Hierarchical Convolution Spatial Propagation Networks for Depth Completion

Praveen Ramesh [1] K Vikas Gowda [1] Kunal A Kathare [1]
Basavaraja Shanthappa Vandrotti [1] Jooyoung Kim[2] Ankit Dhiman[1]
[1]Samsung R&D Institute India- Bangalore [2]Samsung Electronics, Suwon, Korea

***Abstract*—Accurate depth estimation is crucial for applications like AR/VR but faces challenges due to the sparsity of depth sensors like LiDAR and the limitations of monocular and stereo-based methods in texture-less or distant regions. To address these issues, depth completion—converting sparse measurements into dense depth maps—has emerged as a key area of research in both industrial and academic fields. Recent methods for depth completion use RGB images to enhance sparse depth data but often produce blurry and distorted depth maps despite iterative refinements by spatial propagation networks (SPNs), which can be computationally intensive. While approaches like LRRU reduce complexity by refining initial depth maps, they rely on hardware-intensive operations, limiting their applicability for real-time applications on embedded systems. In contrast, we propose an efficient network that is highly suitable for deployment on embedded devices. We propose a dual-encoder architecture followed by a hierarchical update scheme that mitigates the usage of computationally expensive operations like dynamic convolution. Our method achieves competitive performance compared to the state-of-the-art methods on challenging NYUv2 dataset, the TOFDC dataset, and an in-house synthetic dataset: SynT.**

I. Introduction

Accurate depth estimation is critical for a wide range of applications, including AR/VR headsets [1] and autonomous vehicles [2]. However, inherent limitations are faced by existing depth-sensing technologies in capturing detailed depth information across both indoor and outdoor environments. For example, sparse measurements are produced by depth sensors such as LiDAR, dToF, and iToF, and their direct applicability is restricted. Monocular depth estimation methods [3], while promising, often lack the precision required for tasks like surface reconstruction. Additionally, stereo-based approaches [4] struggle with texture-less regions and exhibit reduced accuracy for distant objects. These challenges have led to increasing interest in the depth completion task which transforms sparse distance measurements into dense depth maps.

Traditional methods leverage rich semantic information from RGB images [5] to complete sparse depth data. Other methods rely solely on sparse depth data [6], which often leads to inaccurate results when the input depth is extremely sparse. However, these approaches exhibit blurriness and distorted object boundaries in the estimated depth maps. To address these limitations, popular techniques such as NLSPN [7] and CSPN [8] employ spatial propagation networks (SPNs) to iteratively refine the predicted depth. While effective, this refinement process is computationally intensive, often requiring 20 to 30 iterations to achieve accurate results. To reduce complexity, recent methods like LRRU [9] iteratively refine an initial depth map generated by non-learning-based approaches, avoiding direct regression from sparse inputs. However, these methods rely on computationally demanding operations that are not supported by embedded systems, limiting their applicability in real-time applications.

In this work, we introduce Hi-CSPN, a novel depth completion network specifically designed to leverage operations compatible with embedded systems, making it highly suitable for real-time deployment on resource-constrained devices. Each Hi-CSPN module operates as a locally normalized spatial propagation block, ensuring that the sum of absolute propagation weights is bounded, which ensures numerical stability during iterative refinement. Additionally, multi-scale supervision at intermediate resolutions enhances convergence and reduces gradient instability. Our key contributions include: 1.) the integration of a Convolutional Spatial Propagation Network (CSPN) in a multi-scale framework to enhance depth estimation accuracy, 2.) the refinement of initial depth map estimates through cross-guided and self-guided feature extraction, utilizing operations optimized for embedded systems, and 3.) the demonstration of superior performance compared to state-of-the-art methods on the challenging indoor dataset NYUv2 [10]. By addressing both computational efficiency and accuracy, Hi-CSPN represents a practical solution for enabling high-quality depth completion for real-time applications on embedded platforms.

II. Related Work

Early works such as [11] integrate sparse geometry with image-guided learning. Subsequent methods [12] use self-supervised techniques by enabling training using monocular video and sparse LiDAR signals. Spatial propagation methods, SPN [13] and CSPN [8] refine predictions via affinity-guided propagation, improving boundary sharpness. These approaches evolved into deformable [14], non-local [7], and dynamic SPNs [15] techniques, enhancing flexibility for local structures. CSPN++ [16] introduced context-aware propagation. Further works include bilateral forms [17]. However, the iterative nature makes these methods computationally expensive and unsuitable for real-time applications.

Recent transformer-based methods for depth completion introduce long-range attention mechanisms to capture global context. Approaches like CompletionFormer [18] leverage

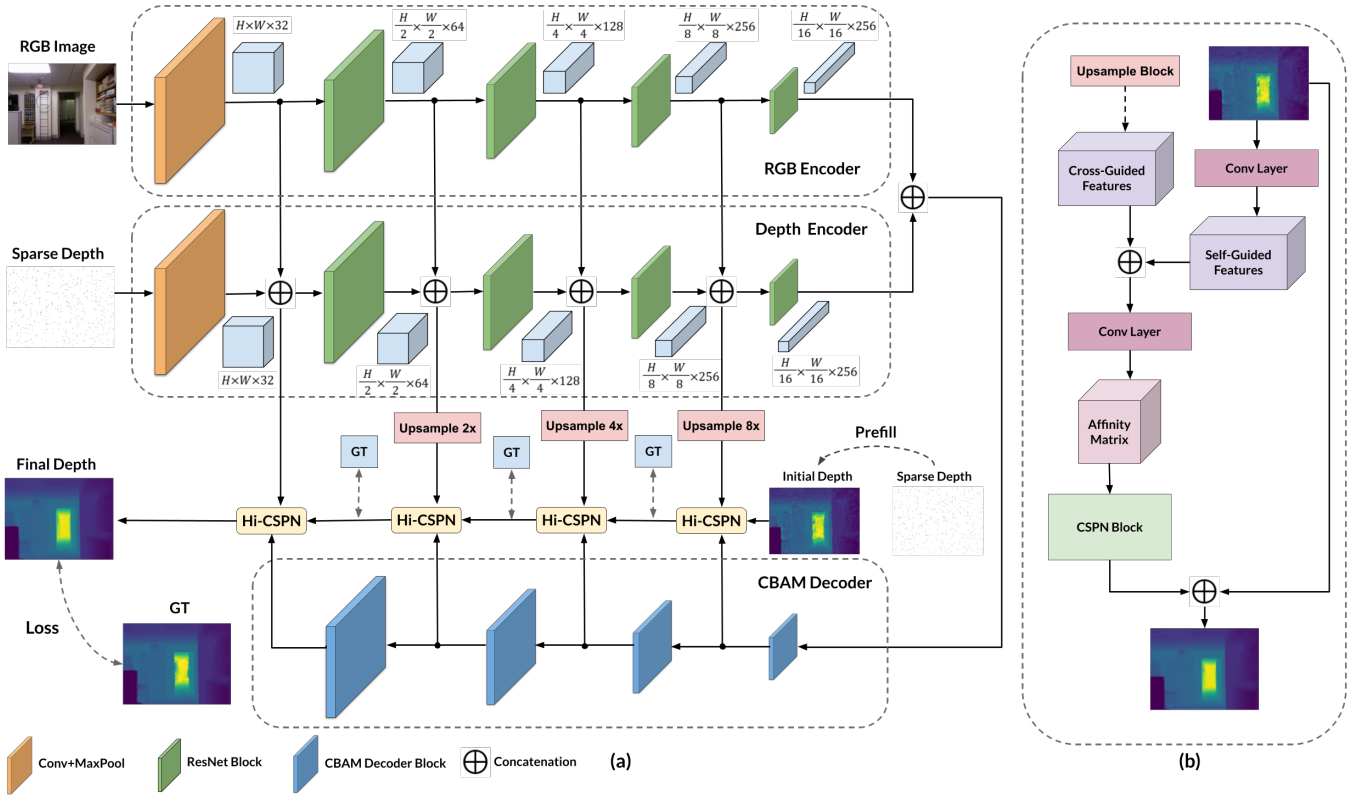

Fig. 1. Overview of the proposed Hi-CSPN architecture. (a) Dual-encoder depth completion framework with hierarchical refinement and multi-scale supervision. (b) Hi-CSPN block combines cross-guided and self-guided features to estimate affinity matrix for spatial propagation.

transformers to enhance semantic guidance and modality interaction. Gated cross-attention [19] further refined these methods by improving adaptability in cross-modal processing. Ntrans-Net [20] incorporates neutrosophic uncertainty modeling within a dual-encoder transformer architecture to improve robustness and generalization for the depth completion task. While these transformer-based techniques achieved state-of-the-art accuracy, they often come with increased computational load and memory overhead.

A recent approach, Indepth [21] targets real-time depth completion for mobile and embedded systems. LRRU [9] simplifies sparse-to-dense regression with recurrent updating. Techniques such as feature upsampling [22] enhance robustness and generalizability. Recent advancements in depth completion leverage generative models that achieve finer geometric details due to diffusion priors. Marigold [23] enhances efficiency with image generators. Other approaches [24] integrate diffusion priors with selective enhancement or distillation, underscoring their transformative impact. Despite their state-of-the-art accuracy, iterative denoising renders them unsuitable for real-time embedded applications.

III. Method

Our method takes sparse depth data and corresponding RGB images as inputs. The initial step involves densifying the input sparse depth using a fast, non-learning-based method [25]. Although the non-learning-based approach produces a coarse depth map, it reduces a significant amount of time and computation spent on directly regressing a dense depth map from a sparse depth map. Subsequently, the method refines the initial depth map through a multi-scale update unit. As illustrated in Fig. 1, the overall architecture of our approach is presented. The initial dense depth map, generated by the fast depth completion method of Ku et al. [25], which densifies sparse depth using classical image-processing operations, is denoted as $\hat{D}^0$. At each scale, the depth map undergoes refinement, resulting in an updated depth map represented as $\hat{D}^t$. The refined depth map $\hat{D}^t$ is the output from the update unit, which takes previous-stage depth map, $\hat{D}^{t-1}$, and features extracted from the dual encoder.

Cross-guided and self-guided features. At each scale, the proposed method integrates features from the RGB encoder (f_θ) and the depth encoder (f_γ), resulting in cross-guided features (F_{CG}). Concurrently, the input depth map $\hat{D}^{t-1}$ is processed through a convolutional layer (f_ω) to extract self-guided features (F_{SG}) specific to that scale. To ensure alignment with the final depth map resolution, the cross-guided features (F_{CG}) are upsampled using an upsampling layer. This architectural design closely resembles TDU [9]. Further, F_{CG} and F_{SG} are concatenated to create F_{feat} which is input to the SPN network.

Update Units. TDU [9] employed a costly non-local SPN module to refine the depth map. This design is computationally

TABLE I
QUANTITATIVE COMPARISON ON *NYUv2* DATASET.

	RMSE ↓	**REL** ↓	$\delta_{1.25}$ ↓
NLSPN [7]	0.092	0.012	99.6
LRRU [9]	0.091	0.011	99.6
Cformer [18]	0.090	0.012	99.5
ImprovingDC [22]	0.091	0.011	99.6
BP-Net [17]	0.088	0.012	99.6
DMD3C [24]	0.085	0.011	99.7
Ours with TDU	**0.086**	**0.010**	**99.6**
Ours	**0.084**	**0.010**	**99.6**

TABLE II
QUANTITATIVE COMPARISON ON *SynT* DATASET.

	RMSE ↓	**MAE** ↓	**RMAE** ↓	**EWMAE** ↓
Indepth [21]	0.158	0.065	0.027	0.070
NLSPN [7]	0.135	0.037	0.013	0.062
LRRU [9]	0.135	0.039	0.014	0.061
Cformer [18]	0.216	0.077	0.025	0.079
Ours	**0.124**	**0.034**	**0.012**	**0.057**

expensive and difficult to deploy on the device. To mitigate this, we propose a simple and computationally efficient approach. We take F_{feat} and pass it through a convolutional layer to predict the affinity matrix, $\mathcal{A} \in \mathcal{R}^{H\times W\times K\times K}$, where H and W are height and width of the target depth, respectively, and K is the kernel size for the CSPN block. A Convolutional Spatial Propagation Network (CSPN) then estimates the residual image using the estimated affinity matrix. Further, we found 4 iterations to provide the best trade-off between accuracy and latency. More precisely, CSPN in the proposed method predicts the residual $\Delta\hat{D}^{t-1}$ as below:

$$\Delta\hat{D}^{t-1}(p) = \sum_{q\in\mathcal{N}(p)} \mathcal{A}(p, p-q)\cdot\hat{D}^{t-1}(q) \tag{1}$$

where p is the pixel location and $\mathcal{N}(p)$ is $K\times K$ neighbourhood around p. Then, the depth map at this stage is updated as:

$$\hat{D}^{t} = \hat{D}^{t-1} + \Delta\hat{D}^{t-1} \tag{2}$$

Affinity weights are locally normalized such that the sum of their absolute values is bounded by one, ensuring stable propagation during iterative refinement.

Loss Function. We take the following loss at each scale level with the ground-truth depth map D_{gt}

$$\mathcal{L} = \sum_{i=1}^{3} ||D_{gt} - \hat{D}^{i}||_1 + \sum_{i=1}^{3} ||\nabla(D_{gt} - \hat{D}^{i})||_1 \tag{3}$$

Training Details. We trained our model on a single A6000 GPU for $150K$ iterations. We use the ADAM [26] optimizer with a batch-size of 16 and learning rate of 10^{-3}. Further, we use $K = 3$ for the CSPN block. Additionally, the multi-scale analysis happens at three scales i.e. $2\times, 4\times, 8\times$ as illustrated in Figure 1.

TABLE III
QUANTITATIVE COMPARISON ON *TOFDC* DATASET.

	RMSE ↓	**REL** ↓
NLSPN [7]	0.174	0.029
Cformer [18]	0.113	0.029
PointDC [28]	0.109	0.021
Rignet++ [29]	0.097	**0.016**
Ours	**0.096**	0.026

TABLE IV
ROBUSTNESS EVALUATION UNDER SALT-AND-PEPPER NOISE CORRUPTION

	RMSE ↓	**% drop** ↓
Cformer [18]	0.120	33.33
BP-Net [17]	0.101	14.7
Ours	**0.092**	**9.5**

IV. RESULTS AND EXPERIMENTS

To evaluate the effectiveness of our proposed depth completion framework, experiments are conducted on a widely used benchmark dataset as well as on an in-house synthetic dataset:

NYUv2 [10]: NYUv2 is an indoor dataset and contains RGB-D data collected with a Kinect sensor. We adopt the official split with 50K training and 654 testing samples, using the official evaluation protocol.

TOFDC [27] Dataset: TOFDC contains RGB-D data collected using a Huawei P30 Pro smartphone equipped with a ToF sensor. We use the official split of 10K training samples and 560 testing samples. This dataset aims to evaluate the performance under practical deployment conditions.

SynT Dataset: We further create a synthetic dataset comprising 150 synthetic scenes and render them using a rendering pipeline to generate photorealistic RGB images and dense ground-truth depth annotations. Each scene contains 500 rendered images. We use 134 scenes for training and 16 scenes for evaluation. This dataset includes diverse room layouts, object arrangements and lighting conditions.

Metrics: We adopt the standard depth completion evaluation metrics to evaluate the accuracy of depth predictions. Root Mean Squared Error (RMSE) measures the square root of the average of squared differences between predicted and ground truth depths. Relative Error (REL) measures how far the predicted depth values deviate from the ground truth relative to the actual depth. δ-threshold accuracy reports the percentage of predictions that lie within a certain tolerance of the ground truth. Mean Absolute Error (MAE) computes the average absolute differences. Relative Mean Absolute Error (RMAE) assesses error relative to ground truth depth, useful for scale-variant evaluation. Edge-Weighted MAE (EWMAE) places greater weight on depth errors near edges, encouraging structural consistency.

Overall Results: Quantitative results are shown on NYUv2 in Table I, TOFDC in Table III and SynT dataset in Table II. As shown in the table, competitive or superior performance is

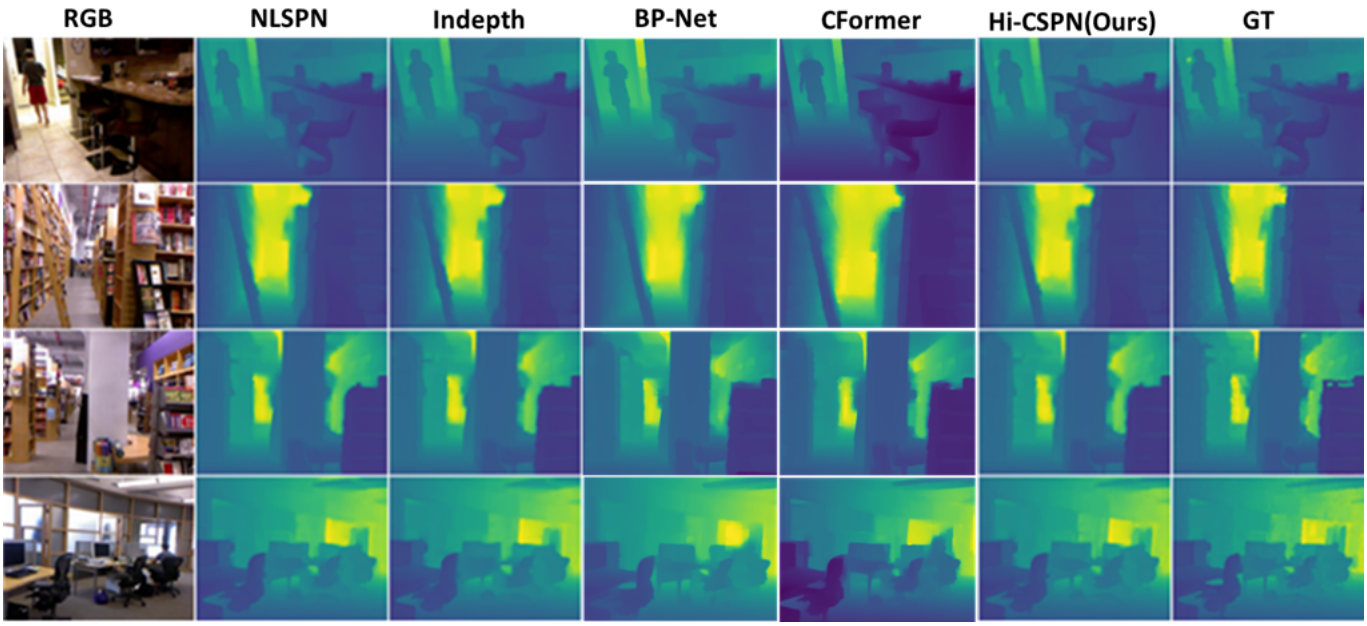

Fig. 2. Qualitative comparison on NYUv2 dataset. Our method produces sharper edges and finer details compared to the existing baselines. The proposed method preserves sharper object boundaries and finer structural details, particularly around the human silhouette in the first row.

TABLE V
ABLATION STUDIES ON TDU VS. HI-CSPN ON THE NYUV2 DATASET.

	RMSE ↓	**REL** ↓	$\delta_{1.25}$ ↓
LRRU (with TDU)	0.091	0.011	99.6
LRRU (with Hi-CSPN)	**0.084**	**0.010**	**99.6**
Our (with TDU)	0.086	0.010	99.6
Our (with Hi-CSPN)	**0.084**	**0.010**	**99.6**

TABLE VI
PEFORMANCE ON EMBEDDED DEVICE.

	Time (in s) ↓
ImprovingDC [22]	5.4
Ours	**1.5**

achieved by our method as compared to the baseline methods on all the datasets. On NYUv2, our method outperforms the baseline method LRRU [9]. Further, qualitative results are shown in Figure 2. Notice, finer details are observed near the human in the first row by our method compared to the baseline methods. Another version of our method with LRRU units is also shown: "Ours with LRRU" which achieves comparable results to the baseline methods. On TOFDC, although the best RMSE is achieved by our method, the REL metric is slightly higher than that of PointDC and Rignet++, indicating larger relative errors in a small subset of near-depth regions.

Performance on Embedded device: Our model is evaluated on the Qualcomm XR2-Gen2 processor. The model was converted using QNN-SDK and the performance on an embedded system is measured. It is observed that most of the baseline methods fail to convert due to their reliance on deformable convolutions, which are not supported by this SDK. Table VI demnostrate that our method achieves a $3.7\times$ speedup compared to the baseline method ImprovingDC on an embedded system.

Robustness to noise: To evaluate the robustness of the proposed method to degradation in visual input, we corrupt the RGB inputs with salt-and-pepper noise at probability of 0.05 during inference. Table IV shows that our method exhibit the least drop compared to the other approaches. This robustness can be attributed to the hierarchical refinement process.

Ablation Studies: To evaluate the effectiveness of the proposed Hi-CSPN refinement module, we perform an ablation study on the NYUv2 dataset by replacing the Temporal Depth Units (TDU) in LRRU [9] with our Hi-CSPN block while keeping the remaining components unchanged. This isolates the contribution of Hi-CSPN in the overall depth completion pipeline. As shown in Table V, replacing TDU with Hi-CSPN units consistently improves depth completion performance on all evaluation metrics. Specifically, the RMSE decreases by 7 mm, demonstrating that Hi-CSPN is more effective at refining coarse dense depth with guidance from RGB features. These improvements suggest that Hi-CSPN offers better edge preservation and smoother propagation compared to the original TDU design.

To further confirm the general effectiveness of Hi-CSPN, the LRRU backbone is also replaced with our proposed backbone. Similar trends are observed with Hi-CSPN achieving lower RMSE than the TDU version. This shows that the advantage of Hi-CSPN arise from the refinement mechanism itself rather than from a particular network architecture.

V. Conclusion

We present Hi-CSPN, a hierarchical convolutional spatial propagation network for depth completion that is designed for efficient deployment on embedded systems. By replacing computationally expensive non-local dynamic convolutions with fixed-window-size convolutions and employing a hierarchical refinement strategy, our approach achieves significant improvements in both accuracy and computational efficiency. Experimental results on NYUv2, TOFDC and SynT demonstrate that Hi-CSPN achieves competitive or superior performance as compared to existing methods. Furthermore, the proposed approach exhibits robustness to noisy RGB inputs. Future work will investigate replacing iterative refinement with a single-step prediction framework which will enhance speed and resource utilization for real-world applications.

References

[1] R. Du, E. Turner, M. Dzitsiuk, L. Prasso, I. Duarte, J. Dourgarian, J. Afonso, J. Pascoal, J. Gladstone, N. Cruces *et al.*, "Depthlab: Real-time 3d interaction with depth maps for mobile augmented reality," in *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, 2020, pp. 829–843.

[2] Y. Huang and Y. Chen, "Autonomous driving with deep learning: A survey of state-of-art technologies," *arXiv preprint arXiv:2006.06091*, 2020.

[3] J. Lu, T. Huang, P. Li, Z. Dou, C. Lin, Z. Cui, Z. Dong, S.-K. Yeung, W. Wang, and Y. Liu, "Align3r: Aligned monocular depth estimation for dynamic videos," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 22 820–22 830.

[4] B. Wen, M. Trepte, J. Aribido, J. Kautz, O. Gallo, and S. Birchfield, "Foundationstereo: Zero-shot stereo matching," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 5249–5260.

[5] K. Lu, N. Barnes, S. Anwar, and L. Zheng, "From depth what can you see? depth completion via auxiliary image reconstruction," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 306–11 315.

[6] X. Wu, L. Peng, H. Yang, L. Xie, C. Huang, C. Deng, H. Liu, and D. Cai, "Sparse fuse dense: Towards high quality 3d detection with depth completion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 5418–5427.

[7] J. Park, K. Joo, Z. Hu, C.-K. Liu, and I. So Kweon, "Non-local spatial propagation network for depth completion," in *European conference on computer vision*. Springer, 2020, pp. 120–136.

[8] X. Cheng, P. Wang, and R. Yang, "Depth estimation via affinity learned with convolutional spatial propagation network," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 103–119.

[9] Y. Wang, B. Li, G. Zhang, Q. Liu, T. Gao, and Y. Dai, "Lrru: Long-short range recurrent updating networks for depth completion," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 9422–9432.

[10] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *European conference on computer vision*. Springer, 2012, pp. 746–760.

[11] J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox, and A. Geiger, "Sparsity invariant cnns," in *2017 international conference on 3D Vision (3DV)*. IEEE, 2017, pp. 11–20.

[12] F. Ma, G. V. Cavalheiro, and S. Karaman, "Self-supervised sparse-to-dense: Self-supervised depth completion from lidar and monocular camera," in *2019 international conference on robotics and automation (ICRA)*. IEEE, 2019, pp. 3288–3295.

[13] S. Liu, S. De Mello, J. Gu, G. Zhong, M.-H. Yang, and J. Kautz, "Learning affinity via spatial propagation networks," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[14] Z. Xu, H. Yin, and J. Yao, "Deformable spatial propagation networks for depth completion," in *2020 ieee international conference on image processing (icip)*. IEEE, 2020, pp. 913–917.

[15] Y. Lin, T. Cheng, Q. Zhong, W. Zhou, and H. Yang, "Dynamic spatial propagation network for depth completion," in *Proceedings of the aaai conference on artificial intelligence*, vol. 36, no. 2, 2022, pp. 1638–1646.

[16] X. Cheng, P. Wang, C. Guan, and R. Yang, "Cspn++: Learning context and resource aware convolutional spatial propagation networks for depth completion," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 10 615–10 622.

[17] J. Tang, F.-P. Tian, B. An, J. Li, and P. Tan, "Bilateral propagation network for depth completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9763–9772.

[18] Y. Zhang, X. Guo, M. Poggi, Z. Zhu, G. Huang, and S. Mattoccia, "Completionformer: Depth completion with convolutions and vision transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 18 527–18 536.

[19] X. Jia, S. Jian, Y. Tan, Y. Che, W. Chen, and Z. Liang, "Gated cross-attention network for depth completion," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[20] A. Ramachandran, A. Dhiman, B. S. Vandrotti, and J. Kim, "Ntrans-net: a multi-scale neutrosophic-uncertainty guided transformer network for indoor depth completion," in *2023 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2023, pp. 905–909.

[21] Y. Zhang, T. Scargill, A. Vaishnav, G. Premsankar, M. Di Francesco, and M. Gorlatova, "Indepth: Real-time depth inpainting for mobile augmented reality," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 1, pp. 1–25, 2022.

[22] Y. Wang, G. Zhang, S. Wang, B. Li, Q. Liu, L. Hui, and Y. Dai, "Improving depth completion via depth feature upsampling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21 104–21 113.

[23] B. Ke, K. Qu, T. Wang, N. Metzger, S. Huang, B. Li, A. Obukhov, and K. Schindler, "Marigold: Affordable adaptation of diffusion-based image generators for image analysis," *arXiv preprint arXiv:2505.09358*, 2025.

[24] Y. Liang, Y. Hu, W. Shao, and Y. Fu, "Distilling monocular foundation model for fine-grained depth completion," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 22 254–22 265.

[25] J. Ku, A. Harakeh, and S. L. Waslander, "In defense of classical image processing: Fast depth completion on the cpu," in *2018 15th conference on computer and robot vision (CRV)*. IEEE, 2018, pp. 16–22.

[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[27] Z. Yan, Y. Lin, K. Wang, Y. Zheng, Y. Wang, Z. Zhang, J. Li, and J. Yang, "Tri-perspective view decomposition for geometry-aware depth completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4874–4884.

[28] Z. Yu, Z. Sheng, Z. Zhou, L. Luo, S.-Y. Cao, H. Gu, H. Zhang, and H.-L. Shen, "Aggregating feature point cloud for depth completion," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 8732–8743.

[29] Z. Yan, X. Li, L. Hui, Z. Zhang, J. Li, and J. Yang, "Rignet++: Semantic assisted repetitive image guided network for depth completion: Z. yan et al." *International Journal of Computer Vision*, vol. 133, no. 9, pp. 6051–6073, 2025.