

HFD-Teacher: High-Frequency Depth Distillation from Depth Foundation Models for Enhanced Depth Completion

Zhiyuan Yang¹, Anqi Cheng¹, Haiyue Zhu², Tianjiao Li¹, Pey Yuen Tao², Kezhi Mao¹

¹Nanyang Technological University, Singapore

²SIMTech, Agency for Science, Technology and Research (A*STAR), Singapore

{zhiyuan002, anqi002, tianjiao.li, ekzmao}@ntu.edu.sg,

{zhu.haiyue, pytao}@simtech.a-star.edu.sg

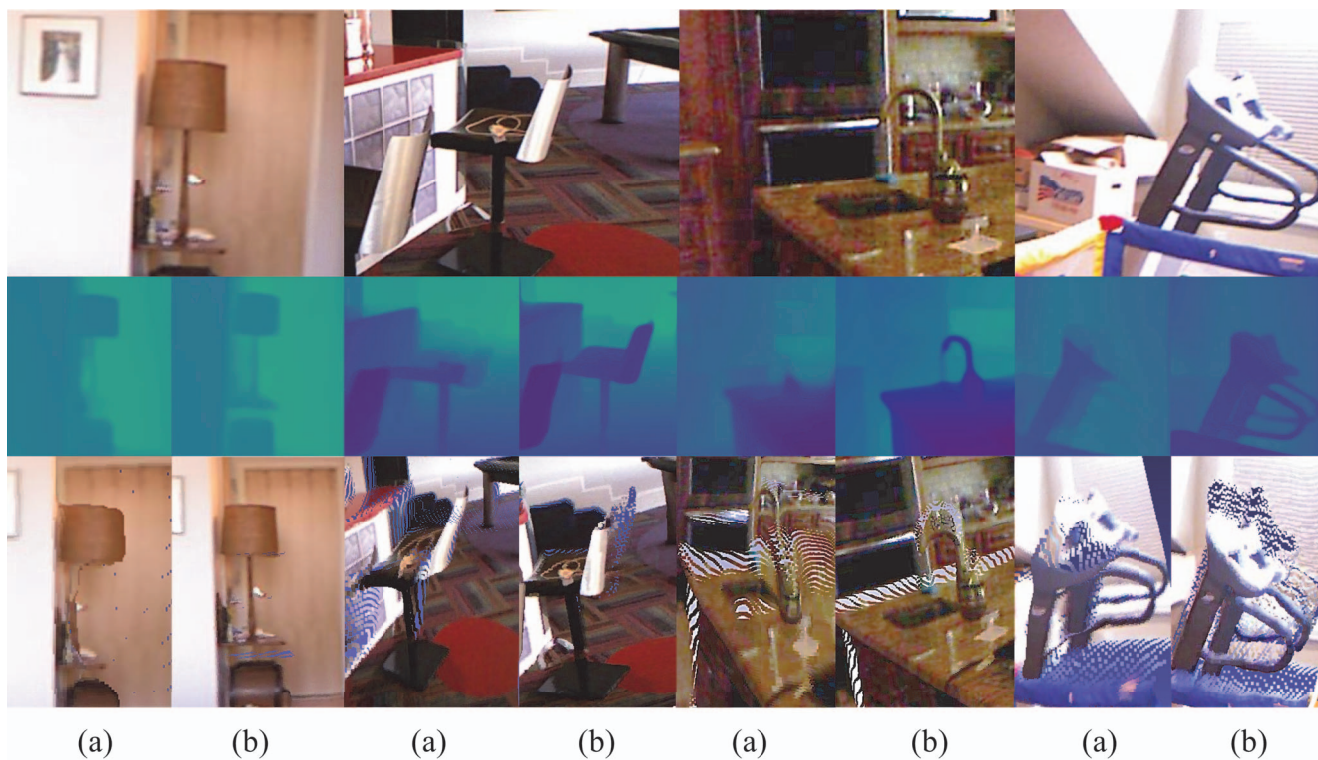


Figure 1. Comparison of our method (b) with BP-Net [32] (a), which is supervised by imperfect depth data. Our approach leverages high-frequency knowledge from depth foundation models, enabling detailed depth reconstruction even in reflective and fine-structured regions, where models trained solely on imperfect ground truth struggle. The *bottom* row displays the 3D point clouds generated from each depth map in a single view. Zoom in for better visualization.

Abstract

Depth completion, the task of reconstructing dense depth maps from sparse depth and RGB images, plays a critical role in 3D scene understanding. However, existing methods often struggle to recover high-frequency details, such as regions with fine structures or weak signals, since depth sensors may fail to capture accurate depth maps in those

regions, leading to imperfect supervision ground truth. To overcome this limitation, it is essential to introduce an alternative training source for the models. Emerging depth foundation models excel at producing high-frequency details from RGB images, yet their depth maps suffer from inconsistent scaling. Therefore, we propose a novel teacher-student framework that enhances depth completion by distilling high-frequency knowledge from depth foundation

models across multiple scales. Our approach introduces two key innovations: *Adaptive Local Wavelet Decomposition*, which dynamically adjusts wavelet decomposition level based on local complexity for efficient feature extraction, and *Topological Constraints*, which apply persistent homology to enforce structural coherence and suppress spurious depth edges. Experiment results demonstrate that our method outperforms state-of-the-art methods, preserving high-frequency details and overall depth fidelity.

1. Introduction

Depth completion, the task of generating dense depth maps from sparse depth measurements and RGB imagery, is a cornerstone of 3D scene understanding in applications such as autonomous driving, robotics, and augmented reality [8, 21, 44]. Accurate depth maps enable precise navigation [21], object detection [24], and environmental interaction [16], but achieving this accuracy remains a significant challenge. This is partially because depth sensors, such as LiDAR or structured light systems, often produce incomplete depth due to lighting conditions, sensor noise, or limited resolution, leading to depth holes and discontinuities, as illustrated in Fig. 2. These imperfections are especially pronounced in high-frequency areas, such as object boundaries, where preserving sharp depth transitions is essential for downstream tasks.

Existing depth completion methods, however, have largely overlooked the impact of training on imperfect ground truth data. By relying on noisy or incomplete depth maps as supervision, these approaches struggle to capture high-frequency details, often yielding blurred or fragmented predictions that fall short of real-world requirements. Techniques such as spatial propagation [7, 22, 26] or specialized feature extraction modules [32, 40] have been proposed to mitigate these issues, yet their effectiveness remains constrained by the quality of available high-frequency supervision during training.

In contrast, recent depth foundation models have demonstrated remarkable generalization ability in predicting high-frequency depth details from RGB imagery alone, such as Depth Anything v2 [43] and Marigold [18]. These models excel at reconstructing depth in challenging regions, including fine-structured and weak-signal areas—such as the edges of steel chairs, thin chair legs, and carpeted floors depicted in Fig. 2—where depth sensors typically falter. This underscores their potential as High-Frequency Teachers for depth completion. By leveraging their predictions, we can address missing information in high-frequency ground truth depth data, guiding the learning of detailed depth boundaries in complex environments.

Inspired by this, we present HFD-Teacher, a novel teacher-student framework designed to enhance high-



Figure 2. Existing depth completion methods, such as BP-Net [32], trained solely on noisy ground truth or simple colored depth, fail to recover high-frequency details. In contrast, depth foundation models like Depth Anything v2 [43] excel at generating detailed pseudo depths. We propose leveraging their high-frequency knowledge to enhance depth completion.

frequency details in depth completion across multiple scales in the prediction head. Our approach integrates two key innovations: (1) adaptive local wavelet transform and (2) topological constraints. Specifically, in the *adaptive local wavelet transform*, we extend traditional wavelet techniques by dynamically adjusting the decomposition level for each local region based on its complexity. This enables fine-grained feature extraction in intricate areas while maintaining efficiency in simpler regions, balancing detail preservation. And in the *topological constraints*, we enforce structural coherence in the teacher’s high-frequency depth maps using persistent homology, mitigating the generation of spurious depth edges. This synergy enhances prediction accuracy and ensures structurally coherent depth maps. Extensive experiments conducted on indoor and outdoor datasets demonstrate the superiority of our methods compared to the state-of-the-art techniques.

In summary, our contributions are threefold:

- We introduce HFD-Teacher, a teacher-student framework that selectively distills high-frequency knowledge from depth foundation models.
- We propose adaptive local wavelet transform for efficient, complexity-aware frequency analysis, and topological constraints for ensuring structural integrity in depth maps.
- We validate through extensive experiments on NYUD-v2, KITTI-DC, and other datasets that HF-Teacher surpasses state-of-the-art methods in both high-frequency detail preservation and overall depth fidelity.

2. Related Work

Depth Completion: Early methods used filters [13, 20] or optimization [9, 41] to fill holes in RGB-D data. With LiDAR’s growing use, deep learning has become dominant for sparse depth completion. CSPN [6, 7] introduced convolutional spatial propagation to refine sparse depth. NL-SPN [26] and DySPN [22] enhanced it with learnable ker-

nels for better accuracy. GuideFormer [28] employed dual-branch transformers to embed RGB and depth, while CompletionFormer [45] combined convolution and transformers. BP-Net [32] applied bilateral propagation early to improve sparse data processing. Despite these advances, existing methods overlook high-frequency loss and rely on incomplete supervision. We address this by leveraging HFD-Teacher, achieving clearer and more detailed depth predictions.

Depth Foundation Models: Recent depth foundation models [3, 18, 42, 43] predict universal depth without fine-tuning on specific datasets. Depth Anything [42, 43], trained on 1.5M labeled and 62M unlabeled images, produces dense depth with rich details from a single RGB image. Marigold [18], built on Stable Diffusion [29], achieves zero-shot performance on synthetic and cross-domain data. Most foundation models, however, target *relative* depth prediction in $[0,1]$. While adding a *metric* head [1, 2] enables metric depth, scale prediction remains unreliable. By contrast, our depth completion model leverages sparse depth as anchors, ensuring accurate metric scale in predicted maps.

Frequency Learning: Frequency analysis has been extensively applied in depth-related vision tasks, including depth estimation [4], depth super-resolution [37] and even depth completion [23]. [4] analyzes the frequency response in self-supervised depth estimation tasks to refine depth through interpretable analysis. SGNet [37] investigates the Fourier frequency response of high and low-resolution depth data to enable frequency-aware depth super-resolution. RigNet [39] also aims to enhance high-frequency components through a repetitive image-guided network. Unlike these methods, our model directly analyzes the frequency insight from depth images, as the frequency response of RGB images often introduces noise in depth-irrelevant regions such as planar textures or non-boundary lines.

3. Proposed Method

In this section, we introduce our HFD-Teacher (High-Frequency Depth Teacher) framework, as illustrated in Fig. 4 (a). Given a pair consisting of an RGB image I and a sparse depth map D_s , our goal is to train a depth completion model that enhances D_s into a dense depth map \hat{D} .

3.1. Teacher-Student Framework

We propose a teacher-student framework for depth completion that leverages frequency distillation to enhance high-frequency details in depth maps. The teacher is a pre-trained frozen depth foundation model (e.g., Depth Anything v2 [43]) that guides the student model to predict fine-detailed dense depth maps. The student model features a transformer-based backbone as the feature extractor and a multi-scale prediction head for final depth prediction. We

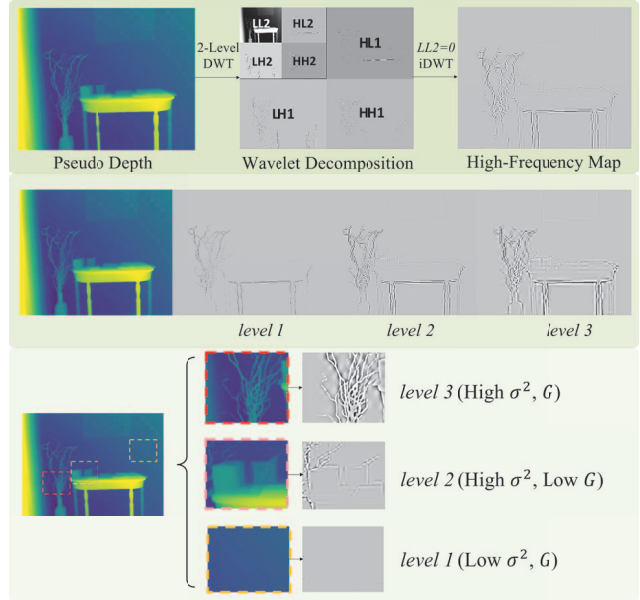


Figure 3. *Top:* A typical 2-level Discrete Wavelet Transform (DWT) applied to a pseudo depth map. By setting $LL_2 = 0$ and performing the inverse DWT, we obtain the level-2 high-frequency map. *Middle:* Deeper decomposition levels yield finer details in the DWT output. *Bottom:* As regions within a single scene vary in complexity, we determine the decomposition level in local regions by their variance (σ^2) and gradient magnitude (G).

use two separate convolution layers to encode the RGB images and sparse depth data, which are then concatenated for further processing.

Within this framework, both the teacher and student models utilize decoders to output multi-scale features, which can be regarded as depth maps layer by layer. As with prevailing prediction heads, such as DPT [27], predict multi-scale feature maps progressively from $\frac{1}{32}$, ..., $\frac{1}{2}$ to 1, each corresponding to a depth map at a specific granularity. Our method distills frequency-domain knowledge from the teacher’s multi-scale depth maps to guide the student’s predictions at each layer. To enable frequency distillation, we extract high-frequency components from these depth maps using the Adaptive Local Wavelet Transform (ALWT).

3.2. Adaptive Local Wavelet Transform

Depth foundation models excel at predicting depth maps rich in high-frequency details, but they often fall short in recovering accurate depth scale, which predominantly resides in the low-frequency components. This limitation positions them as “partial teachers”, effective primarily for high-frequency subbands rather than the full depth image. Therefore, we focus on distilling knowledge from high-frequency depth maps instead of the whole depth images. For this purpose, we utilize the Discrete Wavelet Transform (DWT) to extract frequency components. DWT preserves

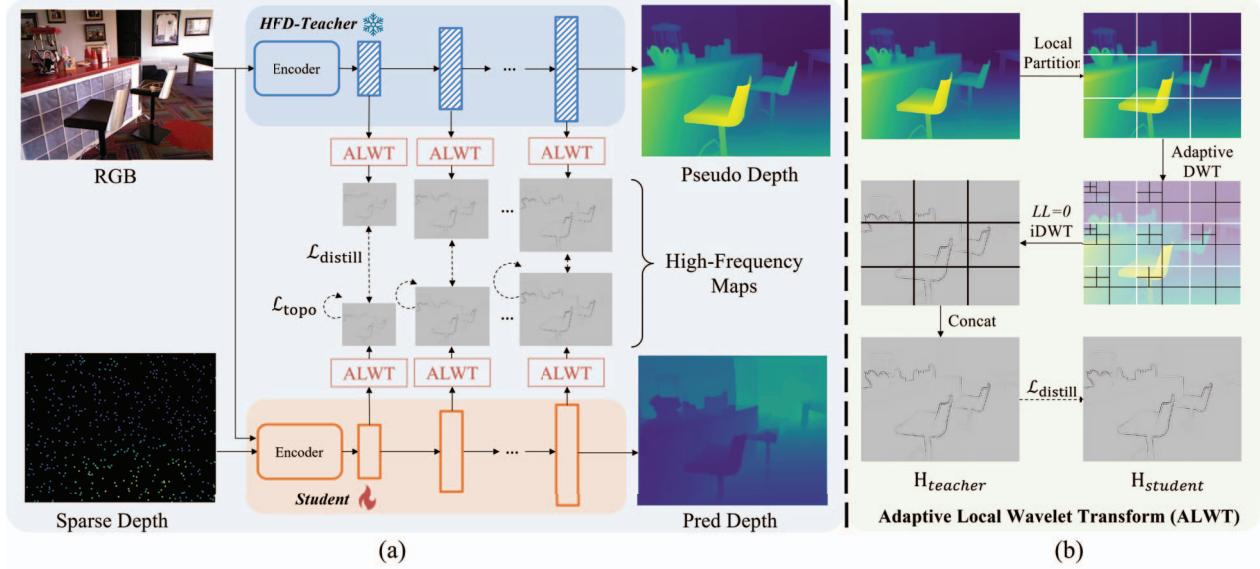


Figure 4. (a): Overview of HFD-Teacher pipeline. To compensate for the lack of high-frequency depth details, we enhance the depth completion student model using high-frequency knowledge from a depth foundation teacher model. The HFD-Teacher guides the student across multiple scales at each layer of the prediction head. (b): Adaptive Local Wavelet Transform (ALWT) module. We apply discrete wavelet transforms (DWT) of different levels within different local blocks and concatenate them to form a global high-frequency map. Distillation is then performed using the high-frequency components from both the teacher and the student models.

spatial structure by decomposing the original 2D image into four half-sized frequency bands: $\{LL, HL, LH, HH\}$. Here, $\{HL_l, LH_l, HH_l\}$ represents the high-frequency knowledge critical for our task. This decomposition can be recursively applied to the LL subband, creating a hierarchy that analyzes depth maps at varying levels of granularity, as shown in Fig. 3. In this work, we adopt the *bior3.3* wavelet, chosen for its linear phase properties that minimize edge artifacts during reconstruction and its use of separate filters for decomposition and reconstruction, enhancing flexibility.

Depth scenes in depth completion tasks often vary greatly in complexity, ranging from flat regions (e.g., walls and floors) to intricate areas crowded with objects (Fig. 3). This diversity necessitates a dynamic approach to wavelet decomposition, as deeper levels extract finer high-frequency details, while shallow levels suffice for simpler regions. However, applying a fixed decomposition level across an entire image is often inadequate for such diverse scenarios: complex regions benefit from deeper decomposition to capture high-frequency details, whereas flat areas require only shallow analysis. To address the intra-scene variability, we first partition each depth map into local blocks. For each block, we evaluate its complexity using two metrics: variance and gradient magnitude. These metrics provide insight into the depth variability and edge intensity within the block, guiding the decision on how deeply it should be decomposed. Variance quantifies the spread of depth values within a block, with higher values indicating significant depth fluctuations. Complementing this, the gra-

dent magnitude measures the strength of depth transitions (i.e., edges), where blocks with pronounced gradients are likely to contain sharp boundaries. These metrics guide the assignment of an appropriate decomposition level l_k for each block B_k with the rule:

$$l_k = \begin{cases} 3 & \text{if } \sigma_k^2 > T_\sigma \text{ and } G_k > T_G \\ 2 & \text{if } \sigma_k^2 > T_\sigma \text{ or } G_k > T_G \\ 1 & \text{otherwise} \end{cases} \quad (1)$$

Here, T_σ and T_G represent the running means of variance and gradient magnitude across all blocks in a batch. This rule assigns the deepest decomposition (level 3) to blocks with both high variance and strong gradients—hallmarks of complex, edge-heavy regions, while uniform, low-complexity blocks are efficiently processed with a shallow decomposition (level 1).

After determining l_k , we apply DWT to each block up to its assigned level, generating subbands such as LL_l (low-frequency) and $\{HL_l, LH_l, HH_l\}$ (high-frequency). To isolate high-frequency details, we set the low-frequency component to zero and reconstruct a local high-frequency-only map H_k by inverse DWT (iDWT) level by level:

$$\hat{H}_{l-1} = \text{iDWT}(0, \{HL_l, LH_l, HH_l\}) \quad (2)$$

$$H_k = \text{iDWT}(\hat{H}_1, \{HL_1, LH_1, HH_1\}) \quad (3)$$

This process discards low-frequency content, retaining only the critical high-frequency information essential for

subsequent depth distillation. These local high-frequency maps are then combined to form a global high-frequency map H for the entire depth map. However, independent block processing can introduce visible seams at boundaries. To mitigate this, we extend each block with a 2-pixel border from adjacent blocks, providing contextual overlap for smoother transitions.

After constructing the teacher’s global high-frequency map H_{teacher} from depth predictions, the student model mirrors this process to generate its own high-frequency map H_{student} . We align H_{student} with H_{teacher} , focusing on valid high-frequency pixels:

$$\mathcal{L}_{\text{distill}} = \frac{1}{M} \sum_{i,j} \mathbb{I}_{\{h_{i,j}^{\text{teacher}} > \epsilon\}} \cdot \|h_{i,j}^{\text{student}} - h_{i,j}^{\text{teacher}}\|_2 \quad (4)$$

where $h_{i,j}^{\text{teacher}}$ and $h_{i,j}^{\text{student}}$ are pixel values from H_{teacher} and H_{student} , respectively, M is the number of valid pixels where $h_{i,j}^{\text{teacher}} > \epsilon$, and ϵ is a small threshold (e.g., 10^{-6}) to exclude near-zero values.

3.3. Topological Constraints

After pixel-level distillation, we observe that high-frequency maps still contain imperfections, such as isolated fake edges, suggesting that pixel-wise distillation alone is insufficient for effective depth detail learning. Beyond pixel-level alignment, the global connectivity and persistence of depth details are also critical for accurate frequency distillation. Motivated by this, we introduce a topological constraint using persistent homology [15], which preserves stable, large-scale features while suppressing transient discontinuities like isolated edges. Specifically, we apply a regularization term to H_{student} that penalizes short-lived features in its persistence diagram:

$$\mathcal{L}_{\text{topo}} = \sum_{\substack{p \in \text{PD} \\ (d_p - b_p) < \theta}} (d_p - b_p), \quad (5)$$

where PD is the persistence diagram of H_{student} , b_p and d_p are the birth and death times of feature p , θ is a threshold for short-lived features. This topological loss encourages the student model to reduce topological noise, thereby improving the structural coherence of the distilled high-frequency map.

3.4. Training Objectives

The student model is trained using a composite loss function with three terms: Frequency Distillation Loss ($\mathcal{L}_{\text{distill}}$), Topological Loss ($\mathcal{L}_{\text{topo}}$) and Ground Truth Supervision Loss (\mathcal{L}_{gt}). Besides high-frequency distillation, we still need to ensure fidelity between the student’s final depth map D_{student} and ground truth D_{gt} , considering only valid depth

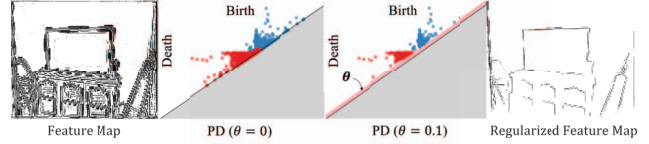


Figure 5. Visualization of high-frequency feature maps and their regularization using persistence homology. Persistence Diagram (PD) displays 0-dimensional topological invariants as red dots and 1-dimensional topological invariants as blue dots. Applying a θ to the persistence $d - b$ could reduce transient discontinuities while preserving significant features.

pixels of ground truth:

$$\mathcal{L}_{\text{gt}} = \frac{1}{N} \sum_{i,j} \mathbb{I}_{\{d_{i,j}^{\text{gt}} > 0\}} \cdot |d_{i,j}^{\text{gt}} - d_{i,j}^{\text{pred}}| \quad (6)$$

where $d_{i,j}^{\text{gt}}$ and $d_{i,j}^{\text{pred}}$ are ground truth and predicted depth values, and N is the number of valid pixels.

The total loss is:

$$\mathcal{L}_{\text{total}} = \alpha \mathcal{L}_{\text{distill}} + \beta \mathcal{L}_{\text{topo}} + \mathcal{L}_{\text{gt}} \quad (7)$$

where α, β are weighting coefficients, and are empirically set to 0.5 and 0.3, respectively.

4. Experiments

4.1. Datasets

We conduct standard depth completion evaluations on the indoor NYU Depth v2 (NYUD-v2) dataset [30] and outdoor KITTI Depth Completion (KITTI-DC) dataset [33]. We utilize the iBims-1 dataset [19] and DDAD [11] dataset to further evaluate the model’s generalization ability. NYUD-v2 is an indoor dataset with a resolution of 640×480 . For a fair comparison, we evaluate only the pixels within the crop defined in [30]. KITTI-DC is an outdoor dataset in the autonomous driving domain. We randomly crop the frames to 1216×256 for training and use the full-resolution frames as input for testing. The iBims-1 dataset comprises 100 indoor RGB-D pairs for the test set. DDAD is another autonomous driving dataset, containing 3,950 validation samples. We use iBims-1 and DDAD for the zero-shot generalization test.

4.2. Implementation Details

Our framework is implemented in Pytorch and trained using the Adam optimizer with an initial learning rate of 5×10^{-4} , $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We use a frozen pre-trained Depth Anything v2 (DAV2) ViT-L model [43] as the teacher model. For the student model, we use a pre-trained DINOv2 [25] encoder as the feature extraction backbone and DPT [27] as the prediction head. The batch size per

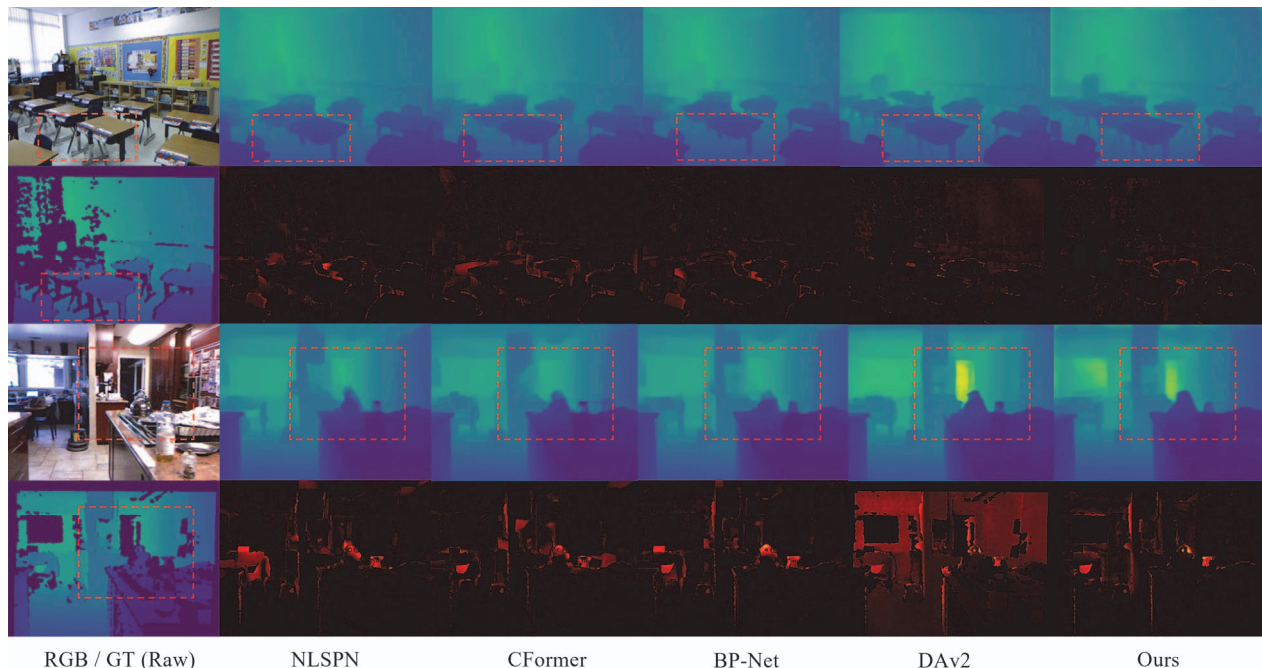


Figure 6. Qualitative results on the NYUD-v2 dataset. We compared our method with three SOTA depth completion models: NLSPN [26], CompletionFormer [45], and BP-Net [32]. The first column shows the RGB image and the raw depth. In each sample, the first row presents the depth prediction, while the second row displays the error map computed between the predicted depth and the ground truth. Areas where our method provides richer depth details are highlighted. Zoom in for better visualization.

GPU is set to 8 for NYUD-v2 and 4 for KITTI-DC respectively. The number of training epochs is set to 50. We follow the same sparse depth sampling strategy [26] for all methods.

4.3. Evaluation Metrics

We evaluate the proposed method from two perspectives: high-frequency performance and fidelity. Since high-frequency details in depth maps commonly refer to the acuity of depth edges, we assess the high-frequency performance of our method using the depth edge accuracy ε_{acc} and completion ε_{comp} metrics from [19]. ε_{acc} and ε_{comp} calculate the Chamfer distance in pixels between the predicted boundaries and the ground truth boundaries. These two metrics have been used in several works [12, 14, 38] to measure the enhanced edge quality.

To ensure the fidelity, we evaluate on the standard evaluation metrics [10]: root mean squared error (RMSE), mean absolute relative error (REL), and the accuracy metrics under threshold values $(\delta_j < 1.25^j)_{j=1,2,3}$ for indoor scenes and RMSE, MAE, root mean squared error of the inverse depth (iRMSE), and mean absolute error of the inverse depth (iMAE) for outdoor scenes.

4.4. Comparison With the State-of-the-Art

On NYUD-v2: Table 1 summarizes the quantitative comparison on NYUD-v2. Our proposed method outper-

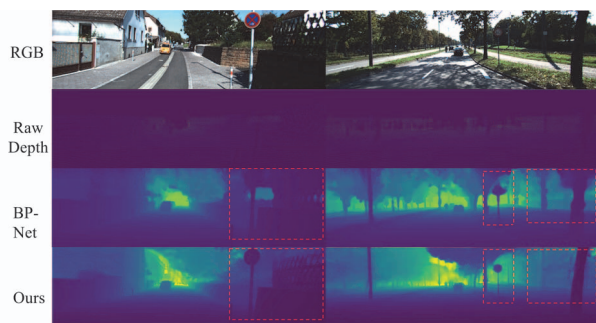


Figure 7. Qualitative Results on KITTI-DC test set. Zoom in for better visualization.

forms current methods in both high-frequency performance (boundary error) and depth fidelity (depth error and accuracy). Our method achieves the lowest ε_{acc} and ε_{comp} of 0.603 and 1.211, indicating its superiority in generating accurate high-frequency details. In terms of depth fidelity, our method surpasses most of the SOTA models with the RMSE of 0.087, which is the primary evaluation metric.

Fig. 6 provides qualitative comparisons on NYUD-v2. Our method effectively learns from the HF-Teacher and generates highly accurate dense depth maps with rich high-frequency details. For instance, in the first example, our method and DAv2 accurately predict the desk legs, while other methods struggle. In comparison of the HFD-Teacher

Methods	HF Error↓		Depth Error↓		Acc↑
	ε_{acc}	ε_{comp}	RMSE	AbsRel	δ_1
CSPN [5]	1.747	5.527	0.117	0.016	99.2
GuideNet [31]	1.255	4.608	0.101	0.015	99.5
NLSPN [26]	1.030	4.643	0.092	0.012	99.6
DySPN [22]	1.204	3.439	0.090	0.012	99.6
RigNet [39]	1.192	4.366	0.090	0.013	99.6
CFormer [45]	1.057	4.103	0.090	0.012	99.6
LRRU [35]	1.109	4.393	0.091	0.011	99.6
OGNI-DC [47]	1.130	2.330	0.089	0.011	99.6
MSPN [17]	1.020	2.041	0.089	0.012	99.6
ImprovingDC [36]	1.135	2.391	0.091	0.011	99.6
TPVD [40]	1.183	2.151	0.086	0.010	99.7
LP-Net [34]	1.041	2.931	0.090	0.012	99.6
BP-Net [32]	1.028	2.872	0.089	0.012	99.6
Ours (ViT-S)	1.014	1.457	0.107	0.016	99.5
Ours (ViT-B)	0.799	1.217	0.098	0.014	99.6
Ours (ViT-L)	0.603	1.211	0.087	0.010	99.7

Table 1. Quantitative Comparison on NYUD-v2. HF Error denotes high-frequency error, where ε_{acc} and ε_{comp} evaluate the boundary errors, reported in pixels. RMSE and AbsRel are reported in meters. The best results are in bold.

Methods	RMSE↓	MAE↓	iRMSE↓	iMAE↓
CSPN [5]	1019.64	279.46	2.93	1.15
GuideNet [31]	736.24	218.83	2.25	0.99
NLSPN [26]	741.68	199.59	1.99	0.84
DySPN [22]	709.12	192.71	1.88	0.82
RigNet [39]	712.66	203.25	2.08	0.90
CFormer [45]	708.87	203.45	2.01	0.88
BEV@DC [46]	697.44	189.44	1.83	0.82
LRRU [35]	696.51	189.96	1.87	0.81
OGNI-DC [47]	747.64	182.29	1.81	0.79
MSPN [17]	835.7	218.5	2.10	0.90
TPVD [40]	693.97	188.60	1.82	0.81
ImprovingDC [36]	686.46	187.95	1.83	0.81
LP-Net [34]	684.71	186.63	1.81	0.80
BP-Net [32]	684.90	194.69	1.82	0.84
Ours (ViT-S)	710.51	196.12	1.99	0.90
Ours (ViT-B)	689.82	194.03	1.87	0.84
Ours (ViT-L)	679.98	193.67	1.81	0.80

Table 2. Quantitative Comparison on KITTI-DC. RMSE and MAE are in millimeters, and iRMSE and iMAE are in 1/km. The best results are in bold.

DAv2, our method outperforms in the low-frequency domain, achieving a smaller overall error, as seen in the error map.

On KITTI-DC: Table 2 summarizes the quantitative comparison on KITTI-DC test set. Our ViT-L model outperforms existing SOTA methods. Notably, we only evaluate the standard metrics in the table, as the KITTI dataset has sparse ground truth, limiting us to assess high-frequency performance. Fig. 7 illustrates that our model generates richer high-frequency details, demonstrating its superiority

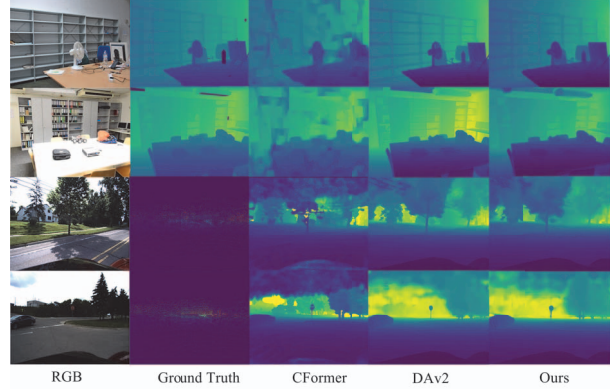


Figure 8. Qualitative results on iBims-1 (row 1-2) and DDAD (row 3-4). Zoom in for better visualization.

over other depth completion models.

Zero-shot Testing: Table 3 summarizes the quantitative comparison on the zero-shot unseen datasets iBims-1 and DDAD. All models are trained on the NYUD-v2 and KITTI-DC datasets without exposure to the test dataset. Our method outperforms the SOTA models in both high-frequency performance and depth fidelity. This demonstrates our method’s superior generalization to unseen data compared to others. Fig. 8 shows visual samples of zero-shot depth completion on these two datasets.

Methods	iBims-1					DDAD	
	$\varepsilon_{acc} \downarrow$	$\varepsilon_{comp} \downarrow$	RMSE↓	AbsRel↓	$\delta_1 \uparrow$	RMSE↓	MAE↓
NLSPN [26]	5.557	6.875	4.847	1.148	56.8	701.9	309.6
CFormer [45]	4.720	5.872	4.072	1.103	59.3	889.3	400.1
BP-Net [32]	4.755	5.037	4.792	1.057	50.1	919.3	452.9
Ours (ViT-S)	1.513	4.102	0.841	0.325	81.7	481.1	348.6
Ours (ViT-B)	1.407	3.928	0.759	0.299	82.9	359.8	227.5
Ours (ViT-L)	1.112	3.689	0.662	0.190	93.2	313.5	207.1

Table 3. Quantitative comparison of zero-shot generalization ability on unseen iBims-1 and DDAD dataset.

4.5. Ablation Study

We conducted ablation studies on the NYUD-v2 dataset with the vanilla ViT-S model as the baseline to evaluate the components of our HFD-Teacher framework, with results summarized in Table 4 to 6.

Decomposition Levels: The first set of experiments explores fixed decomposition levels in the Discrete Wavelet Transform (DWT), as shown in the upper part of Table 4 (first table). Increasing the decomposition level enhances accuracy by reducing error metrics, reflecting better preservation of fine details. However, this improvement comes with a noticeable rise in computational time. We can observe from Fig. 10(a) that $l = 3$ reveals a clear trade-off between precision and efficiency.

Dynamic levels: The lower part of Table 4 evaluates our dynamic decomposition approach, which adjusts levels based on scene complexity. Compared to fixed levels, this

Methods	$\varepsilon_{acc} \downarrow$	$\varepsilon_{comp} \downarrow$	RMSE \downarrow	AbsRel \downarrow	$\delta_1 \uparrow$	Process Time \downarrow
Baseline	1.622	2.185	0.126	0.035	99.3	0.017
fixed $l = 1$	1.572	2.157	0.123	0.023	99.3	0.034
fixed $l = 2$	1.542	2.014	0.118	0.023	99.3	0.047
fixed $l = 3$	1.501	2.010	0.117	0.023	99.4	0.059
fixed $l = 4$	1.495	2.034	0.117	0.023	99.4	0.074
fixed $l = 5$	1.510	2.052	0.114	0.024	99.3	0.090
dynamic $l = 1, 2$	1.400	1.921	0.109	0.019	99.3	0.090
dynamic $l = 1, 2, 3$	1.385	1.903	0.108	0.018	99.5	0.102
dynamic $l = 2, 3$	1.445	1.975	0.112	0.022	99.3	0.098

Table 4. Ablation studies on fixed and dynamic wavelet decomposition levels. Experiments are conducted on baseline (ViT-S without any modules) on NYUD-v2.

strategy improves both accuracy and completeness metrics. We evaluated on several decomposition level combinations and the trend Fig. 10(b) highlights $level = 1, 2, 3$ optimizes detail extraction efficiently by adapting to local variations. We do not include combinations with $l > 3$ as decomposition greater than level 3 could significantly increase the batch process time.

Local Partitioning: We assess the effectiveness of local decomposition compared to global decomposition. As demonstrated in Table 5, local frequency analysis yields superior performance over global analysis. Additionally, we examine the effect of window size in local partitioning, with results detailed in Table 5. Our analysis shows that a moderate window size of 14×14 achieves optimal performance. Smaller windows overly concentrate on local details, whereas larger windows diminish the advantages of localized analysis, underscoring the importance of selecting an appropriate scale.

Methods	$\varepsilon_{acc} \downarrow$	$\varepsilon_{comp} \downarrow$	RMSE \downarrow	AbsRel \downarrow	$\delta_1 \uparrow$
w.o. partition	1.385	1.903	0.108	0.018	99.5
8×8	1.300	1.820	0.110	0.017	99.5
14×14	1.285	1.805	0.109	0.017	99.5
16×16	1.315	1.835	0.110	0.018	99.5
24×24	1.330	1.850	0.111	0.018	99.5

Table 5. Ablation studies on the effectiveness of local partitioning and local window sizes.

Topological Constraints: Table 6 assesses the effect of topological constraints via persistent homology. Applying these constraints with a suitable threshold reduces discontinuities. Performance peaks at an optimal $\theta = 0.1$, beyond which additional regularization yields diminishing benefits, underscoring its role in enhancing structural coherence.

5. Conclusion

In this paper, we introduce HFD-Teacher, a novel teacher-student framework that elevates depth completion by distilling high-frequency knowledge from depth foundation models across multiple scales. By integrating adaptive local wavelet decomposition for efficient, complexity-aware

Methods	$\varepsilon_{acc} \downarrow$	$\varepsilon_{comp} \downarrow$	RMSE \downarrow	AbsRel \downarrow	$\delta_1 \uparrow$
w.o. \mathcal{L}_{topo}	1.285	1.805	0.109	0.017	99.5
$\theta = 0.01$	1.240	1.760	0.108	0.017	99.5
$\theta = 0.05$	1.145	1.665	0.109	0.017	99.5
$\theta = 0.1$	1.014	1.457	0.107	0.016	99.5
$\theta = 0.15$	1.190	1.710	0.110	0.017	99.4
$\theta = 0.2$	1.280	1.800	0.113	0.017	99.4

Table 6. Ablation study on the influence of topological constraints and optimal θ values on performance.

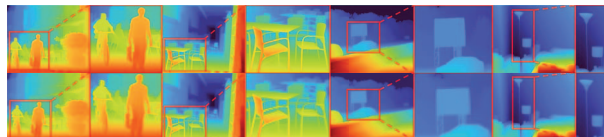


Figure 9. Top: Results with topological constraints applied. Bottom: Results without topological constraints, exhibiting spurious edges in the depth map.

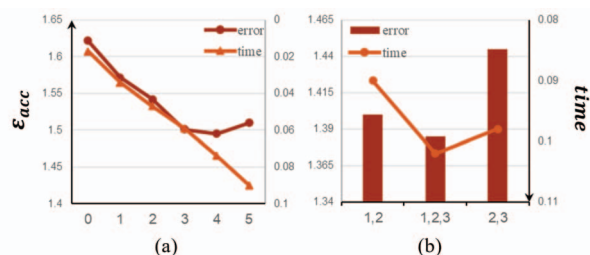


Figure 10. Trade-off between performance and efficiency. The primary vertical axis represents ε_{acc} , while the secondary vertical axis represents batch processing time. (a) For fixed decomposition level, $l = 3$ achieves the optimal balance between error and computational time. (b) Among dynamic decomposition sets, the combination of $l = 1, 2, 3$ yields the most optimal results.

feature extraction and topological constraints to enforce structural coherence, our method effectively addresses the challenges of recovering fine-grained details from imperfect ground truth data. Extensive experiments on diverse datasets, including NYUD-v2, KITTI-DC, iBims-1, and DDAD, demonstrate superior performance in high-frequency detail preservation, depth fidelity, and zero-shot generalization compared to state-of-the-art approaches.

Acknowledgements: This research is supported by A*STAR under its "RIE2025 IAF-PP Advanced ROS2-native Platform Technologies for Cross sectorial Robotics Adoption (M21K1a0104)" programme.

References

- [1] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Localbins: Improving depth estimation by learning local distributions. In *European Conference on Computer Vision*, pages 480–496. Springer, 2022. 3

- [2] Shariq Farooq Bhat, Reiner Birkel, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 3
- [3] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. *arXiv preprint arXiv:2410.02073*, 2024. 3
- [4] Xingyu Chen, Thomas H Li, Ruonan Zhang, and Ge Li. Frequency-aware self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5808–5817, 2023. 3
- [5] Xinjing Cheng, Peng Wang, and Ruigang Yang. Depth estimation via affinity learned with convolutional spatial propagation network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 103–119, 2018. 7
- [6] Xinjing Cheng, Peng Wang, and Ruigang Yang. Learning depth with convolutional spatial propagation network. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2361–2379, 2019. 2
- [7] Xinjing Cheng, Peng Wang, Chenye Guan, and Ruigang Yang. Cspn++: Learning context and resource aware convolutional spatial propagation networks for depth completion. In *Proceedings of the AAAI conference on artificial intelligence*, pages 10615–10622, 2020. 2
- [8] Catherine Diaz, Michael Walker, Danielle Albers Szafer, and Daniel Szafer. Designing for depth perceptions in augmented reality. In *2017 IEEE international symposium on mixed and augmented reality (ISMAR)*, pages 111–122. IEEE, 2017. 2
- [9] James Diebel and Sebastian Thrun. An application of markov random fields to range sensing. *Advances in neural information processing systems*, 18, 2005. 2
- [10] David Eigen, Christian Puhusch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014. 6
- [11] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2485–2494, 2020. 5
- [12] Praful Hambarde, Gourav Wadhwa, Santosh Kumar Vipparthi, Subrahmanyam Murala, and Abhinav Dhall. Occlusion boundary prediction and transformer based depth-map refinement from single image. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024. 6
- [13] Kaiming He, Jian Sun, and Xiaoou Tang. Guided image filtering. *IEEE transactions on pattern analysis and machine intelligence*, 35(6):1397–1409, 2012. 2
- [14] Junjie Hu, Mete Ozay, Yan Zhang, and Takayuki Okatani. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. In *2019 IEEE winter conference on applications of computer vision (WACV)*, pages 1043–1051. IEEE, 2019. 6
- [15] Xiaoling Hu, Fuxin Li, Dimitris Samaras, and Chao Chen. Topology-preserving deep image segmentation. *Advances in neural information processing systems*, 32, 2019. 5
- [16] Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Dustin Freeman, Andrew Davison, et al. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 559–568, 2011. 2
- [17] Jinyoung Jun, Jae-Han Lee, and Chang-Su Kim. Masked spatial propagation network for sparsity-adaptive depth refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19768–19778, 2024. 7
- [18] 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
- [19] Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Korner. Evaluation of cnn-based single-image depth estimation methods. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 5, 6
- [20] Johannes Kopf, Michael F Cohen, Dani Lischinski, and Matt Uyttendaele. Joint bilateral upsampling. *ACM Transactions on Graphics (ToG)*, 26(3):96–es, 2007. 2
- [21] Wei Li, CW Pan, Rong Zhang, JP Ren, YX Ma, Jin Fang, FL Yan, QC Geng, XY Huang, HJ Gong, et al. Aads: Augmented autonomous driving simulation using data-driven algorithms. *Science robotics*, 4(28):eaaw0863, 2019. 2
- [22] 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. 2, 7
- [23] Qiang Liu, Haosong Yue, Zhanggang Lyu, Wei Wang, Zhong Liu, and Weihai Chen. Sehl-net: separate estimation of high- and low-frequency components for depth completion. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 668–674. IEEE, 2022. 3
- [24] Tanguy Ophoff, Kristof Van Beeck, and Toon Goedemé. Exploring rgb+ depth fusion for real-time object detection. *Sensors*, 19(4):866, 2019. 2
- [25] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 5
- [26] 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. 2, 6, 7
- [27] 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, 5
- [28] Kyeongha Rho, Jinsung Ha, and Youngjung Kim. Guideformer: Transformers for image guided depth completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6250–6259, 2022. 3
- [29] 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. 3
- [30] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgb-d images. In *Computer Vision—ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7–13, 2012, Proceedings, Part V 12*, pages 746–760. Springer, 2012. 5
- [31] Jie Tang, Fei-Peng Tian, Wei Feng, Jian Li, and Ping Tan. Learning guided convolutional network for depth completion. *IEEE Transactions on Image Processing*, 30:1116–1129, 2020. 7
- [32] Jie Tang, Fei-Peng Tian, Boshi An, Jian Li, and Ping Tan. Bilateral propagation network for depth completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9763–9772, 2024. 1, 2, 3, 6, 7
- [33] 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. 5
- [34] Kun Wang, Zhiqiang Yan, Junkai Fan, Jun Li, and Jian Yang. Learning inverse laplacian pyramid for progressive depth completion. *arXiv preprint arXiv:2502.07289*, 2025. 7
- [35] Yufei Wang, Bo Li, Ge Zhang, Qi Liu, Tao Gao, and Yuchao Dai. Lrru: Long-short range recurrent updating networks for depth completion. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9422–9432, 2023. 7
- [36] Yufei Wang, Ge Zhang, Shaoqian Wang, Bo Li, Qi Liu, Le Hui, and Yuchao Dai. Improving depth completion via depth feature upsampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21104–21113, 2024. 7
- [37] Zhengxue Wang, Zhiqiang Yan, and Jian Yang. Sgnet: Structure guided network via gradient-frequency awareness for depth map super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5823–5831, 2024. 3
- [38] Feng Xue, Junfeng Cao, Yu Zhou, Fei Sheng, Yankai Wang, and Anlong Ming. Boundary-induced and scene-aggregated network for monocular depth prediction. *Pattern Recognition*, 115:107901, 2021. 6
- [39] Zhiqiang Yan, Kun Wang, Xiang Li, Zhenyu Zhang, Jun Li, and Jian Yang. Rignet: Repetitive image guided network for depth completion. In *European Conference on Computer Vision*, pages 214–230. Springer, 2022. 3, 7
- [40] Zhiqiang Yan, Yuankai Lin, Kun Wang, Yupeng Zheng, Yufei Wang, Zhenyu Zhang, Jun Li, and Jian Yang. Tri-perspective view decomposition for geometry-aware depth completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4874–4884, 2024. 2, 7
- [41] Jingyu Yang, Xinchun Ye, Kun Li, Chunping Hou, and Yao Wang. Color-guided depth recovery from rgb-d data using an adaptive autoregressive model. *IEEE transactions on image processing*, 23(8):3443–3458, 2014. 2
- [42] 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
- [43] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2, 2024. 2, 3, 5
- [44] Ekim Yurtsever, Jacob Lambert, Alexander Carballo, and Kazuya Takeda. A survey of autonomous driving: Common practices and emerging technologies. *IEEE access*, 8:58443–58469, 2020. 2
- [45] Youmin Zhang, Xianda Guo, Matteo Poggi, Zheng Zhu, Guan Huang, and Stefano Mattoccia. Completionformer: Depth completion with convolutions and vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18527–18536, 2023. 3, 6, 7
- [46] Wending Zhou, Xu Yan, Yinghong Liao, Yuankai Lin, Jin Huang, Gangming Zhao, Shuguang Cui, and Zhen Li. Bev@dc: Bird’s-eye view assisted training for depth completion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9233–9242, 2023. 7
- [47] Yiming Zuo and Jia Deng. Ogni-dc: Robust depth completion with optimization-guided neural iterations. In *European Conference on Computer Vision*, pages 78–95. Springer, 2024. 7