pages 709–726. Springer, 2022. 2, 8

[98] Zhizhuo Zhou and Shubham Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *CVPR*, 2023. 3

[99] Joseph Zhu and Peiye Zhuang. Hifa: High-fidelity text-to-3d with advanced diffusion guidance. *arXiv preprint arXiv:2305.18766*, 2023. 3