Enhancing Stereo Depth Maps through RGBD-Conditioned Generative Models
CS598 3D-Vision Project Final Report

Jose Cuaran Asher Mai Kulbir Ahluwalia Junzhe Wu
University of Illinois Urbana-Champaign
{jrc9, hanlinm2, ksa5, junzhew3}@illinois.edu

1. Motivation

The cost-effectiveness and deployment simplicity of inexpensive RGB-D sensors have led to their growing use in robotics and agricultural applications. These sensors offer depth (D) and dense color (RGB) information, which is essential for applications like 3D reconstruction, object perception, and navigation. Low-cost sensors, however, frequently have substantial noise, occlusions, missing regions, and sensor-specific aberrations that skew the raw depth data they collect. These imperfections hinder downstream tasks, limiting the sensors' utility in real-world scenarios where accurate and reliable depth information is essential.

By reducing noise and filling in missing values, traditional depth completion algorithms seek to transform unprocessed sensor outputs into clear, dense depth maps. The varied and intricate artifacts found in field settings, particularly in agricultural settings, are frequently difficult for these approaches to handle. Simultaneously, monocular depth estimation methods have made impressive strides in generating depth predictions that are both visually smooth and semantically rich. Despite their popularity, these learning-based monocular techniques fall short of geometrically accurate depth measurements and scale consistency due to reliance on monocular pictures rather than depth signals.

Our work proposes a novel generative depth completion method that bridges this gap by combining conditional generative techniques with raw depth sensor data. We mitigate the drawbacks of monocular depth estimation while preserving important geometric clues by conditioning a diffusion-based generative model on the noisy but actual sensor depth.

Improved depth completion performance can enable downstream tasks such as geometrically accurate depth maps using low-cost RGBD sensors. For instance, enhanced field robot mapping and navigation, enhanced manipulation and detection of regions of interest, resilience to changing lighting conditions, leaf angle estimation in precision farming, 3D object reconstruction, and keypoint matching for safer, more effective robot operation.

2. Related Work

2.1. Sim-to-Real

Depth images have lower sim-to-real gap as compared to RGB images since they are not prone to inaccurate simulated visual features in simulations[4, 5, 14]. For passive stereo sensors, it is harder to simulate external lighting conditions which is why active infrared based stereo sensors is predominantly used to measure depth indoors.

However, for active infrared depth measurement, the errors in depth depend on the surface properties of the object such as transparency, reflectance and transmission which are harder to recreate in simulations. A noise free approximation of a scene's depth image can be obtained by averaging multiple frames of the same static scene. The MAP (Maximum a priori) estimate can be used for noise model parameter estimation[19].

For reducing the sim-to-real gap, object material parameters can be approximated for aligned simulated Visible and Infrared spectrum visual images using a multispectral matching loss function[19] which adds a perceptual loss term to the L2 loss between simulation and real RGB image features. The perceptual loss is defined as the L2 difference between the AlexNet visual features that reduces the need for color, exposure and lighting condition alignment thereby improving the accuracy of material acquisition and results in better rendering.

2.2. Simulating sensor noise

Simulations have often used simplified models for thermal camera noise, surface lighting, reflectance, radiance, refraction and soft shadows which leads to a large domain gap between real world data and synthetic simulation data. For infrared sensors, Sensor specific noise such as laser speckle and thermal noise can be modelled[3] by using a element wise multiplicative term for laser speckle and a element wise additive term for camera thermal gaussian noise[7].

Prior approaches for simulating depth sensor noise and estimating depth from monocular RGB images have limited transferability due to lack of geometric accuracy and scale.

This results in unrealistically smooth depth maps with an unkown scale and cannot be used for tasks that require precise geometric information.

Geometric and semantic information from RGB and depth have only been combined using simple fusion techniques without focussing on the common and complementary information from both 'RGB' and 'Depth' modalities for a improved concatenated representation. This can be crucial in scenarios with varying visual RGB features but stable depth features or vice versa.

Geometric information in the generated depth can be preserved during generation by making the differentiable features invariant[13]. However, it lacks surface property dependent depth errors. Differentiable ray tracing has used to estimate depth in simulations by optimizing for rendering and stereo matching parameters[8].

2.3. Depth Completion

For the task of depth completion, missing depth values for input RGBD images can be generated by predicting surface normals and occlusion boundaries from color to get complete output depth[20].

Early depth completion methods often lacked geometric constraints, leading to inconsistencies in the predicted depth maps, especially in complex scenes with intricate geometries. Ignoring geometric information can result in depth predictions that do not align with the underlying 3D structure of the scene. To address this limitation, recent approaches have incorporated surface normal guidance to enforce geometric consistency in depth completion [9, 16, 21].

For instance, methods like DeepLiDAR [9] integrate surface normal information derived from color images to improve the accuracy and reliability of the completed depth maps. By leveraging surface normals and geometric constraints, these methods produce depth predictions that are more consistent with the scene geometry, leading to better performance in regions with intricate structures and discontinuities.

3. Method

We condition a generative model on noisy and incomplete RGBD sensor data to generate accurate and smooth depth maps. Specifically, a latent diffusion model (Stable Diffusion v2 [11]) is trained to iteratively denoise a latent representation of our depth map. We employ pretrained diffusion models for text-to-image generation, trained on large datasets, to exploit semantic and geometric priors that can improve the performance of our method while reducing the required training data and time. A robust conditioning strategy is implemented to ensure that these pretrained models do not suffer from overfitting or catastrophic forgetting.

Training. Fig. 1 shows an overview of our training pipeline. Following Marigold [2], we adapt the pretrained

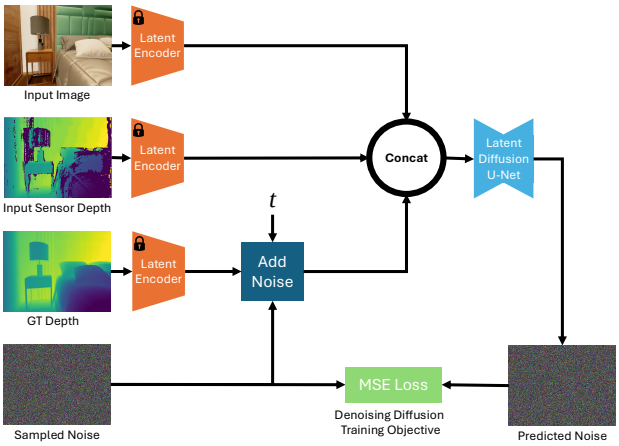


Figure 1. System overview for training

U-Net from Stable Diffusion V2 [11] as the latent denoiser and use the corresponding VAE to encode the input image, the input sensor depth map, and the ground-truth depth map. Since the VAEs were designed for 3-channel RGB inputs, we replicate the single-channel depth inputs into 3 channels each. The VAE is not retrained for depth; as demonstrated in [2], the Stable Diffusion VAE can encode and decode these replicated depth maps with minimal reconstruction error.

We then concatenate all three resulting latent codes (RGB input image, noisy incomplete sensor depth, and ground-truth depth with added sampled noise), each of which has 4 latent channels, resulting in a total of 12 input channels along the feature dimension to the U-Net. To preserve the pretrained weights of the U-Net's first layer, we duplicate its original input layer weights (trained on 4-channel inputs) three times along the channel dimension, and then divide all the weights by 3. This ensures that each newly added set of 4 channels is initialized similarly to the original, preventing a sudden change in activation magnitudes and ensure stable training. As a result, the scale of the network's inputs and activations is kept consistent, allowing the U-Net to rely on its previously learned feature extraction capabilities and prevent catastrophic forgetting. The latent denoiser is trained to predict the added Gaussian noise with uniformly sampled noise level t.

Inference. Fig. 2 shows an overview of our pipeline for inference. At inference time, we concatenate the latent codes of input image, input sensor depth and depth map, which is a normally-distributed Gaussian noise initially. The latent depth map is iteratively denoised by the latent denoiser T steps before being decoded, and we take the average of the three channels to be the predicted depth map.

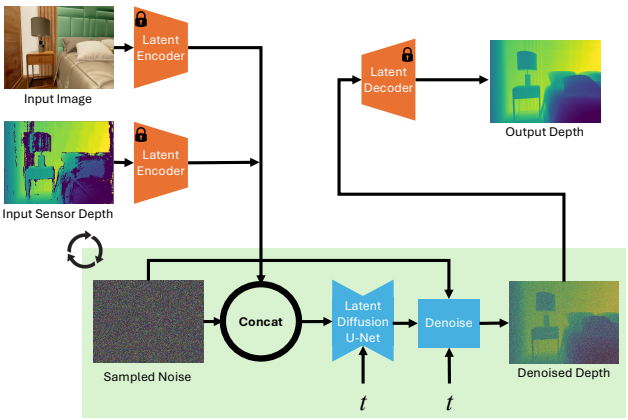


Figure 2. System overview for inference

4. Experiments

4.1. Baselines

We compare our method with state-of-the-art monocular depth estimation methods including:

- **Marigold [2]**. A diffusion-based monocular depth estimation method pre-trained on the LAION-5B dataset [12] for promt-image generation, and finetuned for depth estimation on 74k image samples from the Hypersim [10] and VKITTI [1] syntethic datasets.
- **Depth Anything V2 [17](small)**. A transformer-based discriminative method for monocular depth estimation, trained on 65 million image samples from different indoor and outdoor datasets.
  The predicted depth maps are aligned to the ground truth using least squares method before evaluation.

4.2. Datasets

Training datasets. We train our model on the IRS [15] and VKITTI [1] datasets. Both datasets contain synthetic stereo images, that are used to generate a total of 35k training samples with RGB, ground truth depth and sensor depth. The IRS dataset [15] includes images from indoor scenes like hotel rooms and stores. The VKITTI [1] dataset contains images from urban environments collected from a vehicle in a simulated environment.

1. Indoor Robotics Stereo (IRS) Dataset [15]
2. NYUv2 dataset [6]
3. Virtual KITTI (VKITTI) [1]

Evaluation datasets. We evaluate our approach on a subset of 350 samples from the IRS dataset, 410 samples from the VKITTI [1] dataset, both seen during training, and 200 samples from the NYU-V2 dataset [6] for zero-shot evaluation. The NYU-V2 dataset [6] contains RGBD images from indoor environments, captured with the Kinect sensor.

4.3. Metrics

We use standard metrics for evaluation of depth estimation methods, including:

- **Mean Absolute relative error (MAE)**

$$\text{AbsRel} = \frac{1}{T}\sum_{i \in T}\frac{|d_i^{gt} - d_i|}{d_i^{gt}}$$

- **Mean squared Relative error (REL)**

$$\text{SqRel} = \frac{1}{T}\sum_{i \in T}\frac{|d_i^{gt} - d_i|^2}{d_i^{gt}}$$

- **Root-mean-squared error (RMSE)**

$$\text{RMSE} = \sqrt{\frac{1}{T}\sum_{i \in T}(d_i^{gt} - d_i)^2}$$

- **Accuracy with threshold:** δ_t. Percentage of pixels such that:

$$\max\left(\frac{d_i^{gt}}{d_i}, \frac{d_i}{d_i^{gt}}\right) < t$$

Where d_i^{gt} is the ground-truth depth for pixel i, d_i is the estimated depth, and T is the total number of pixels in all the evaluated images.

The evaluation is conducted after aligning the predicted depth map to the ground truth depth maps by least-squares fitting. In addition, as an alternative strategy, we perform scale and shift correction for our method leveraging the sensor depth map. This is done by denormalizing the predicted depth map with the scale and shift factors used during the normalization of the sensor depth map.

5. Results

Fig. 3 shows sample depth predictions by our method and the baselines. We can see that all the methods predict similar depth maps in terms of smoothness and consistency with the ground truth. However, the error maps indicated that our model predictions exhibit lower error (darker error maps) than the baselines. This is more noticeable for samples from the IRS [15] and VKITTI [1] datasets, seen during training. Unlike these datasets, the NYUV2 dataset [6] contains real-world RGBD images, causing a performance drop due to distribution shift in color and sensor depth patterns.

Table 1 indicates that our method outperforms the baselines in most cases, resulting in accuracy improvements (delta1 metric) of 7%, 0.5% and 2.4% on the IRS [15], VKITTI [1] and NYUV2 [6] datasets. This results suggests that sensor depth maps that condition the diffusion model can effectively improve the accuracy of depth estimation methods. Furthermore, we found that our proposed strategy for scale and shift correction using the sensor depth map

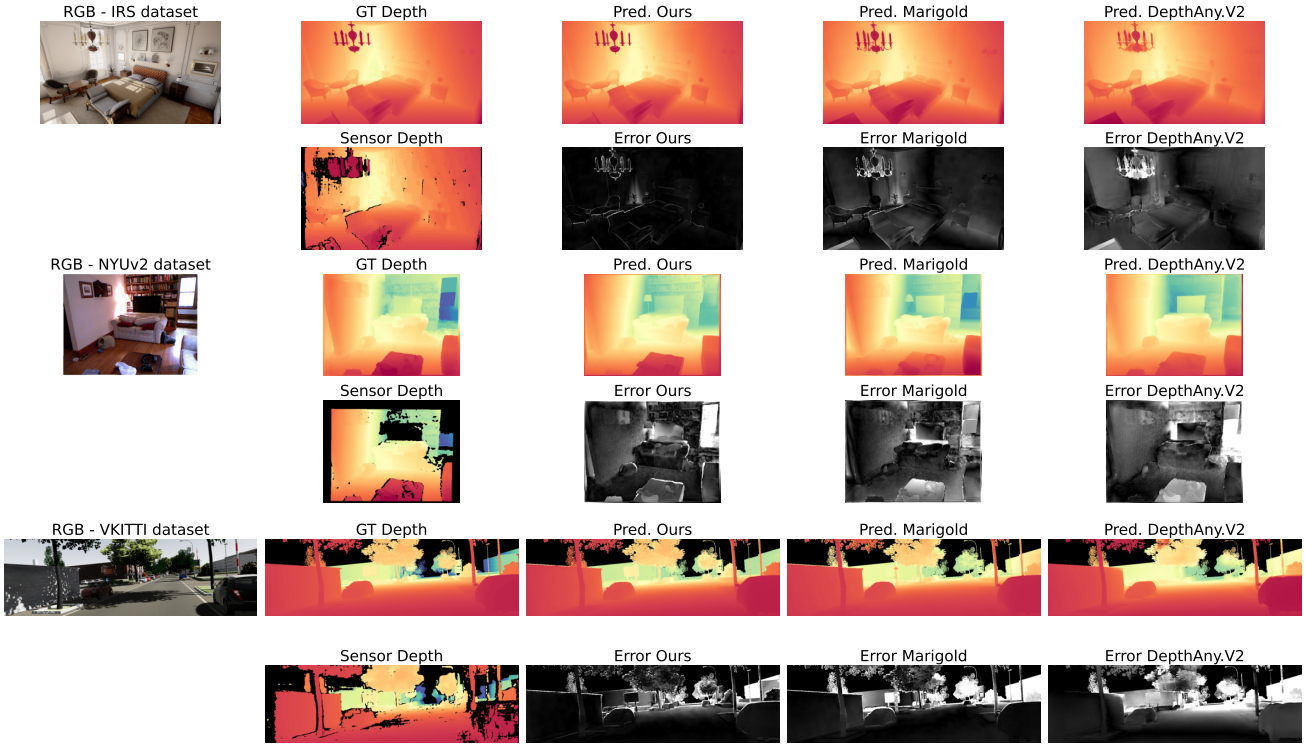


Figure 3. Depth prediction and absolute error for different methods. The error maps go from zero (black) to 10% of maximum depth (white). Note that our method exhibits darker error maps, meaning higher accuracy on depth estimation.

leads to more accurate depth predictions (2% on average) than the commonly used least-squares fitting method, in particular for the indoor datasets (IRS [15] and NYUv2 [6]). However, in the case of the VKITTI [1] dataset, the least-squares alignment surpasses our method. An explanation for this could be that the least squares method explicitly reduces larger errors, which in the case of ourdoor datasets like VKITTI [1], are concentrated in farther distances. As our method does not penalize large errors directly, it exhibits lower accuracy. However, further experiments are necessary to explain this discrepancy.

5.1. Ablation Studies

We train and evaluate two additional conditioning strategies: **(i) RGB + sensor depth + mask simple conditioning**. This uses the same pipeline shown in Fig. 1, but includes an inpainting mask to constrain the loss function to missing pixels only. This approach is inspired by color inpainting models. Instead of passing the mask through a decoder, we downsample the original mask to a 64x64 resolution which is directly concatenated with the latent vector.

(ii) ControlNet [18] conditioned on RGB and sensor depth. By leveraging zero convolutions, ControlNet has demonstrated superior performance on various conditioned generative tasks. Specifically, we concatenate the color channels with the sensor depth map and use this combined tensor as input to a CNN encoder. Unlike the original ControlNet[18], where the controlled network is frozen during training, we fix only the encoder weights and fine-tune the decoder. This adjustment is necessary because the original stable diffusion model[11] is pretrained for image generation rather than depth generation.

Fig. 7 shows sample predictions for our three conditioning strategies. While the two simple conditioning strategies (with or without inpainting masks) generate consistent depth maps, ControlNet[18] tends to produce artifacts and distort the original objects' shapes. This could be attributed to the fact that part of the controlled network remains frozen during training, making it challenging for the model to learn new geometric features. These inconsistencies result in significant performance drops for ControlNet[18], as shown in Table 2. Overall, the simple conditioning strategies (with or without the inpainting mask) outperform ControlNet[18] by more than 7% in the delta1 accuracy metric across all datasets. Finally, while the inpainting mask does not significantly improve the model's accuracy, slight improvements are observed in some cases.

Dataset	Method	Scale and shift Correction	AbsRel	SqRel	delta1_acc	delta2_acc
IRS [15]	Marigold[2]	Least Squares	0.117	0.276	0.900	0.964
	Depth Anything V2[17]	Least Squares	0.121	0.261	0.886	0.961
	Depth Completion (ours)	Least Squares	0.064	0.212	0.951	0.970
	Depth Completion (ours)	Sensor depth scale	**0.046**	**0.099**	**0.976**	**0.988**
VKITTI [1]	Marigold[2]	Least Squares	0.120	1.079	0.886	0.967
	Depth Anything V2[17]	Least Squares	0.239	2.132	0.588	0.881
	Depth Completion (ours)	Least Squares	**0.110**	**0.910**	**0.891**	**0.972**
	Depth Completion (ours)	Sensor depth scale	0.270	2.929	0.574	0.937
NYUv2 [6]	Marigold[2]	Least Squares	0.093	0.080	0.910	0.970
	Depth Anithing V2[17]	Least Squares	0.098	0.064	0.904	**0.974**
	Depth Completion (ours)	Least Squares	0.084	**0.062**	0.906	0.971
	Depth Completion (ours)	Sensor depth scale	**0.065**	0.063	**0.924**	0.965

Table 1. Comparison of depth estimation methods across IRS [15], VKITTI [1] and NYUV2 [6] datasets.

Table 2. Evaluation metrics for different conditioning strategies

Dataset	Conditioning Method	Scale and shift correction	AbsRel ↓	SqRel ↓	rmse_linear ↓	delta1_acc ↑	delta2_acc ↑
IRS [15]	RGB + sensor depth	Least Squares	0.064	0.212	**0.464**	0.951	0.970
		Sensor depth scale	**0.046**	**0.099**	0.540	**0.976**	**0.988**
	RGB + sensor depth + mask	Least Squares	0.076	0.228	0.503	0.944	0.969
		Sensor depth scale	0.070	0.179	0.601	0.964	0.984
	ControlNet (RGB + sensor depth)	Least Squares	0.143	0.342	0.654	0.885	0.941
		Sensor depth scale	0.182	0.639	0.818	0.876	0.949
NYUv2 [6]	RGB + sensor depth	Least Squares	0.084	0.062	0.388	0.906	0.971
		Sensor depth scale	**0.065**	0.063	0.416	**0.924**	0.965
	RGB + sensor depth + mask	Least Squares	0.083	**0.057**	**0.372**	0.911	**0.977**
		Sensor depth scale	0.071	0.058	0.394	0.924	0.972
	ControlNet (RGB + sensor depth)	Least Squares	0.140	0.143	0.527	0.847	0.942
		Sensor depth scale	0.140	0.160	0.567	0.853	0.942
VKITTI [1]	RGB + sensor depth	Least Squares	0.110	**0.910**	4.967	0.891	0.972
		Sensor depth scale	0.270	2.929	8.369	0.574	0.937
	RGB + sensor depth + mask	Least Squares	**0.105**	0.955	**4.855**	**0.920**	**0.974**
		Sensor depth scale	0.191	1.360	5.333	0.804	0.967
	ControlNet (RGB + sensor depth)	Least Squares	0.154	1.852	6.716	0.861	0.948
		Sensor depth scale	0.302	3.197	7.930	0.524	0.915

6. Conclusion and Future Work

Diffusion models conditioned on RGB and noisy sensor depth exhibit a slight improvement in accurate depth estimation compared to pure monocular depth estimation methods. This suggests that sensor depth provides valuable guidance to the model for estimating accurate depth.

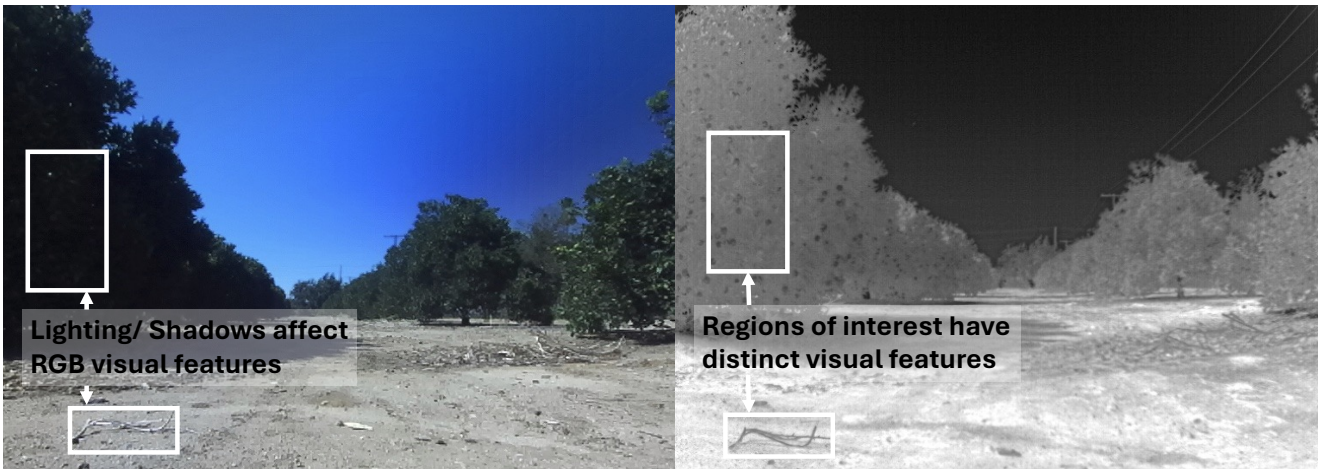
Our approach demonstrates generalization to real-world indoor scenes, despite being fine-tuned on synthetic data only. This can be attributed to the fact that the original Stable Diffusion model was trained on large datasets, capturing strong semantic and geometric priors from diverse domains.

Future work aims to expand the conditioning inputs beyond RGB and standard depth by incorporating additional modalities such as thermal (Fig 4), monochrome, near-infrared (NIR) images, semantic and instance masks to capture both local and global features at the object part level to improve generalization to other datasets and camera sensors. To incorporate more conditioning inputs and modalities, such as point clouds, text, thermal images, and near-infrared images, the frozen VAE encoder can be selectively chosen for each specific input modality to improve the resultant embeddings.

References

[1] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*, 2020. 3, 4, 5

[2] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9492–9502, 2024. 2, 3, 5, 8, 9



(a) Left RGB Camera Image (Zed 2i) (b) Thermal Camera Image (FLIR ADK)

Figure 4. Comparison of (a) RGB and (b) Thermal camera images for feature detection in outdoor environments. The RGB image highlights challenges such as lighting and shadows affecting visual features, while the thermal image provides distinct and consistent features for regions of interest, such as obstacles or fruits, aiding in robust perception.

[3] Michael J. Landau, Bryan Y. Choo, and Peter A. Beling. Simulating kinect infrared and depth images. *IEEE Transactions on Cybernetics*, 46(12):3018–3031, 2016. 1

[4] Yoav Litvak, Ariel Biess, and Aharon Bar-Hillel. Learning pose estimation for high-precision robotic assembly using simulated depth images. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 3521–3527. IEEE, 2019. 1

[5] Jeff Mahler, Jacky Liang, Surat Tejomurtula Niyaz, Michael Laskey, Ryan Doan, Xinyu Liu, Jose Ojea, and Ken Goldberg. Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics. In *Proceedings of Robotics: Science and Systems (RSS)*, 2017. 1

[6] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 3, 4, 5

[7] D. V. Perepelitsa. Johnson noise and shot noise. MIT Junior Lab (Course 8.13/8.14) Experiment, 2006. Available at http://web.mit.edu/8.13/www/JLExperiments/JLExp05.pdf. 1

[8] Baptiste Planche and Ram Vasudevan. Physics-based differentiable depth sensor simulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14387–14397. IEEE, 2021. 2

[9] Jiagang Qiu, Zhaopeng Cui, Yao Zhang, Shuaicheng Zhang, Xiaowei Gao, Bing Zeng, Marc Pollefeys, and Dengxin Dai. Deeplidar: Deep surface normal guided depth prediction for outdoor scene from sparse lidar data and single color image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3313–3322, 2019. 2

[10] 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. 3

[11] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 4, 8

[12] 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

[13] Yecheng Shen, Yao Yang, Ying Zheng, C. Karen Liu, and Leonidas J. Guibas. DCL: Differential contrastive learning for geometry-aware depth synthesis. *IEEE Robotics and Automation Letters*, 7(2):4845–4852, 2022. 2

[14] Uwe Viereck, Arjun Pas, Kate Saenko, and Robert Platt. Learning a visuo-motor controller for real world robotic grasping using simulated depth images. In *Proceedings of the Conference on Robot Learning (CoRL)*, pages 291–300, 2017. 1

[15] 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, 2021. 3, 4, 5

[16] Yan Xu, Xinzhu Zhu, Jianping Shi, Guangming Zhang, Hujun Bao, and Hujun Li. Depth completion from sparse lidar data with depth-normal constraints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2811–2820, 2019. 2

[17] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024. 3, 5, 8, 9

[18] 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. 4, 11, 12

[19] Xiaoshuai Zhang, Rui Chen, Ang Li, Fanbo Xiang, Yuzhe Qin, Jiayuan Gu, Zhan Ling, Minghua Liu, Peiyu Zeng, Songfang Han, Zhiao Huang, Tongzhou Mu, Jing Xu, and Hao Su. Close the optical sensing domain gap by physics-grounded active stereo sensor simulation. *IEEE Transactions on Robotics*, 39 (3):2429–2447, 2023. 1

[20] Yinda Zhang and Thomas Funkhouser. Deep depth completion of a single rgb-d image. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 175–185, 2018. 2

[21] Yinda Zhang and Thomas Funkhouser. Deep depth completion of a single rgb-d image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 175–185, 2018. 2

Appendix

A. Code

All code references, evaluation scripts, and configuration files are publicly available at our GitHub repository: [CS598SHW_Depth_Completion].

B. Implementation details

- **Pretrained Models:** We use the Stable Diffusion V2 weights for both U-Net and VAE encoder. [11].
- **Baselines:** We compare against Marigold [2] and Depth Anything V2 (DAV2) [17]. Pre-trained models and code for implementation were obtained from their respective official repositories:
  1. [Marigold code and documentation]
  2. [Depth Anything V2 code and documentation]
- **Training:** Training our method takes 36K iterations using a batch size of 32. Training our method to convergence takes approximately 3 days on a single Nvidia A100 GPU.
- **Evaluation Metrics:** We implement standard metrics (MAE, REL, RMSE, $\delta 1$, $\delta 2$) as described in the main paper.
- **Hyperparameters:** We use a learning rate of 3×10^{-5}, Adam optimizer, and a noise schedule adapted from Stable Diffusion V2 for noise injection. For further details, please refer to the repository's configuration files.

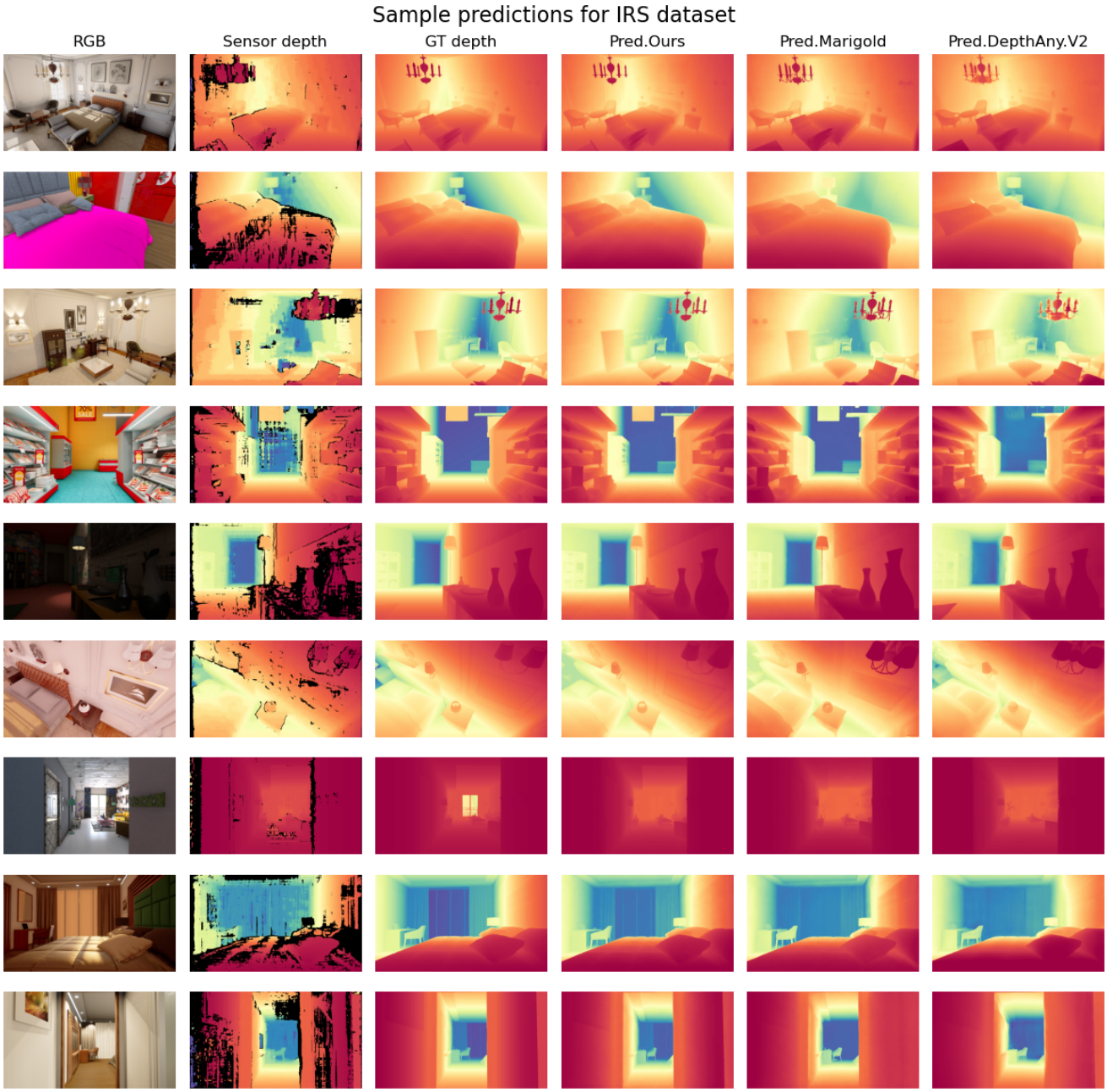


Figure 5. Comparison of Depth Predictions on the IRS Dataset. Predictions using our approach (Pred. Ours) leverages sensor depth to produce more geometrically consistent depth maps than monocular-only baselines. Even in challenging indoor scenes with cluttered objects and complex lighting, our method preserves fine details such as complex geometry lanterns, thin light tubes, thin structures in furniture and mitigates artifacts in many cases that persist in monocular depth maps from Marigold[2] and DepthAnyV2[17].

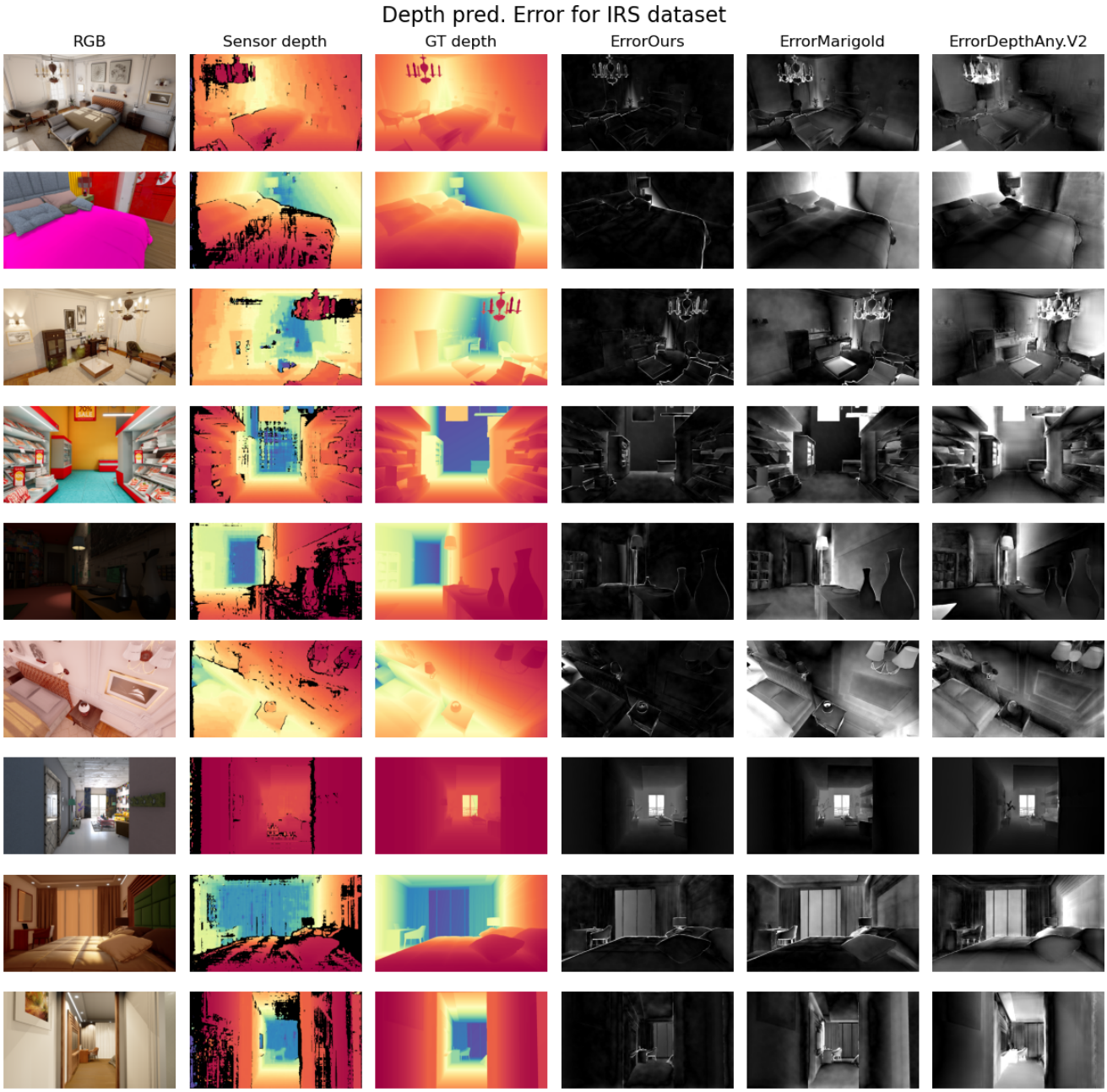


Figure 6. Depth Error Visualization for IRS Dataset. The error maps (calculated as absolute differences from ground truth) highlight the regions where predictions deviate from ground truth. Our method's error maps exhibit fewer high-error regions compared to monocular baselines. This suggests that sensor conditioning provides valuable geometric cues, especially in areas with low texture, thin geometry, small objects or ambiguous structures.

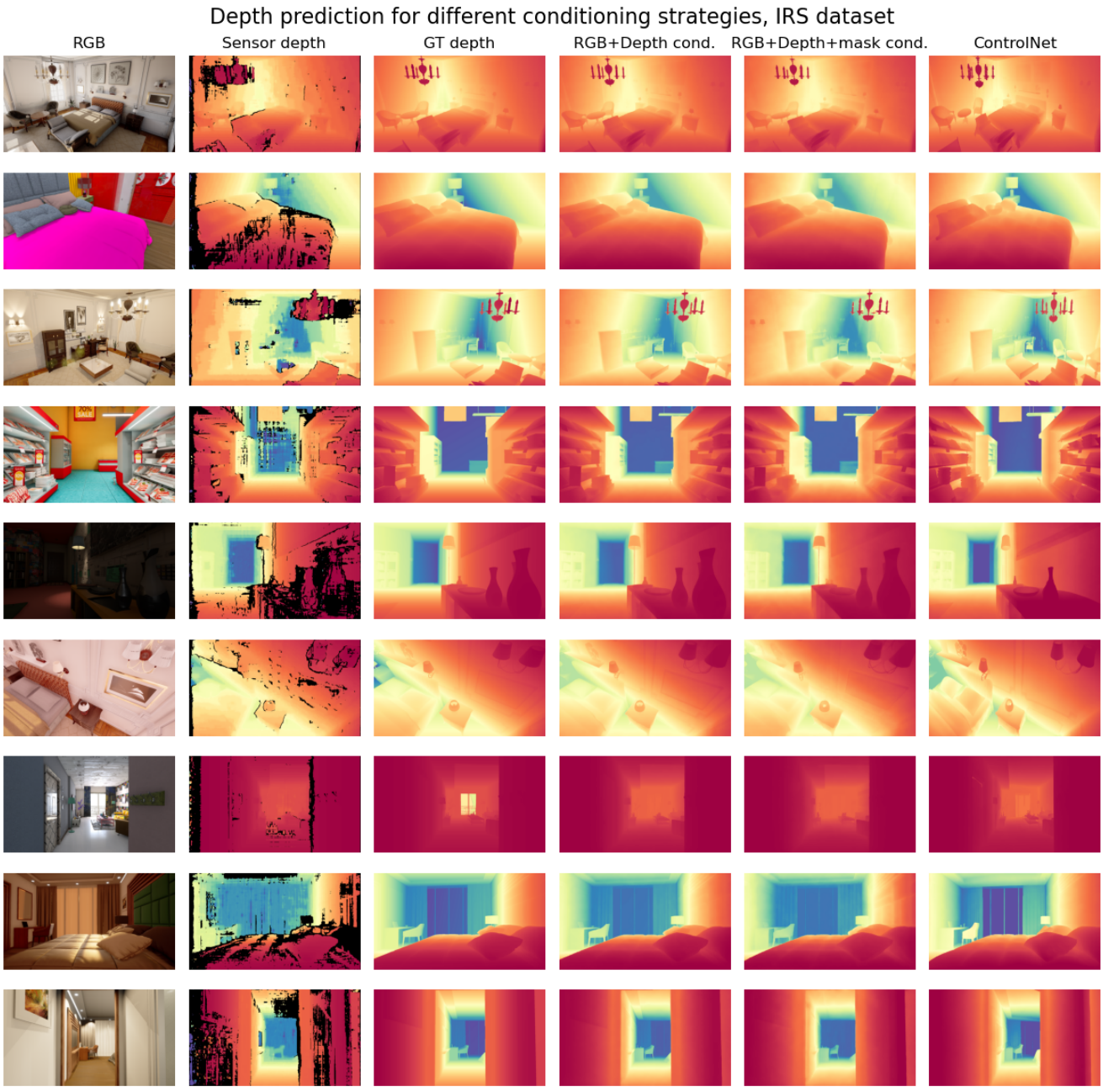


Figure 7. Impact of different conditioning strategies on depth prediction. Simple conditioning with RGB and sensor depth predicts more consistent depth maps than ControlNet[18]. ControlNet[18] tends to predict wrong object's shape and artifacts.

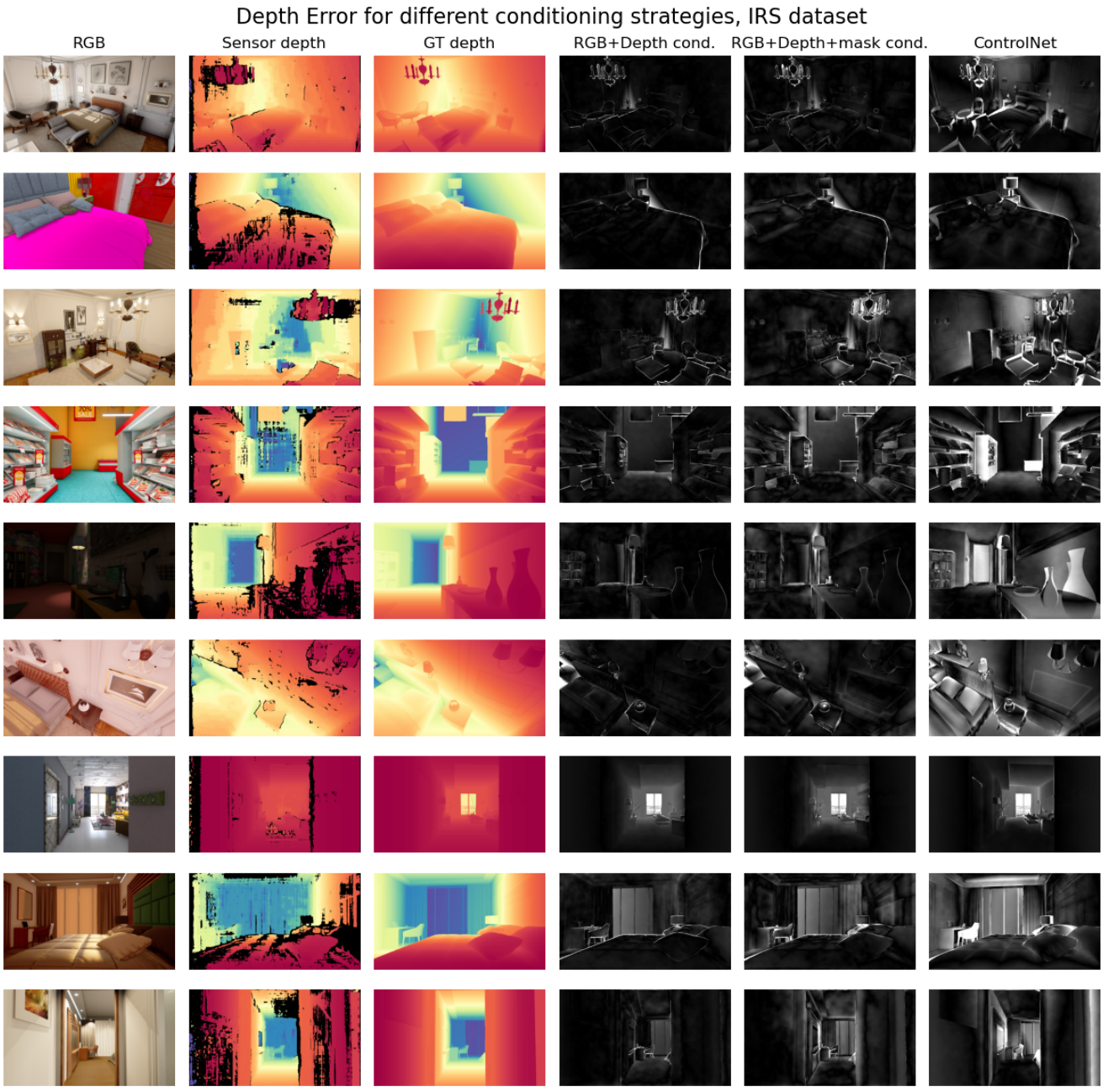


Figure 8. Depth prediction error for multiple conditioning strategies. Simple conditioning with RGB and sensor depth leads to more accurate depth predictions than complex conditioning strategies like ControlNet[18].