SUPPLEMENTARY MATERIALS for Calibration Tokens

Suchisrit Gangopadhyay[1*], Jung-Hee Kim[2*], Xien Chen[1*], Patrick Rim[1], Hyoungseob Park[1], Alex Wong[1]
Yale University[1], Michigan State University[2]
{rit.gangopadhyay,xien.chen,patrick.rim,hyoungseob.park,alex.wong}@yale.edu, kimjun84@msu.edu

Table 1. **Additional experiments.**

	Experiment	Model	RMSE↓	δ_1↑
ScanNet++	Self-supervised (ours)	UniDepth	<u>0.244</u>	<u>0.766</u>
	Supervised (ours)	UniDepth	**0.242**	**0.769**
	Fisheye space	UniDepth	0.280	0.755
	Same token added	UniDepth	0.290	0.752
KITTI-360	Self-supervised (ours)	UniDepth	<u>2.040</u>	**0.664**
	Supervised (ours)	UniDepth	**1.994**	<u>0.651</u>
	Fisheye space	UniDepth	2.110	0.618
	Same token added	UniDepth	2.062	0.631

A. Additional Experiments

To further validate our claims and design choices, we evaluated the performance of some other possible designs, which can be seen in Tab. 1.

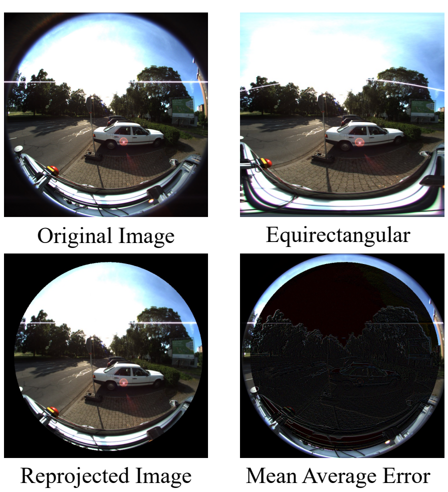

Figure 1. **Visualization of lossy training objective.**

Fisheye Frame Loss. In the main paper, we claimed that computing loss in the fisheye reference frame would perform worse because we would need to transform the perspective output, which would give us a lossy training objective. We have validated that claim with another experiment in the table. Furthermore, Fig. 1 shows the information loss caused by distorting to the equirectuangular space, which is used by some baseline methods. In this example with an image from KITTI-360, there is a 17.23% loss in the image pixels.

Same Token Added. In addition to the "Layer-wise" and "Single Token" approaches for adding our calibration tokens that we discussed in the main paper, we tried taking the same token, but adding and removing it after each transformer block, so it remains unchanged for each transformer block. We found that this approach still does not outperform the "Layer-wise" approach.

Supervised Loss. Because our loss is self-supervised (using output from a pretrained model as the training objective), we also evaluate the performance of our method when training with perspective ground truth instead of the perspective model output. As expected, there is a slight performance increase. However, it would be more cost-effective to use the self-supervised approach because the improvement is limited, especially in the indoor setting. This further validates the robustness of the baseline foundation model for perspective images.

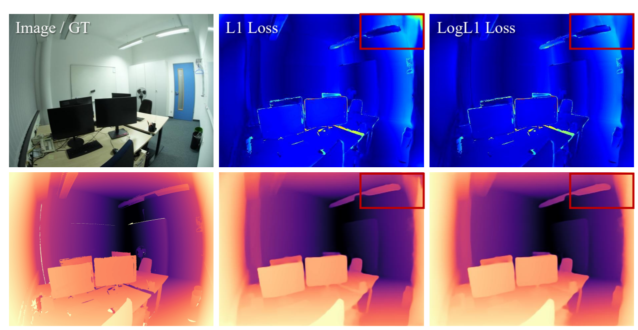

Figure 2. **Validation on LogL1 loss.** We evaluate the effectiveness of our LogL1 loss by comparing a single-layer token baseline with an additional LogL1 loss. Incorporating LogL1 loss helps model to mitigate artifacts in the highlighted border regions of fisheye images, leading to improved visual consistency.

*Equal contribution

Additional Qualitative Results. We further demonstrate our contribution with the 3D reconstruction results as shown in Fig. 3. This result provides evidence of our contribution toward foundational model latent embeddings to be aligned to fisheye images with our fully self-supervised training. Additionally, we provide qualitative results to validate our LogL1 loss. As can be seen with the Fig. 2, the logL1 loss helps the model mitigate the impact of artifacts caused by severe distortions, leading to more stable improvements on fisheye images, as reflected in the depth map and error map results. Fig. 4 and Fig. 5 visualize the depth estimation comparison with and without the calibration token (C.T.) on the ScanNet++ and KITTI-360 datasets, respectively.

B. Additional Details

B.1. Foundational Depth Estimation Models

MiDAS, DepthAnything-V1(ViT-L). Following the pipeline of [31, 61], these models utilizes a Vision Transformer Large encoder and a specialized decoder head for single-view depth estimation. Its training covers a massive corpus of perspective images drawn from both indoor and outdoor domains, aiming at robust zero-shot performance. Despite strong generalization within pinhole-camera distributions, it lacks dedicated mechanisms for counteracting severe lens distortions (e.g., fisheye or panoramic).

UniDepth-V2(ViT-S). UniDepth-V2 [30] leverages a Vision Transformer Small backbone, paired with a camera self-prompting routine to address moderate discrepancies in intrinsic parameters. However, when confronted with extreme distortions typical of ultra-wide or fisheye lenses, it is insufficient to recover geometry reliably. In both cases, we demonstrate how a small set of learnable calibration tokens (see main paper) can bridge the gap from perspective to fisheye images without retraining the full models.

B.2. Datasets

We provide further details on the datasets used for both training and testing.

Training Datasets: **NYU Depth V2** [37] ("NYUv2") consists of 464 diverse indoor scenes (e.g., living rooms, offices). It contains about 400,000 aligned RGB–depth pairs at 640×480 resolution. Following standard practice, approximately 1,500 depth points are chosen in each map via the Harris corner detector [14]. NYUv2 is a common benchmark for indoor depth tasks and serves here as one of our primary training sets.

IRS [46] compiles a large number of synthetic indoor environments, from small apartments to commercial interiors—each scene offering ground-truth depth rendered at resolutions comparable to 640×480. Its scale (up to 103,316 frames) and variety of virtual layouts supplement real data.

VOID [52] (Visual Odometry with Inertial and Depth) features about 58,000 frames taken in hallways, classrooms, and shared spaces, each accompanied by a sparse depth map at roughly 0.5% density (≈1,500 points).

Hypersim [36] is a photo-realistic synthetic dataset offering about 77,400 RGB–depth pairs. These scenes incorporate meticulously rendered geometry and lighting across various architectural styles (e.g., residential, museum-like structures). Hypersim's controlled yet visually realistic design helps our model see a wide spectrum of interior layouts even before encountering real-world test sets.

Waymo Open Dataset [40] contributes ∼230,000 camera–LiDAR frames across urban and suburban roads. Though heavily used for self-driving applications (e.g., detection, tracking), we leverage it here to extend our token training beyond the pure indoor scenario. The inclusion of Waymo frames exposes our method to outdoor scenes with larger view ranges and more complex lighting.

Testing Datasets: Our proposed approach is primarily evaluated on two real-world datasets that each incorporate fisheye or wide-FOV imaging. **ScanNet++** [63] is an extended collection of indoor RGB-D sequences, building on the popular ScanNet dataset but augmented with additional scenes and fisheye captures. We use the fisheye depth estimation ground truth to verify how our framework handles substantial lens distortion indoors.

KITTI-360 [24] is an outdoor dataset focusing on large-scale mapping and autonomous driving. It contains 360° fisheye cameras and high-grade LiDAR depth. Scenes encompass suburban roads, semi-rural stretches, and detailed 3D annotations. Testing on KITTI-360 lets us measure the ability of our approach to generalize to wide-FOV imagery in challenging real-world driving contexts.

B.3. Implementations

All experiments used the same training hyperparameters: Adam optimizer with learning rate of 10^{-4} and $\beta_1 = 0.9, \beta_2 = 0.999$. For random fisheye distortion synthesis, we leveraged the polynomial distortion model introduced by Kannala & Brandt [16], using four distortion parameters (i.e., $N_k = 4$) within the range of $[-1.0, -0.01]$.

B.4. Evaluation Metrics

For the evaluation, we used metrics proposed by Eigen et al.[7]. Since our focus is on adapting monocular depth estimation to different visual modalities, we measure relative depth estimation performance to mitigate the gap introduced by fisheye images. This is crucial, as foundation models often suffer from a loss of general performance in such cases. Tab. 2 provides detailed equations used for evaluation. The *root mean squared error* (RMSE) measures deviation in the linear depth space. We further report a threshold-based accuracy, δ_1, which represents the percentage of pixels whose predicted depth is within a tight bound of the ground-truth

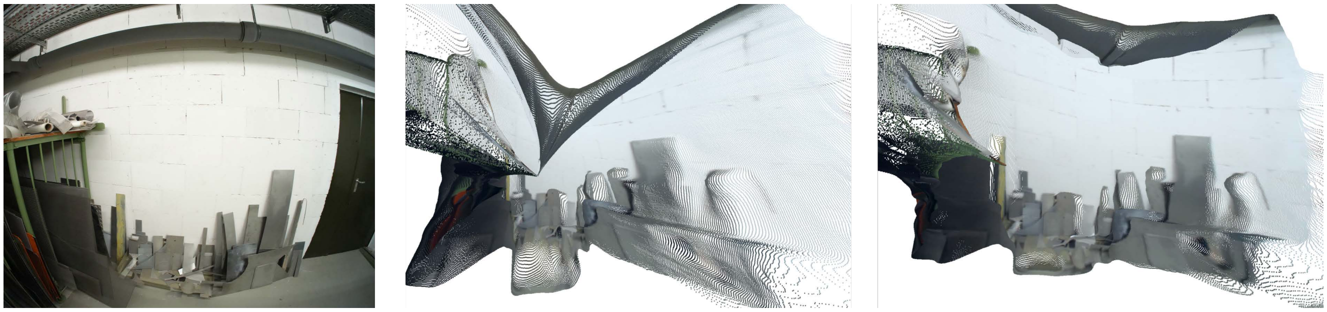

Figure 3. **3D reconstruction result of UniDepth predictions on ScanNet++ dataset.**

depth.

Metric	Definition
RMSE$\downarrow$	$\sqrt{\frac{1}{\lvert\Omega\rvert}\sum_{p\in\Omega}\left(\hat{d}(p)-d(p)\right)^2}$
$\delta_1\uparrow$	$\frac{1}{\lvert\Omega\rvert}\sum_{p\in\Omega}\mathbf{1}\Big(\max\Big(\frac{\hat{d}(p)}{d(p)},\frac{d(p)}{\hat{d}(p)}\Big)<1.25^1\Big)$

Table 2. **Error metrics for depth estimation.** These evaluation metrics compute the error between predicted depth values $\hat{d}(x)$ and ground truth depth values $d(x)$.

C. Discussion

Spatial applications are typically deployed on platforms (e.g., robots, autonomous vehicles, extended reality headsets) with multi-camera systems. Naturally, data collection is done on a specific platform that may differ from those used during deployment. This introduces a domain or covariate shift between the training and testing distributions. The focus of this paper is on the covariate shift introduced by fisheye cameras, which are common to many spatial platforms. While we demonstrate our method on monocular depth estimation [9–11, 22, 23, 43, 44, 47, 51, 65, 70], it is just one of many perception tasks that are affected by this covariate shift: We see further applications in optical flow [18–21, 39, 41, 67, 68], semantic segmentation [4, 13, 17, 49, 50, 53, 58, 69], image restoration [1, 64, 66] and stereo [2, 12, 45, 56, 60]. Further, many perception tasks follow the convention of projecting different sensor modalities onto the image reference frame for fusion. We envision our method to be applicable towards perception tasks [33, 57, 59] on multi-sensor platforms with camera and LiDAR and 3D reconstruction with camera and LiDAR [3, 6, 8, 15, 25, 26, 28, 29, 48, 52, 54, 55, 62] or radar [34, 38]. Finally, we see a connection between our method and continual learning [5, 27, 32, 35, 42] as our method aims extend to models to different cameras, e.g. perspective to fisheye, instead of 3D scenes while maintaining previously learned information, e.g., backward-compatibility.

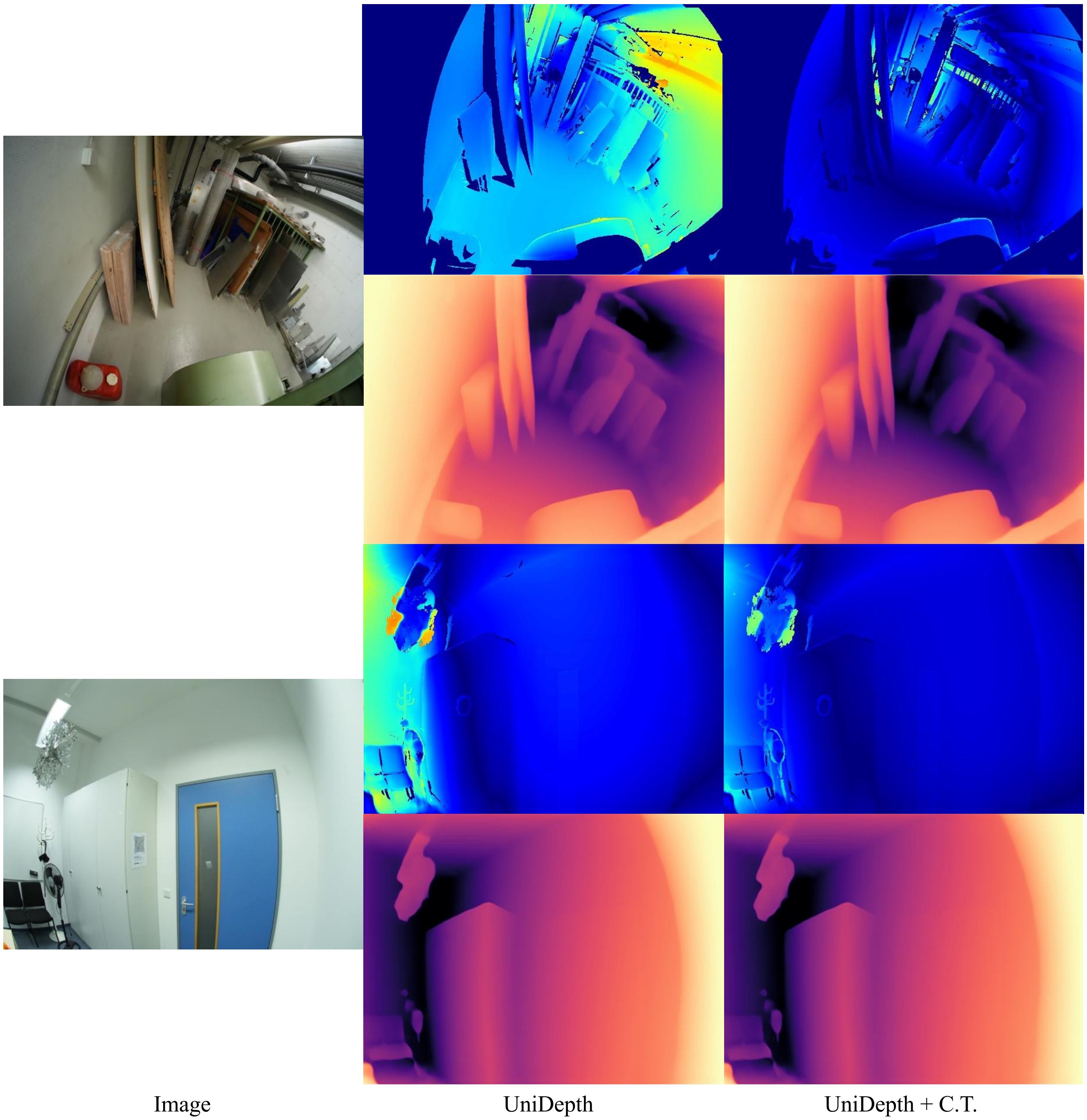

Figure 4. **Additional comparison results on ScanNet++ dataset.**

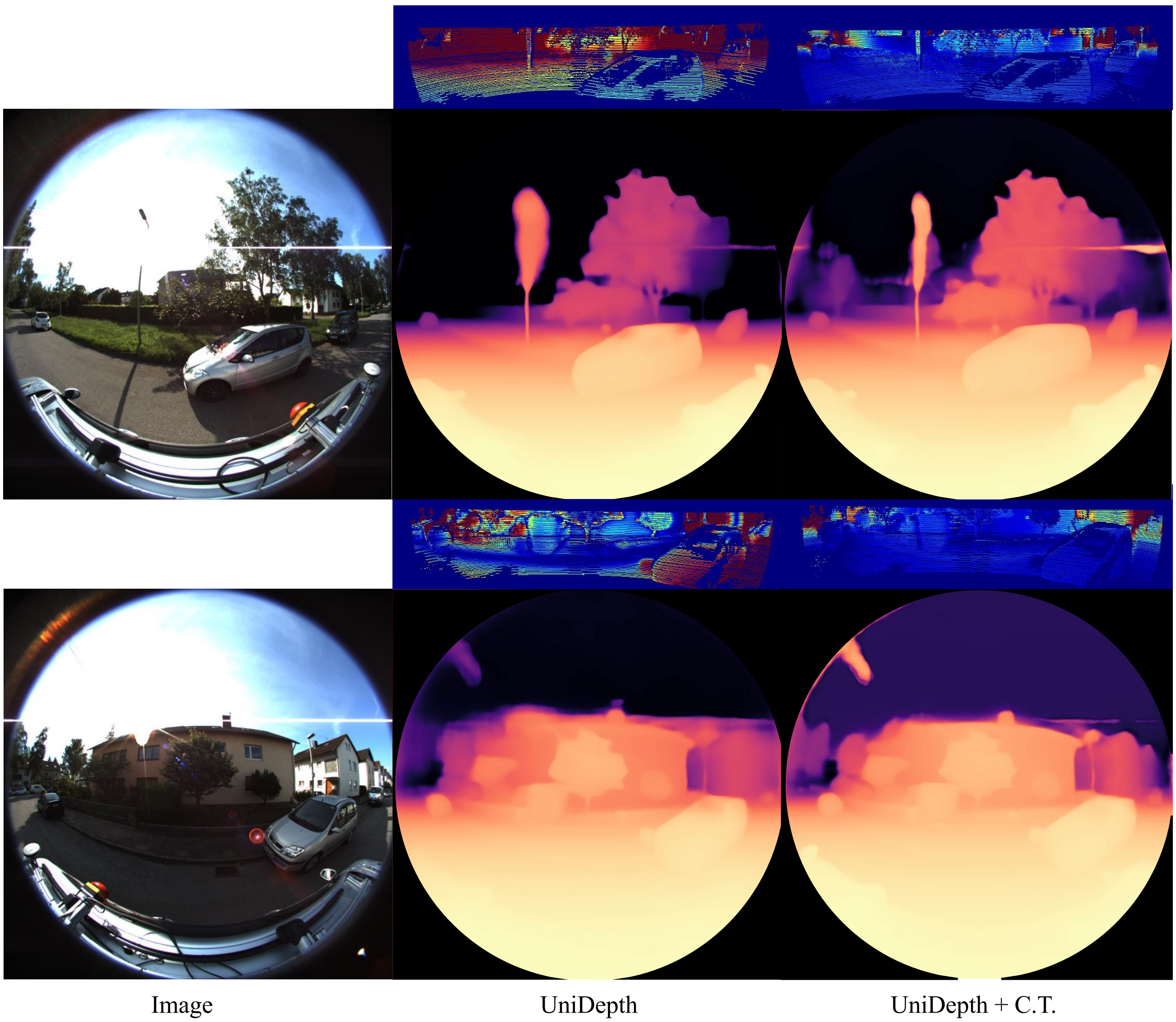

Figure 5. **Additional comparison results on KITTI-360 dataset.**

References

[1] Yunhao Ba, Howard Zhang, Ethan Yang, Akira Suzuki, Arnold Pfahnl, Chethan Chinder Chandrappa, Celso M de Melo, Suya You, Stefano Soatto, Alex Wong, and Achuta Kadambi. Not just streaks: Towards ground truth for single image deraining. In *European Conference on Computer Vision*, pages 723–740. Springer, 2022. 3

[2] Zachary Berger, Parth Agrawal, Tian Yu Liu, Stefano Soatto, and Alex Wong. Stereoscopic universal perturbations across different architectures and datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2022. 3

[3] Marvin Chancán, Alex Wong, and Ian Abraham. 3d reprojection-driven robot navigation improves depth sensing. In *2025 International Conference on Advanced Robotics and Mechatronics (ICARM)*. IEEE, 2025. 3

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin P. Murphy, and Alan Loddon Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40:834–848, 2016. 3

[5] Xien Chen, Suchisrit Gangopadhyay, Michael Chu, Patrick Rim, Hyoungseob Park, and Alex Wong. Uncle: Benchmarking unsupervised continual learning for depth completion. *arXiv preprint arXiv:2410.18074*, 2024. 3

[6] Younjoon Chung, Hyoungseob Park, Patrick Rim, Xiaoran Zhang, Jihe He, Ziyao Zeng, Safa Cicek, Byung-Woo Hong, James S. Duncan, and Alex Wong. Eta: Energy-based test-time adaptation for depth completion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. 3

[7] 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. 2

[8] Vadim Ezhov, Hyoungseob Park, Zhaoyang Zhang, Rishi Upadhyay, Howard Zhang, Chethan Chinder Chandrappa, Achuta Kadambi, Yunhao Ba, Julie Dorsey, and Alex Wong. All-day depth completion. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024. 3

[9] Xiaohan Fei, Alex Wong, and Stefano Soatto. Geo-supervised visual depth prediction. *IEEE Robotics and Automation Letters*, 4(2):1661–1668, 2019. 3

[10] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 270–279, 2017.

[11] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3828–3838, 2019. 3

[12] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2495–2504, 2020. 3

[13] Mingqi Han, Eric A Bushong, Mayuko Segawa, Alexandre Tiard, Alex Wong, Morgan R Brady, Milica Momcilovic, Dane M Wolf, Ralph Zhang, Anton Petcherski, et al. Spatial mapping of mitochondrial networks and bioenergetics in lung cancer. *Nature*, 615(7953):712–719, 2023. 3

[14] Christopher G. Harris and M. J. Stephens. A combined corner and edge detector. In *Alvey Vision Conference*, 1988. 2

[15] Mu Hu, Shuling Wang, Bin Li, Shiyu Ning, Li Fan, and Xiaojin Gong. Penet: Towards precise and efficient image guided depth completion. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13656–13662. IEEE, 2021. 3

[16] Juho Kannala and Sami Brandt. A generic camera calibration method for fish-eye lenses. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, pages 10–13. IEEE, 2004. 2

[17] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3992–4003, 2023. 3

[18] Dong Lao and Ganesh Sundaramoorthi. Minimum delay moving object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4250–4259, 2017. 3

[19] Dong Lao and Ganesh Sundaramoorthi. Extending layered models to 3d motion. In *Proceedings of the European conference on computer vision (ECCV)*, pages 435–451, 2018.

[20] Dong Lao and Ganesh Sundaramoorthi. Minimum delay object detection from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5097–5106, 2019.

[21] Dong Lao, Congli Wang, Alex Wong, and Stefano Soatto. Diffeomorphic template registration for atmospheric turbulence mitigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25107–25116, 2024. 3

[22] Dong Lao, Yangchao Wu, Tian Yu Liu, Alex Wong, and Stefano Soatto. Sub-token vit embedding via stochastic resonance transformers. In *International Conference on Machine Learning*. PMLR, 2024. 3

[23] Dong Lao, Fengyu Yang, Daniel Wang, Hyoungseob Park, Samuel Lu, Alex Wong, and Stefano Soatto. On the viability of monocular depth pre-training for semantic segmentation. In *European Conference on Computer Vision*. Springer, 2024. 3

[24] Yiyi Liao, Jun Xie, and Andreas Geiger. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3292–3310, 2022. 2

[25] Yuankai Lin, Tao Cheng, Qi Zhong, Wending Zhou, and Hua Yang. Dynamic spatial propagation network for depth completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1638–1646, 2022. 3

[26] Tian Yu Liu, Parth Agrawal, Allison Chen, Byung-Woo Hong, and Alex Wong. Monitored distillation for positive congruent depth completion. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II*, pages 35–53. Springer, 2022. 3

[27] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, pages 109–165. Elsevier, 1989. 3

[28] Hyoungseob Park, Anjali Gupta, and Alex Wong. Test-time adaptation for depth completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20519–20529, 2024. 3

[29] Jinsun Park, Kyungdon Joo, Zhe Hu, Chi-Kuei Liu, and In So Kweon. Non-local spatial propagation network for depth completion. In *ECCV*, 2020. 3

[30] Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. Unidepthv2: Universal monocular metric depth estimation made simpler. *arXiv preprint arXiv:2502.20110*, 2025. 2

[31] Rene Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 44(03):1623–1637, 2022. 2

[32] Roger Ratcliff. Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological review*, 97(2):285, 1990. 3

[33] Patrick Rim, Kun He, Kevin Harris, Braden Copple, Shangchen Han, Sizhe An, Ivan Shugurov, Tomas Hodan, He Wen, and Xu Xie. Ego-exo 3d hand tracking in the wild with a mobile multi-camera rig. *arXiv preprint arXiv:2510.02601*, 2025. 3

[34] Patrick Rim, Hyoungseob Park, Vadim Ezhov, Jeffrey Moon, and Alex Wong. Radar-guided polynomial fitting for metric depth estimation. *arXiv preprint arXiv:2503.17182*, 2025. 3

[35] Patrick Rim, Hyoungseob Park, Ziyao Zeng, Younjoon Chung, and Alex Wong. Protodepth: Unsupervised continual depth completion with prototypes. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6304–6316, 2025. 3

[36] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10912–10922, 2021. 2

[37] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision*, 2012. 2

[38] Akash Deep Singh, Yunhao Ba, Ankur Sarker, Howard Zhang, Achuta Kadambi, Stefano Soatto, Mani Srivastava, and Alex Wong. Depth estimation from camera image and mmwave radar point cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9275–9285, 2023. 3

[39] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943, 2018. 3

[40] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020. 2

[41] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020. 3

[42] Sebastian Thrun. Is learning the n-th thing any easier than learning the first? *Advances in neural information processing systems*, 8, 1995. 3

[43] Rishi Upadhyay, Howard Zhang, Yunhao Ba, Ethan Yang, Blake Gella, Sicheng Jiang, Alex Wong, and Achuta Kadambi. Enhancing diffusion models with 3d perspective geometry constraints. *ACM Transactions on Graphics (TOG)*, 42(6): 1–15, 2023. 3

[44] Daniel Wang, Patrick Rim, Tian Tian, Alex Wong, and Ganesh Sundaramoorthi. Ode-gs: Latent odes for dynamic scene extrapolation with 3d gaussian splatting. *arXiv preprint arXiv:2506.05480*, 2025. 3

[45] Fangjinhua Wang, Silvano Galliani, Christoph Vogel, Pablo Speciale, and Marc Pollefeys. Patchmatchnet: Learned multi-view patchmatch stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14194–14203, 2021. 3

[46] Qiang Wang, Shizhen Zheng, Qingsong Yan, Fei Deng, Kaiyong Zhao, and Xiaowen Chu. Irs: A large naturalistic indoor robotics stereo dataset to train deep models for disparity and surface normal estimation. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021. 2

[47] Alex Wong and Stefano Soatto. Bilateral cyclic constraint and adaptive regularization for unsupervised monocular depth prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5644–5653, 2019. 3

[48] Alex Wong and Stefano Soatto. Unsupervised depth completion with calibrated backprojection layers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12747–12756, 2021. 3

[49] Alex Wong and Alan L Yuille. One shot learning via compositions of meaningful patches. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1197–1205, 2015. 3

[50] Alex Wong, Brian Taylor, and Alan L. Yuille. Exploiting protrusion cues for fast and effective shape modeling via ellipses. In *BMVC*, 2017. 3

[51] Alex Wong, Safa Cicek, and Stefano Soatto. Targeted adversarial perturbations for monocular depth prediction. *Advances in neural information processing systems*, 33:8486–8497, 2020. 3

[52] Alex Wong, Xiaohan Fei, Stephanie Tsuei, and Stefano Soatto. Unsupervised depth completion from visual inertial odometry. *IEEE Robotics and Automation Letters*, 5(2):1899–1906, 2020. 2, 3

[53] Alex Wong, Allison Chen, Yangchao Wu, Safa Cicek, Alexandre Tiard, Byung-Woo Hong, and Stefano Soatto. Small lesion segmentation in brain mris with subpixel embedding. In *International MICCAI Brainlesion Workshop*, pages 75–87. Springer, 2021. 3

[54] Alex Wong, Safa Cicek, and Stefano Soatto. Learning topology from synthetic data for unsupervised depth completion. *IEEE Robotics and Automation Letters*, 6(2):1495–1502, 2021. 3

[55] Alex Wong, Xiaohan Fei, Byung-Woo Hong, and Stefano Soatto. An adaptive framework for learning unsupervised depth completion. *IEEE Robotics and Automation Letters*, 6 (2):3120–3127, 2021. 3

[56] Alex Wong, Mukund Mundhra, and Stefano Soatto. Stereopagnosia: Fooling stereo networks with adversarial perturbations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2879–2888, 2021. 3

[57] Chao Xia, Chenfeng Xu, Patrick Rim, Mingyu Ding, Nanning Zheng, Kurt Keutzer, Masayoshi Tomizuka, and Wei Zhan. Quadric representations for lidar odometry, mapping and localization. *IEEE Robotics and Automation Letters*, 8 (8):5023–5030, 2023. 3

[58] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems*, 2021. 3

[59] Yichen Xie, Chenfeng Xu, Marie-Julie Rakotosaona, Patrick Rim, Federico Tombari, Kurt Keutzer, Masayoshi Tomizuka, and Wei Zhan. Sparsefusion: Fusing multi-modal sparse representations for multi-sensor 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17591–17602, 2023. 3

[60] Haofei Xu and Juyong Zhang. Aanet: Adaptive aggregation network for efficient stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1959–1968, 2020. 3

[61] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[62] Yanchao Yang, Alex Wong, and Stefano Soatto. Dense depth posterior (ddp) from single image and sparse range. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3353–3362, 2019. 3

[63] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12–22, 2023. 2

[64] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022. 3

[65] Ziyao Zeng, Jingcheng Ni, Daniel Wang, Patrick Rim, Younjoon Chung, Fengyu Yang, Byung-Woo Hong, and Alex Wong. Priordiffusion: Leverage language prior in diffusion models for monocular depth estimation. *arXiv preprint arXiv:2411.16750*, 2024. 3

[66] Howard Zhang, Yunhao Ba, Ethan Yang, Varan Mehra, Blake Gella, Akira Suzuki, Arnold Pfahnl, Chethan Chinder Chandrappa, Alex Wong, and Achuta Kadambi. Weatherstream: Light transport automation of single image deweathering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13499–13509, 2023. 3

[67] Xiaoran Zhang, Daniel H Pak, Shawn S Ahn, Xiaoxiao Li, Chenyu You, Lawrence H Staib, Albert J Sinusas, Alex Wong, and James S Duncan. Heteroscedastic uncertainty estimation framework for unsupervised registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 651–661. Springer, 2024. 3

[68] Xiaoran Zhang, John C Stendahl, Lawrence H Staib, Albert J Sinusas, Alex Wong, and James S Duncan. Adaptive correspondence scoring for unsupervised medical image registration. In *European Conference on Computer Vision*, pages 76–92. Springer, 2024. 3

[69] Xiaoran Zhang, Byung-Woo Hong, Hyoungseob Park, Daniel H. Pak, Anne-Marie Rickmann, Lawrence H. Staib, James S. Duncan, and Alex Wong. Progressive test time energy adaptation for medical image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. 3

[70] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1851–1858, 2017. 3