Generative Data Augmentation for Dual-Plane Monocular Depth Estimation of Non-Lambertian Surfaces

Vilius Bankauskas [1] Andrius Kriščiūnas [2]

[1] *Kaunas University of Technology, K. Donelaičio str. 73, Kaunas, LT-44249, Lithuania*

Abstract

Monocular depth estimation struggles significantly with non-Lambertian surfaces such as glass and mirrors, leading to critical failures in depth maps. Traditional data acquisition methods, including LiDAR, fail to capture accurate ground truth for these objects due to physical light-bending properties. To address this lack of data, a generative data augmentation pipeline is proposed, utilizing advanced diffusion models to synthesize a dual-plane dataset from existing high-quality RGB-D scans. By either inpainting new transparent objects (forcing the model to estimate the background depth) or performing a material transfer on existing objects (providing ground truth for the transparent surface depth), two distinct training datasets are created. The proposed methodology uses the latest generative architectures for their superior curvature preservation and inpainting accuracy. Evaluated using the high-resolution DIODE indoors dataset, this generative augmentation offers an optimized, scalable solution to overcome the lack of data caused by sensor limitations, enabling training of MDE models capable of accurately perceiving complex scenes with transparent elements.

Keywords

Monocular Depth Estimation, Generative AI, Data Augmentation, Non-Lambertian Surfaces, Synthetic Data, Diffusion Models, Depth Mapping

1. Introduction

Monocular depth estimation (MDE) has been an ever-developing field in computer science, especially encouraged by steady advancements in artificial intelligence. MDE is a computer vision task using neural networks to predict a 3D depth map from a single 2D RGB image. This enables a variety of applications ranging from robotics to augmented reality. There has been a surge of research in the field due to advancements in neural network architecture and availability of high-quality data. High accuracy sensors that are able to acquire the required depth data, such as LiDAR, used to be exclusive to specialized applications due to their price and availability. The reduction in sensor costs and the integration of LiDAR into consumer smartphones have democratized access to depth data, accelerating the advancement of dependent models.

There is a variety of well-performing MDE models that are open sourced and well documented. *Depth Anything*[1], *UniDepth*[2] and *MiDaS*[3] are just a few to name that have pushed the whole field of MDE forward and enabled accessibility of this technology to the masses. Even though these models perform well in naturally occurring scenes there are few flaws within both the logic of the task itself and elements that the task is comprised of. Monocular depth estimation is an inherently ill-posed problem, as a single 2D projection can map to multiple 3D geometries.

Another problem arises from the naturally occurring scenes themselves. Everyday objects such as doors, walls and floors are all well documented and depth data for these objects is easy to get and for a model to learn the visual cues to associate with continuous planar surfaces. Not all objects in our everyday life are so simple as mentioned before. Some doors have glass components, others such as a closet door might have a mirror attached to it. This is an obvious problem due to several reasons. First, these objects are very common such as kitchen glasses, windows, plastic bottles, daylight curtains. The visual representation of these objects does not reflect their actual shape or distance, an

*IVUS 2026: Information Society and University Studies 2026, May 12, Kaunas, Lithuania

vilius.bankauskas@ktu.edu (V. Bankauskas. 1), bankauskasvil@gmail.com (V. Bankauskas. 1); andrius.krisciunas@ktu.lt (A. Krisciunas. 2)

0009-0002-0826-4142 (V. Bankauskas); 0000-0002-3691-0423 (A. Krisciunas)

Copyright © 2026 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-06

CEUR-WS.org/Vol-4317/paper21.pdf

object that is largely transparent and lacking in features, like a fully glass door, is visually almost indistinguishable from regular empty space. For humans seeing the slight color changes and floating door handle is sufficient evidence of a door even though visually it is invisible. Reflective objects present a greater challenge; well-maintained, featureless surfaces often perfectly reflect a scene that is not physically present, heavily distorting sensor readings.

For humans these difficulties are innate, unavoidable and are learned quickly. Similar adaptations could potentially be achieved by neural networks; however, a fundamental sensor limitation prevents this process. Sensors used to gather depth data are equally as blind and unfamiliar to these non-Lambertian bodies and their light bending properties as a human eye. Because time-of-flight sensors merely measure the return time of light, they cannot differentiate between diffuse surfaces and non-Lambertian bodies that bend or reflect the signal. Consequently, scans produce data "black holes" where data is misrepresented by an object warping scene by bending the light or detecting a whole "phantom" scene by scanning a mirror or similar reflective surface [4,5]. Therefore, it is extremely difficult to produce quality image and depth map combinations when they have such problematic objects in it. If there is no data for models to learn from – models do not learn.

This sort of problem is faced frequently in the field of artificial intelligence. Quality of data is a concern for all types of models. Usually, these data quality issues can be mitigated without requiring of expensive resources. There have been attempts at dealing with reflective and transparent surfaces in the past both from synthetic data generation perspective and physical assistance for the sensors [6,7]. Rather than treating light-bending properties as noise, this dual-plane information can be used to determine what is behind said object. This would enable to maximize the usage of available information and detect both the "first plan" that is the objects distance and the "second plan" that would be the surface or objects behind the transparent object. In actuality this would require two different datasets and two different models trained to perform two different tasks. Instead of solving the problem this doubles it. However, with advancements in the field of generative AI, especially top-of-the-line diffusion models, it might be possible to avoid complex collection of data all together. Instead of inventing new ways to collect difficult data it might be wise to change the way the data is used and augment it in such a manner that expands on already existing datasets.

Attempts at dealing with reflective and transparent surfaces have heavily relied on depth completion techniques [8,9] and CGI rendering [10], often referred to as sim-to-real approaches. While rendering engines can mathematically simulate light bending properties, these simulations frequently lack photorealism or make such simulations more computationally and financially expensive than the real acquisition of such data [11]. Models trained on CGI data learn to recognize synthetic artifacts rather than the actual visual cues of glass, mirrors or plastic [12–14]. This leads to a significant domain shift when applied in real-world scenarios. Consequently, these sim-to-real methods often fail to generalize effectively outside of their digital environments.

In recent years there is a clear trend forming of using synthetic data augmentation to increase hard to access or impossible to get data [15–18]. This has its drawbacks, first, the quality of the whole process depends on one element, quality of synthetic data. Quality of the generated data can only be as good as diffusion model or rendering engine allows it to be. For simulated environments most lack photorealism and have synthesis defining artifacts that models learn to recognize and this leads to domain shift. Diffusion models struggle with repeatability and geometric consistency leading to distorted objects which leads to increased error margins.

2. Methodology

As seen in Figure 1 the goal is to take already existing high quality RGB-D dataset and augment it in such a way that it results in two distinct types of data:

1. Seeing through an object (background)
2. Seeing the surface of an object (foreground)

This is achieved using generative artificial intelligence to inpaint new objects that are not originally in the scene as well as to swap the already existing objects in a scene to a new see-through material while maintaining as much of the original geometry as possible. Doing so creates two different datasets, one with inpainted new transparent objects that are not "visible" from the

perspective of a depth map, another with object that is in the original depth map but changed to visually appear transparent. Having these two datasets allows for training of two different models with different methods of dealing with transparent objects.

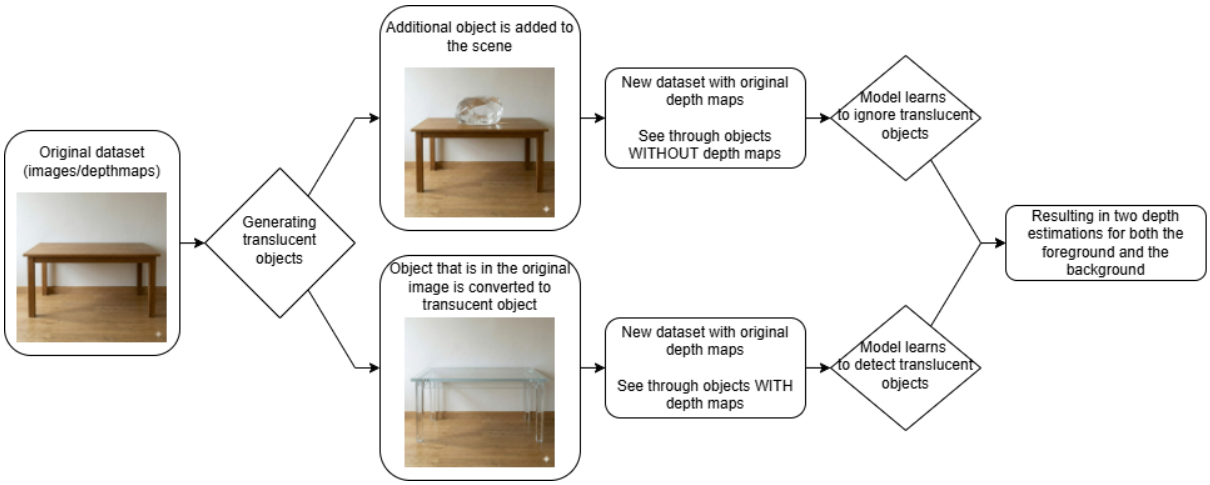


Figure 1: Flowchart of the proposed generative data augmentation pipeline. The original RGB dataset is split into two parallel streams: inpainting transparent objects to force the model to learn background depth and applying material style transfer to existing objects to provide ground truth for transparent surface depth.

The new approach has its own drawbacks. It is fully dependent on the quality of the selected diffusion model to produce photorealistic new objects and even more dependent on the model when it comes to maintaining the shape of an object when transforming its material.

Model selection is critical for optimizing generation quality. However, as mentioned before some of the tasks are impossible by nature, for example testing the geometry of the newly transformed object. It is possible to check for the outline contours to match the original image however when it comes to the actual internal shape and small delicate curves and irregularities it is a far more difficult task. Color palette is another option; however, it must be heavily monitored to allow only a small variation due to transparent objects being see-through, this is also heavily dependent on the background of which front the object is placed.

To ensure the highest quality of augmented data, a comparative analysis was performed between several prominent diffusion models: Stable Diffusion 1.5 [19], Stable Diffusion XL [20], InstructPix2Pix [21], Flux.1[dev] [22], and Kandinsky 5.0 Lite Inpainting [23].

Earlier models frequently failed to produce a convincing glass look, instead generating gray, opaque artifacts. Newer iterations demonstrated an improvement in photorealism but struggled heavily with geometric consistency, often losing the core identity and shape constraints of the original objects.

Ultimately, the Gemini image models (Google Gemini 2.5 Image and Google Gemini 3 Image [24]) were selected for the final pipeline. Justification for this choice relies on three key factors: curvature preservation, inpainting accuracy, and overall efficiency. These models demonstrate superior ability to maintain delicate internal curves and structural integrity without additional finetuning required. This is especially apparent when swapping object material to glass. Furthermore, their high inpainting accuracy ensures that newly generated non-Lambertian objects integrated seamlessly into the existing scene without heavily distorting the surrounding environment or bleeding into the background.

Finally, an important factor that cannot be dismissed is ease of use and the streamlined nature of prompt driven augmentation. The forementioned models use prompts instead of masks to determine what part of an image should be augmented [21]. This presents a complex trade-off that in some cases can become cumbersome and require extensively detailed descriptions of scenes to correctly choose an object or part of the scene to augment. However, in the case of proposed pipeline, this approach enables the prompts to stay intentionally vague and be applicable to all scenes and require no additional cropping or masking.

Synthetic dataset part 1 (background)

When using diffusion models to generate new objects into the scene and keeping the ground truth depth maps unchanged, trained models are forced to deal with this visual inconsistency in the training data. Because the objects are transparent or semi-transparent, features about the background are mostly maintained and can be used to construct an approximation of the scene. Training a model on translucent objects that are not in the original scene forces model to "ignore" any transparent object and instead provide a depth map of scene behind them. This mimics human visual perception where the valuable dual data of transparent objects is not wasted but used [25,26].

Synthetic dataset part 2 (foreground)

Transforming already existing object into a new transparent or semi-transparent material leads to a more standard approach in depth estimation where a translucent object is still detected as an object in the scene and is used and seen in scene's depth map. Using this technique enables synthesis of data that is difficult to gather. This enables models to learn from many more examples of high-quality data with depth maps generated using lidar sensors or other high-fidelity sensors.

3. Experiments and validation

DIODE indoors dataset [27] was selected for this research due to several factors, one of which is relatively high resolution of the scene images. Higher resolution means more data to work with and more realistic scenario, this is important for both the diffusion models and for trained MDE models.

While alternative datasets such as NYU Depth V2 [28] or various outdoor datasets are widely used, the DIODE indoors dataset was prioritized specifically over these alternatives due to its laser-scanned ground truth. Unlike datasets that rely heavily on estimated or lower-resolution depth maps, DIODE provides the geometric constraints required for diffusion models to work with, ensuring the trained MDE models learn from highly realistic, noise-free scenarios.

When models fail, they typically show a "failed glass" look rendering the object flat and lacking in detail, localized blur, or they suffer a complete loss of geometric constraints, warping the object's identity beyond recognition. For the generated data to be viable for MDE training, the inserted or transformed objects must seamlessly fit into the scene and fit the shadows, lighting realistically bend through the object and cast shadows of its own. If generated object fails to replicate how light realistically bends through the object and cast shadows, the result is a visual contradiction. MDE model would then learn incorrect cues, rendering the augmented dataset unlearnable for the training process.

It is important to note that the task of inpainting is complicated by itself. Due to objects being transparent there are a lot of variables to account for. Lighting must match the scene as well as the unusual properties of objects of non-Lambertian nature.

4. Current results

Generated images are currently of varying quality. Images were generated using Google Gemini 2.5 Image and Google Gemini 3 Image [24]. Due to some of the scenes in the DIODE dataset being empty and not featuring any objects, the chosen approach uses images that have a parallel angle to the floor. Some of the chosen images already have non-Lambertian objects present in the scenes. This was deliberately ignored for the sake of acquiring first results and to test the feasibility of implementation.

When analyzing presented figures, the layout of examples follows the same structure. As demonstrated in Figure 2, the pipeline processes three distinct scene scenarios.



Figure 2: Visual comparison of the augmentation pipeline across three distinct scenes from the DIODE indoors dataset. Left: Original non-augmented scene. Middle: Scene with a novel non-Lambertian object (glass vase) inpainted, accurately reflecting scene lighting and casting shadows. Right: Existing scene object transformed into a transparent material. Note the minor loss of geometric detail in the top row, where the transformed electrical cover loses its physical indentations.

Examining Figure 2 three different scene scenarios are demonstrated. The synthetic objects introduced into these scenes consist of three floating vases. These new scenes are not commonly occurring and are quite difficult to recreate, however examining each of them they do seem to be visually fitting in the scenes lighting and all of them generate a shadow of their own as well as keep the scenes information that was originally in the image but is now covered by the non-Lambertian body. When examining the transformed objects, the results are more nuanced. Two of the three transformed scenes suffered some minor loss of geometric identity, and the newly generated objects did not fully represent their real counterparts. In the image 00000_00003_indoors_170_000.png (first row) metal cover of the electrical components gets converted to a glass plane. This new glass plane does not fully fit into the geometric constraints of the original object; it keeps the silhouette and general relative location in the scene however it loses some minor indentations that will cause a discrepancy between ground truth depth maps and the visual data.

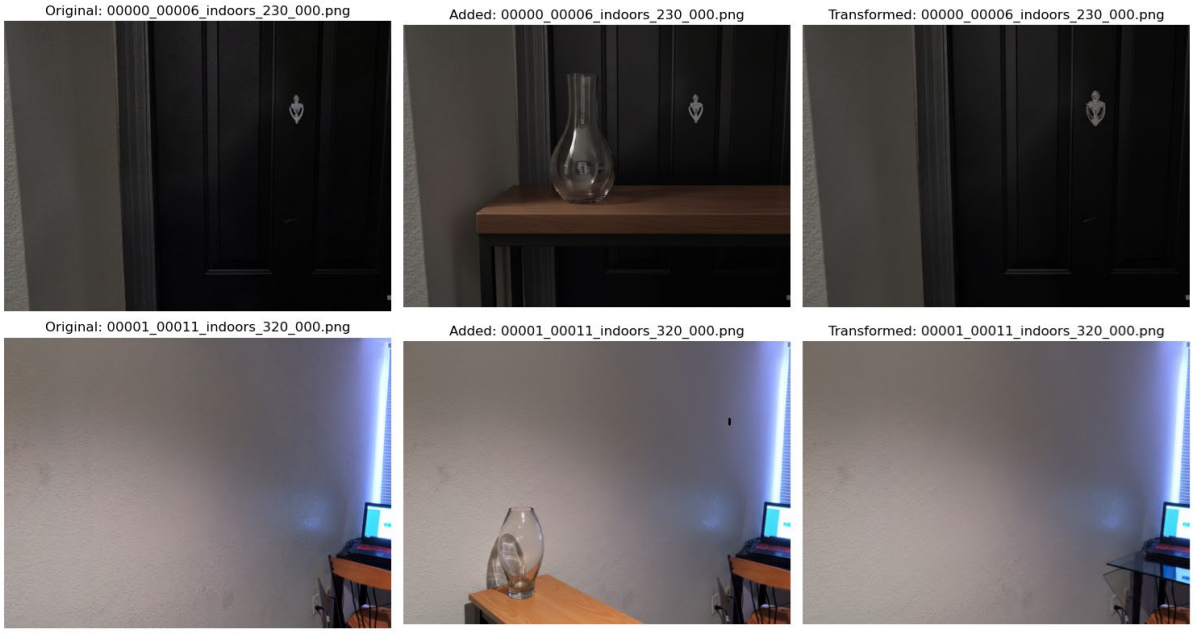


Figure 3: Example of generation artifacts during inpainting. In addition to the requested transparent object, the model occasionally generates unprompted supporting structures (e.g., a wooden table), which alters the original color palette and introduces severe geometric discrepancies with the ground truth depth map.

A more problematic issue is visible in Figure 3, where adding an additional nonnative object to the scene, the model unprompted adds a supporting object to the scene, in presented cases that object is a table to support or to place the vase on. This happens around 4% of the time when requesting the model with exactly same prompt. It might be possible to reduce this by tweaking the hyperparameters of the model, however it is also possible to filter out these new additions due to them introducing radically different colors to the scene and by doing so changing scenes color palette noticeably more than a transparent object alone. A more advanced and robust approach could use auxiliary neural networks to detect artifacts or geometric distortions.



Figure 4: Severe cases of geometric identity loss during material transformation. Top row: The model improperly flattens the structure, replacing complex architectural geometry with a flat glass pane.

Bottom row: The model introduces unnatural indentations and rotations to the glass blocks, directly conflicting with the original scene's depth data.

Whereas mentioned above, examples are facing minor changes in loss of geometric Figure 4 examples are more severe cases of this mismatch. Either covering up half of the scene with glass and making it flat (top row) or making indentations and rotating the object in an unnatural way (bottom row) these cases are severe enough to throw off any model during training. Automating checking these images would require additional intelligent system to evaluate the integrity and consistency of already difficult to work with non-Lambertian objects [29,30].



Figure 5: Failure case in an empty scene. When the model fails to detect a prominent 3D furniture piece, it incorrectly applies the requested material style transfer to 2D wall art, rendering the painted chair as a transparent physical object.

Finally, in Figure 5 an issue arises from lack of objects present in the scene. When given the prompt to change an object that is prevalent to the scene and failing to detect a valid foreground object leads to unintended transformations, such as the material alteration of 2D wall art. Wall painting is featuring a woman sitting on a chair, this leads to the model changing the material of this chair into a transparent glass like material. This fails to transform an object that is actually in the scene as a physical object and therefore is considered a failure.

To perform initial feasibility testing on the “background” focused model, first round of training was performed. This training involved a limited number of training images in a very static environment. Training dataset contained 200 images and all of them were in the same room but at different angles. Model that was chosen to finetune was *Depth Anything V2 (Small)* [1]. Results of this initial training can be seen in Figure 6 where the generated image is presented in the left most case, pretrained models depth maps are shown in the middle and the finetuned models predictions are present at the rightmost column.

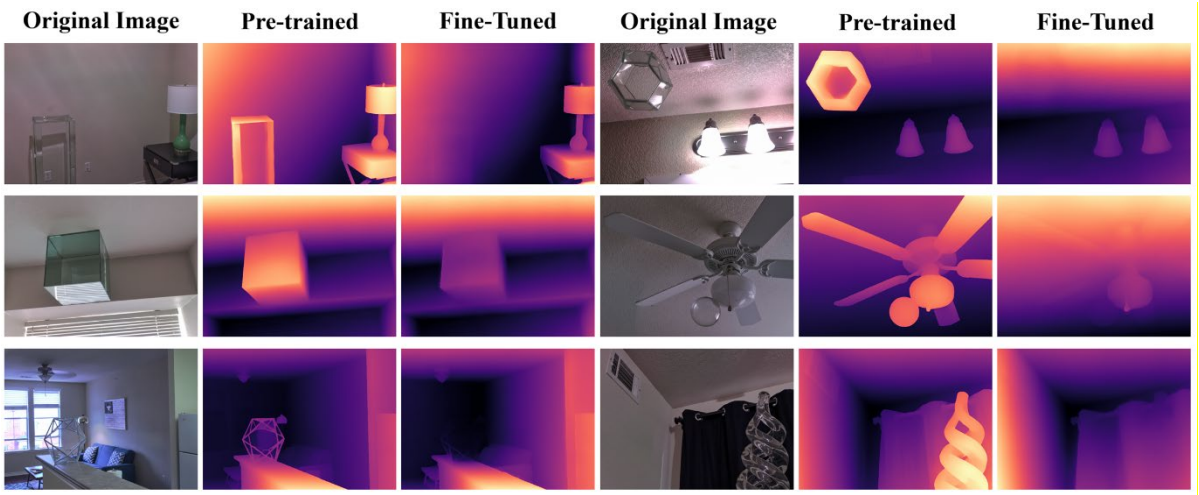


Figure 6: Initial validation results from fine-tuning the Depth Anything V2 (Small) model on the synthesized background-focused dataset. Left: Original input image. Middle: Depth map predicted

by the baseline pre-trained model. Right: Depth map predicted by the fine-tuned model, demonstrating its emerging ability to ignore the transparent object and map the scene behind it.

When analyzing samples, a background focused model is feasible and in some example cases already satisfactory, however in other cases the model fails to eliminate the transparent object and loses fine detail around the area of this object.

5. Issues

Due to the nature of diffusion models, issues arise in both expected and unexpected, unprompted ways. One of the biggest and most difficult to avoid issues comes from transformation of material required for the foreground dataset. When given a prompt "Identify a prominent furniture piece. Change its material to transparent glass. Preserve shape, shadow, position. Do not change background." The resulting generations are not consistently predictable. Object, for example a bowl, can have infinite amount of 3D scenes to represent the same 2D image that is seen. After converting objects into glass material sometimes, diffusion model also adds some geometric quirks such as a hole or a hollower point or a curve that is not in the original photo. This leads to changes in 3D objects shape which directly conflicts with ground truth depth maps. A constraint-aware diffusion models or hybrid approaches combining generative methods with explicit 3D priors could be integrated to minimize forementioned issues.

Future research must focus on maintaining strict geometric identity during style transfer. Some tools are available, for example, using color palettes to ensure no new unwanted object were not added to the scene because the new object is supposed to be transparent color palettes should remain similar to the original picture, but because the transparent object are also reflective by nature it is also possible that the new generated object will reflect light in a specific way and drastically alter the color palettes. Another tool that could be used to determine the shape of the generated object is contours. However, this does not work so well when it comes to transparent or reflective objects because the new transparent object possesses more complex visual features than the original opaque object, because it is seen through it inherits all the features of the background as well as light bending features of a glass object. Using contouring might provide the main outline contour of the object but cannot ensure the internal structure of the object.

6. Conclusion

In conclusion, the physical limitations of LiDAR and other traditional depth sensors on non-Lambertian surfaces create a critical data void in monocular depth estimation. Generative artificial intelligence offers a powerful, optimization-driven solution to this problem. Instead of investing heavily in acquiring new, difficult to capture data with specialized sensors, proposed pipeline expands upon already acquired, high-quality datasets. By use of advanced diffusion models, this methodology enables to automate creation of large quantities of photorealistic, dual-plane datasets. This approach successfully provides the necessary data to train MDE models to perceive both the "foreground" and the "background" of transparent objects, pushing the boundaries of what depth estimation networks can achieve in complex, everyday environments full of non-Lambertian bodies. A promising new step would be to jointly train unified models capable of reasoning about both foreground and background depth simultaneously.

7. Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, H. Zhao, Depth Anything V2, (2024). https://doi.org/10.48550/arXiv.2406.09414.

[2] L. Piccinelli, Y.-H. Yang, C. Sakaridis, M. Segu, S. Li, L.V. Gool, F. Yu, UniDepth: Universal Monocular Metric Depth Estimation, (2024). https://doi.org/10.48550/arXiv.2403.18913.

[3] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, V. Koltun, Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer, (2020). https://doi.org/10.48550/arXiv.1907.01341.

[4] W. Choi, K. Hwang, W. Peng, M. Choi, S. Im, Self-supervised Monocular Depth Estimation Robust to Reflective Surface Leveraged by Triplet Mining, (2025). https://doi.org/10.48550/arXiv.2502.14573.

[5] A. Costanzino, P.Z. Ramirez, M. Poggi, F. Tosi, S. Mattoccia, L.D. Stefano, Learning Depth Estimation for Transparent and Mirror Surfaces, (2023). https://doi.org/10.48550/arXiv.2307.15052.

[6] M. Shao, C. Xia, D. Duan, X. Wang, Polarimetric Inverse Rendering for Transparent Shapes Reconstruction, (2022). https://doi.org/10.48550/arXiv.2208.11836.

[7] M.W. Tao, T.-C. Wang, J. Malik, R. Ramamoorthi, Depth Estimation for Glossy Surfaces with Light-Field Cameras, in: L. Agapito, M.M. Bronstein, C. Rother (Eds.), Computer Vision - ECCV 2014 Workshops, Springer International Publishing, Cham, 2015: pp. 533–547. https://doi.org/10.1007/978-3-319-16181-5_41.

[8] S.S. Sajjan, M. Moore, M. Pan, G. Nagaraja, J. Lee, A. Zeng, S. Song, ClearGrasp: 3D Shape Estimation of Transparent Objects for Manipulation, (2019). https://doi.org/10.48550/arXiv.1910.02550.

[9] H. Xu, Y.R. Wang, S. Eppel, A. Aspuru-Guzik, F. Shkurti, A. Garg, Seeing Glass: Joint Point Cloud and Depth Completion for Transparent Objects, (2021). https://doi.org/10.48550/arXiv.2110.00087.

[10] Q. Dai, J. Zhang, Q. Li, T. Wu, H. Dong, Z. Liu, P. Tan, H. Wang, Domain Randomization-Enhanced Depth Simulation and Restoration for Perceiving and Grasping Specular and Transparent Objects, (2022). https://doi.org/10.48550/arXiv.2208.03792.

[11] H. Wen, Y. Zuo, V. Subramanian, P. Chen, J. Deng, Seeing and Seeing Through the Glass: Real and Synthetic Data for Multi-Layer Depth Estimation, (2025). https://doi.org/10.48550/arXiv.2503.11633.

[12] T. Tommasi, N. Patricia, B. Caputo, T. Tuytelaars, A Deeper Look at Dataset Bias, (2015). https://doi.org/10.48550/arXiv.1505.01257.

[13] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, P. Abbeel, Domain Randomization for Transferring Deep Neural Networks from Simulation to the Real World, (2017). https://doi.org/10.48550/arXiv.1703.06907.

[14] J. Tremblay, A. Prakash, D. Acuna, M. Brophy, V. Jampani, C. Anil, T. To, E. Cameracci, S. Boochoon, S. Birchfield, Training Deep Networks with Synthetic Data: Bridging the Reality Gap by Domain Randomization, (2018). https://doi.org/10.48550/arXiv.1804.06516.

[15] A. Mumuni, F. Mumuni, N.K. Gerrar, A survey of synthetic data augmentation methods in computer vision, Mach. Intell. Res. 21 (2024) 831–869. https://doi.org/10.1007/s11633-022-1411-7.

[16] J. Guo, S. Yang, Y. Huang, Y. Long, X. Xia, X. Su, B. Zhao, Z. Xie, L. Nie, UtilGen: Utility-Centric Generative Data Augmentation with Dual-Level Task Adaptation, (2025). https://doi.org/10.48550/arXiv.2510.24262.

[17] G. Valvano, A. Agostino, G.D. Magistris, A. Graziano, G. Veneri, Controllable Image Synthesis of Industrial Data Using Stable Diffusion, (2024). https://doi.org/10.48550/arXiv.2401.03152.

[18] Z. Weng, L. Bravo-Sánchez, S. Yeung-Levy, Diffusion-HPC: Synthetic Data Generation for Human Mesh Recovery in Challenging Domains, (2023). https://doi.org/10.48550/arXiv.2303.09541.

[19] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-Resolution Image Synthesis with Latent Diffusion Models, (2022). https://doi.org/10.48550/arXiv.2112.10752.

[20] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, R. Rombach, SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis, (2023). https://doi.org/10.48550/arXiv.2307.01952.

[21] T. Brooks, A. Holynski, A.A. Efros, InstructPix2Pix: Learning to Follow Image Editing Instructions, (2023). https://doi.org/10.48550/arXiv.2211.09800.

[22] Black Forest Labs, FLUX.1-dev, (2024). https://huggingface.co/black-forest-labs/FLUX.1-dev.

[23] V. Arkhipkin, V. Korviakov, N. Gerasimenko, D. Parkhomenko, V. Vasilev, A. Letunovskiy, N. Vaulin, M. Kovaleva, I. Kirillov, L. Novitskiy, D. Koposov, N. Kiselev, A. Varlamov, D. Mikhailov, V. Polovnikov, A. Shutkin, J. Agafonova, I. Vasiliev, A. Kargapoltseva, A. Dmitrienko, A. Maltseva, A. Averchenkova, O. Kim, T. Nikulina, D. Dimitrov, Kandinsky 5.0: A Family of Foundation Models for Image and Video Generation, (2025). https://doi.org/10.48550/arXiv.2511.14993.

[24] Gemini Team, Gemini: A Family of Highly Capable Multimodal Models, (2025). https://doi.org/10.48550/arXiv.2312.11805.

[25] M. Ohara, J. Kim, K. Koida, The Role of Specular Reflections and Illumination in the Perception of Thickness in Solid Transparent Objects, Front. Psychol. 13 (2022) 766056. https://doi.org/10.3389/fpsyg.2022.766056.

[26] A. Blake, H. Bülthoff, Does the brain know the physics of specular reflection?, Nature 343 (1990) 165–168. https://doi.org/10.1038/343165a0.

[27] I. Vasiljevic, N. Kolkin, S. Zhang, R. Luo, H. Wang, F.Z. Dai, A.F. Daniele, M. Mostajabi, S. Basart, M.R. Walter, G. Shakhnarovich, DIODE: A Dense Indoor and Outdoor DEpth Dataset, (2019). https://doi.org/10.48550/arXiv.1908.00463.

[28] N. Silberman, D. Hoiem, P. Kohli, R. Fergus, Indoor Segmentation and Support Inference from RGBD Images, in: A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, C. Schmid (Eds.), Computer Vision – ECCV 2012, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012: pp. 746–760. https://doi.org/10.1007/978-3-642-33715-4_54.

[29] A. Mumuni, F. Mumuni, N.K. Gerrar, A survey of synthetic data augmentation methods in computer vision, Mach. Intell. Res. 21 (2024) 831–869. https://doi.org/10.1007/s11633-022-1411-7.

[30] X. Dong, M.A. Garratt, S.G. Anavatti, H.A. Abbass, Towards Real-Time Monocular Depth Estimation for Robotics: A Survey, (2021). https://doi.org/10.48550/arXiv.2111.08600.