SUPPLEMENTARY MATERIALS for Energy-based Test-time Adaptation

Younjoon Chung* Hyoungseob Park* Patrick Rim* Xiaoran Zhang Jihe He
Ziyao Zeng Safa Cicek† Byung-Woo Hong‡ James S. Duncan Alex Wong

Yale University †UCLA ‡Chung-Ang University

{younjoon.chung, hyoungseob.park, patrick.rim, xiaoran.zhang}@yale.edu
safacicek@ucla.edu, hong@cau.ac.kr, {james.duncan, alex.wong}@yale.edu

A. Datasets

KITTI dataset [17] provides calibrated RGB images synchronized with Velodyne lidar point clouds, GPS, and inertial data, collected from over 61 driving scenes. It includes ≈80K raw image frames paired with sparse depth maps of ≈5% density, commonly used for depth completion [69]. Semi-dense depth data is available for the bottom 30% of the image space, while ground-truth depth maps combine 11 consecutive raw lidar scans. We trained our model on ≈86K single images, without using the test or validation sets.

VOID dataset [77] consists of 640×480 RGB images synchronized with sparse depth maps captured in indoor settings like classrooms and laboratories, and outdoor gardens. Sparse depth maps (≈0.5% density, ≈1,500 points) were created with the XIVO VIO system [14], while dense ground-truth maps were obtained using active stereo. VOID introduces challenging 6 DoF motion due to rolling shutter effects in 56 sequences, contrasting with KITTI's planar motion. Our model was trained on ≈46K images.

NYUv2 dataset [45] contains 372K synchronized 640×480 RGB images and depth maps captured using Microsoft Kinect across 464 indoor scenes, including homes, offices, and stores. To simulate SLAM/VIO-style sparse depth maps, we employed the Harris corner detector [22] to extract ≈1,500 points from the depth maps. We evaluated adaptation performance on 654 test images.

ScanNet dataset [10] offers 2.5 million images with dense depth maps across 1,513 indoor scenes. SLAM/VIO-style sparse depth maps were simulated by applying the Harris corner detector [22], sampling ≈1,500 points from the dense maps. Our experiments utilized ≈21K test images for adaptation.

Virtual KITTI (VKITTI) dataset [15] includes ≈17K 1242×375 synthetic images across 35 videos, derived from 5 original KITTI videos augmented with 7 variations in lighting, weather, and camera perspectives [69]. To minimize the large domain gap between RGB images from VKITTI and KITTI despite Unity's virtual similarity to KITTI scenes [15], we used VKITTI's dense depth maps only to reduce the domain gap in photometric variations, while sparse depth maps were simulated to match KITTI's lidar-generated distribution in terms of marginal distribution of sparse points. A test set of ≈2,300 images was used for adaptation.

nuScenes dataset [5] provides 1600×900 RGB images synchronized with sparse point clouds, featuring 27.4K training images from 1,000 driving scenes and 5.8K test images from 150 scenes. For the test set, ground truth was created by merging projected sparse depth from forward-backward frames. Setup details will be provided with released code for reproducibility.

SceneNet dataset [37] comprises 5 million 320×240 RGB images with depth maps captured in simulated indoor environments with randomized room arrangements. Due to the lack of sparse depths, sparse depth maps were derived using the Harris corner detector [22] simulating SLAM/VIO outputs, followed by k-means clustering to reduce the sampled points to 375 (0.49% total pixel density). We used ≈2,300 test images for adaptation from a single split (out of 17 available) of 1,000 sequences of 300 images each. Each sequence is generated by recording the same scene over a trajectory.

Waymo Open Dataset [65] includes 1920×1280 RGB images and lidar scans collected at 10Hz in autonomous vehicle scenes. It features ≈158K training images from 798 scenes, and ≈40K validation images from 202 scenes with sampling frquency of 0.6 seconds. Objects are annotated across full 360° field. Each top lidar sensor's point cloud is projected onto camera frame. Ground truth was generated by merging top and front lidar scans projected over 10 forward-backward frames, corresponding to 1-second intervals, with moving objects removed using annotations. Outliers in depth points were filtered out for accuracy.

*Equal contribution

B. Implementation and training details

Model Architecture. Energy model is implemented as a convolutional neural network that takes a two-channel input of sparse depth and the dense prediction. It uses six 5x5 convolutional layers (stride 2) with LeakyReLU activations to increase channel depth from 2 to 512. A final 3x3 convolutional layer then maps these features to a single-channel energy map to score input regions.

Hyperparameters. Model and dataset specific hyperparameters for test-time adaptation are noted in 2.

Training energy models. We take baseline depth completion models pre-trained on KITTI and VOID from [?]. For each model, we train patch-based energy model on the corresponding source dataset *i.e*. KITTI, VOID. All models were trained for 5 epochs with a batch size of 32. Specific learning rates and hyperparameters for data augmentation will be released with the code.

Evaluation. For outdoor datasets, test-time adaptation performances are evaluated on bottom-cropped regions to exclude regions where no corresponding sparse depth exists. For VKITTI, we evaluate on 1240×240 bottom-cropped regions, 1600×544 for nuScenes, and 1920×640 for Waymo. For indoor datasets, models are evaluated on the entire region. The error metrics used for evaluation are defined in 1. For outdoor, we evaluate the models on depth range from 0.0 to 80.0 meters. For indoor, we evaluate on 0.2 to 5.0 meters.

C. Extended Related Work

As we utilize adversarial perturbations in our method, we present a related works on the topic as an extended discussion.

Adversarial Perturbations. Small input perturbations can significantly alter classification outputs [66]. Goodfellow et al. [19] introduced Fast Gradient Sign Method (FGSM), later extended to iterative variants for increased effectiveness [12, 27, 36]. Minimal perturbations were studied in [38], and lower bounds on their magnitudes were analyzed in [49]. Adversarial examples can yield high-confidence outputs from unrecognizable inputs [46], and are attributed to non-robust features [26]. Transferability across models and datasets was explored in [44, 86].

Universal perturbation, which can be applied even without knowledge of the trained model, and generalize across domains [6], was proposed in [39]. Data-independent and data-free constructions have been studied in [40, 41], and generative methods has been explored in [23, 42, 52]. [55] extends the concept to non-Euclidean domains.

Adversarial defense includes adversarial training [27, 68], universal training [43, 62], gradient discretization [4, 84], input randomization [47, 54, 83], purification [1, 21, 53, 60], and denoising [33]. Other strategies include normalization [87] and object detection [8].

Adversarial robustness has also been studied in dense prediction tasks. Prior works addressed detection and segmentation [24, 85], monocular depth [41, 76], and optical flow [57, 61]. Recent studies examined physical patch attacks [95] and synthetic augmentations [11]. Stereo attacks were considered in [?], and [3] studies universal perturbations for stereo depth estimation. We exploit the adversarial perturbations as a mean of exploring the data space, where the perturbed samples simulates the out-of-distribution samples with source data. The out-of-distribution samples enable the energy model to learn to assign high energy to the predictions on target distribution.

Metric	Definition
MAE	$\frac{1}{\lvert\Omega\rvert}\sum_{x\in\Omega} \lvert\hat{d}(x) - d_{gt}(x)\rvert$
RMSE	$\left(\frac{1}{\lvert\Omega\rvert}\sum_{x\in\Omega} \lvert\hat{d}(x) - d_{gt}(x)\rvert^2\right)^{1/2}$

Table 1. *Error metrics.* d_{gt} means the ground-truth depth.

Dataset	LR	w_{sm}	w_z	w_{energy}	Inner Iter.
MSG-CHN					
Waymo	3e-3	3.0	1.0	0.001	3
VKITTI-FOG	5e-4	6.0	1.0	0.5	5
nuScenes	3e-3	5.0	1.0	0.5	3
SceneNet	1e-3	8.0	1.0	0.1	3
NYUv2	5e-4	7.5	1.0	0.004	3
ScanNet	5e-3	8.0	1.0	0.001	3
NLSPN					
Waymo	6e-3	1.0	1.0	0.001	1
VKITTI-FOG	1e-3	1.0	1.0	0.001	1
nuScenes	6e-3	1.0	1.0	0.002	1
SceneNet	3e-3	1.5	1.0	2.0	3
NYUv2	4e-3	5.0	1.0	1.0	3
ScanNet	1e-4	2.0	1.0	0.3	3
CostDCNet					
Waymo	5e-3	3.0	1.0	0.1	1
VKITTI-FOG	5e-3	3.0	1.0	0.04	1
nuScenes	5e-3	3.0	1.0	0.003	1
SceneNet	6e-3	2.5	1.0	0.001	3
NYUv2	3e-3	3.5	1.0	0.0001	3
ScanNet	2e-3	2.0	1.0	0.0002	3

Table 2. *Hyperparameters.* Model specific hyperparameters used at test-time.

E. Discussion

In the pursuit of building embodied AI agents, we must equip them with the capability of efficient and robust egocentric 3D reconstruction [31, 70, 71, 82, 88] that can generalize to different domains via adaptation. We view energy-based methods, such as ours, as a tool with unlocked potential to push the frontiers of many critical sub-tasks under this broader vision of domain adaptation, including

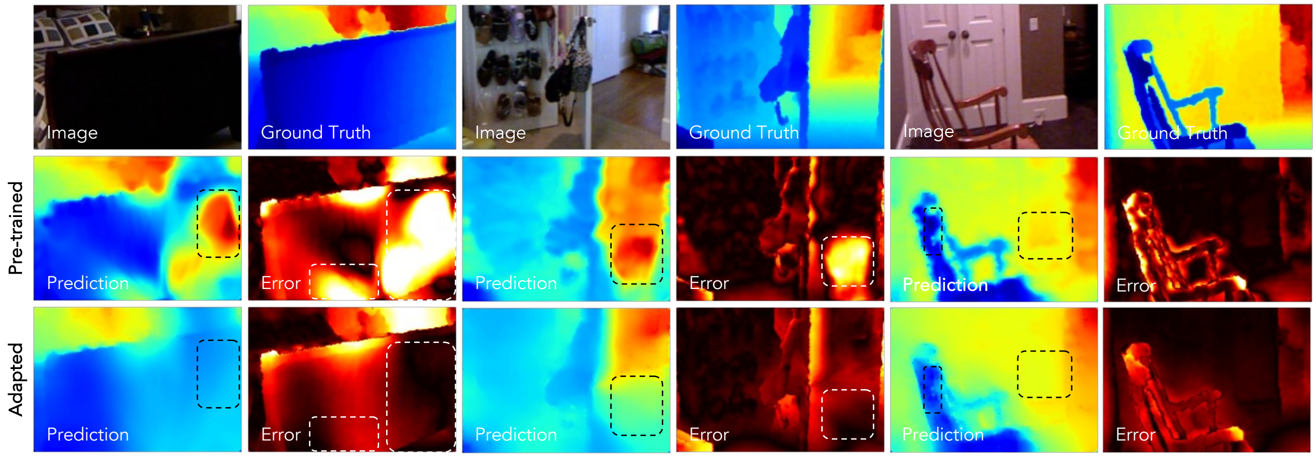

Figure 1. *Qualitative results on NYUv2.* We adapt CostDCNet from VOID→NYUv2.

monocular depth estimation (MDE) [32, 74, 76, 81]. To address the inherent scale ambiguity in the task of estimating 3D depth from a single 2D image, one can explore the multimodal tasks of predicting depth from image, in addition to one or multiple of: radar [58, 63], lidar [9, 13, 35, 59, 75, 77–79], language [93], inertial sensors [14], additional cameras [3, 16, 80], and other modalities that encode semantic and/or geometric information about a three-dimensional scene.

F. Limitations

While this paper proposes an energy-based test-time adaptation method for depth completion and demonstrates an energy model trained on both in-distribution and adversarially perturbed out-of-distribution samples, there are limitations in scope and generality. Our focus is restricted to depth completion [25, 34, 48, 75, 77–79, 90, 94]; however, the energy model, the core component of our approach, can be applied to other geometric tasks such as optical flow [2, 28–30, 64, 67, 89], monocular depth prediction [14, 18, 50, 51, 56, 73, 74], and multi-view stereo [7, 20, 72, 91, 92], where adaptation mechanisms using energy models remain underexplored. We hope our findings contribute to the adaptation of geometric models in real-time, resource-constrained settings to unforeseen environmental conditions.

References

[1] Naveed Akhtar, Jian Liu, and Ajmal Mian. Defense against universal adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3389–3398, 2018. 2

[2] Filippo Aleotti, Matteo Poggi, Fabio Tosi, and Stefano Mattoccia. Learning end-to-end scene flow by distilling single tasks knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10435–10442, 2020. 3

[3] 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. 2, 3

[4] Jacob Buckman, Aurko Roy, Colin Raffel, and Ian Goodfellow. Thermometer encoding: One hot way to resist adversarial examples. In *International Conference on Learning Representations*, 2018. 2

[5] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 1

[6] Ashutosh Chaubey, Nikhil Agrawal, Kavya Barnwal, Keerat K Guliani, and Pramod Mehta. Universal adversarial perturbations: A survey. *arXiv preprint arXiv:2005.08087*, 2020. 2

[7] Rui Chen, Songfang Han, Jing Xu, and Hao Su. Point-based multi-view stereo network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1538–1547, 2019. 3

[8] Xiangning Chen, Cihang Xie, Mingxing Tan, Li Zhang, Cho-Jui Hsieh, and Boqing Gong. Robust and accurate object detection via adversarial learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16622–16631, 2021. 2

[9] 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

[10] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 1

[11] Tom van Dijk and Guido de Croon. How do neural networks see depth in single images? In *Proceedings of the IEEE*

International Conference on Computer Vision, pages 2183–2191, 2019. 2

[12] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193, 2018. 2

[13] 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

[14] Xiaohan Fei, Alex Wong, and Stefano Soatto. Geo-supervised visual depth prediction. *IEEE Robotics and Automation Letters*, 4(2):1661–1668, 2019. 1, 3

[15] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4340–4349, 2016. 1

[16] Suchisrit Gangopadhyay, Jung-Hee Kim, Xien Chen, Patrick Rim, Hyoungseob Park, and Alex Wong. Extending foundational monocular depth estimators to fisheye cameras with calibration tokens. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. 3

[17] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32:1231 – 1237, 2013. 1

[18] 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

[19] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 2

[20] 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

[21] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens Van Der Maaten. Countering adversarial images using input transformations. *arXiv preprint arXiv:1711.00117*, 2017. 2

[22] Christopher G. Harris and M. J. Stephens. A combined corner and edge detector. In *Alvey Vision Conference*, 1988. 1

[23] Jamie Hayes and George Danezis. Learning universal adversarial perturbations with generative models. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 43–49. IEEE, 2018. 2

[24] Jan Hendrik Metzen, Mummadi Chaithanya Kumar, Thomas Brox, and Volker Fischer. Universal adversarial perturbations against semantic image segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2755–2764, 2017. 2

[25] 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

[26] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*, pages 125–136, 2019. 2

[27] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016. 2

[28] 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

[29] 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.

[30] 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. 3

[31] 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. 2

[32] 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

[33] Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. Defense against adversarial attacks using high-level representation guided denoiser. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1778–1787, 2018. 2

[34] 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

[35] 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

[36] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 2

[37] John McCormac, Ankur Handa, Stefan Leutenegger, and Andrew J Davison. Scenenet rgb-d: 5m photorealistic images of synthetic indoor trajectories with ground truth. *arXiv preprint arXiv:1612.05079*, 2016. 1

[38] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE con-*

ference on computer vision and pattern recognition, pages 2574–2582, 2016. 2

[39] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773, 2017. 2

[40] KR Mopuri, U Garg, and R Venkatesh Babu. Fast feature fool: A data independent approach to universal adversarial perturbations. In *British Machine Vision Conference 2017, BMVC 2017*. BMVA Press, 2017. 2

[41] Konda Reddy Mopuri, Aditya Ganeshan, and R Venkatesh Babu. Generalizable data-free objective for crafting universal adversarial perturbations. *IEEE transactions on pattern analysis and machine intelligence*, 41(10):2452–2465, 2018. 2

[42] Konda Reddy Mopuri, Utkarsh Ojha, Utsav Garg, and R Venkatesh Babu. Nag: Network for adversary generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 742–751, 2018. 2

[43] Chaithanya Kumar Mummadi, Thomas Brox, and Jan Hendrik Metzen. Defending against universal perturbations with shared adversarial training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4928–4937, 2019. 2

[44] Muhammad Muzammal Naseer, Salman H Khan, Muhammad Haris Khan, Fahad Shahbaz Khan, and Fatih Porikli. Cross-domain transferability of adversarial perturbations. In *Advances in Neural Information Processing Systems*, pages 12905–12915, 2019. 2

[45] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 1

[46] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436, 2015. 2

[47] Tianyu Pang, Kun Xu, and Jun Zhu. Mixup inference: Better exploiting mixup to defend adversarial attacks. *arXiv preprint arXiv:1909.11515*, 2019. 2

[48] Jinsun Park, Kyungdon Joo, Zhe Hu, Chi-Kuei Liu, and In So Kweon. Non-local spatial propagation network for depth completion. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 120–136. Springer, 2020. 3

[49] Jonathan Peck, Joris Roels, Bart Goossens, and Yvan Saeys. Lower bounds on the robustness to adversarial perturbations. In *Advances in Neural Information Processing Systems*, pages 804–813, 2017. 2

[50] Matteo Poggi, Filippo Aleotti, Fabio Tosi, and Stefano Mattoccia. On the uncertainty of self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3227–3237, 2020. 3

[51] Matteo Poggi, Fabio Tosi, Filippo Aleotti, and Stefano Mattoccia. Real-time self-supervised monocular depth estimation without gpu. *IEEE Transactions on Intelligent Transportation Systems*, 2022. 3

[52] Omid Poursaeed, Isay Katsman, Bicheng Gao, and Serge Belongie. Generative adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4422–4431, 2018. 2

[53] Aaditya Prakash, Nick Moran, Solomon Garber, Antonella DiLillo, and James Storer. Deflecting adversarial attacks with pixel deflection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8571–8580, 2018. 2

[54] Edward Raff, Jared Sylvester, Steven Forsyth, and Mark McLean. Barrage of random transforms for adversarially robust defense. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6528–6537, 2019. 2

[55] Arianna Rampini, Franco Pestarini, Luca Cosmo, Simone Melzi, and Emanuele Rodola. Universal spectral adversarial attacks for deformable shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3216–3226, 2021. 2

[56] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12179–12188, 2021. 3

[57] Anurag Ranjan, Joel Janai, Andreas Geiger, and Michael J Black. Attacking optical flow. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2404–2413, 2019. 2

[58] 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

[59] 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

[60] Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-gan: Protecting classifiers against adversarial attacks using generative models. *arXiv preprint arXiv:1805.06605*, 2018. 2

[61] Simon Schrodi, Tonmoy Saikia, and Thomas Brox. What causes optical flow networks to be vulnerable to physical adversarial attacks. *arXiv preprint arXiv:2103.16255*, 2021. 2

[62] Ali Shafahi, Mahyar Najibi, Zheng Xu, John Dickerson, Larry S Davis, and Tom Goldstein. Universal adversarial training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5636–5643, 2020. 2

[63] 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

[64] 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

[65] 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. 1

[66] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 2

[67] 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

[68] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017. 2

[69] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *2017 international conference on 3D Vision (3DV)*, pages 11–20. IEEE, 2017. 1

[70] 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. 2

[71] 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. 2

[72] 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

[73] 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. 3

[74] 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

[75] 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

[76] Alex Wong, Safa Cicek, and Stefano Soatto. Targeted adversarial perturbations for monocular depth prediction. *Advances in neural information processing systems*, 33:8486–8497, 2020. 2, 3

[77] 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. 1, 3

[78] 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.

[79] 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

[80] 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

[81] Yangchao Wu, Tian Yu Liu, Hyoungseob Park, Stefano Soatto, Dong Lao, and Alex Wong. Augundo: Scaling up augmentations for monocular depth completion and estimation. In *European Conference on Computer Vision*, pages 274–293. Springer, 2024. 3

[82] 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. 2

[83] Chang Xiao and Changxi Zheng. One man's trash is another man's treasure: Resisting adversarial examples by adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 412–421, 2020. 2

[84] Chang Xiao, Peilin Zhong, and Changxi Zheng. Enhancing adversarial defense by k-winners-take-all. *arXiv preprint arXiv:1905.10510*, 2019. 2

[85] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. Adversarial examples for semantic segmentation and object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1369–1378, 2017. 2

[86] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2730–2739, 2019. 2

[87] Cihang Xie, Mingxing Tan, Boqing Gong, Jiang Wang, Alan L Yuille, and Quoc V Le. Adversarial examples improve image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 819–828, 2020. 2

[88] 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. 2

[89] Yanchao Yang and Stefano Soatto. Conditional prior networks for optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 271–287, 2018. 3

[90] Yanchao Yang, Alex Wong, and Stefano Soatto. Dense depth posterior (ddp) from single image and sparse range. In *Pro-*

ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3353–3362, 2019. 3

[91] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 767–783, 2018. 3

[92] Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent mvsnet for high-resolution multi-view stereo depth inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5525–5534, 2019. 3

[93] 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

[94] Yufan Zhu, Weisheng Dong, Leida Li, Jinjian Wu, Xin Li, and Guangming Shi. Robust depth completion with uncertainty-driven loss functions. *arXiv preprint arXiv:2112.07895*, 2021. 3

[95] Alon Zolfi, Moshe Kravchik, Yuval Elovici, and Asaf Shabtai. The translucent patch: A physical and universal attack on object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15232–15241, 2021. 2