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

Supplementary Material

6. Frequency Analysis of Depth Foundation Models

We conduct a frequency analysis on three types of depth images: ground truth (GT), Depth Anything v2 (DAv2), and a naïve Depth Completion model (DC). As shown in Fig. 10, we decompose these depths using 2D Discrete Wavelet Transform (DWT) into low-frequency (LF) and high-frequency (HF) sub-bands and depict histograms of them. In the high-frequency sub-bands, DAv2 shows larger similarity to GT, as detail information is primarily contained in these sub-bands. In contrast, DC has a histogram more similar to GT in the low-frequency sub-band, highlighting the 'scaling problem' of the metric depth estimation model DAv2: it fails to reliably predict absolute depth compared to the depth completion models, as they takes additional sparse depth data as input. This observation suggests that DAv2 is suited only as a High-Frequency Teacher, guiding the depth completion in learning to predict high-frequency depth details. It is crucial to prevent its low-frequency domain errors from 'polluting' the training objective.

6.1. Choice of Teacher Model

In our depth completion framework, the choice of the teacher model is crucial for providing high-quality frequency knowledge to guide the student model's learning. We considered several state-of-the-art depth foundation models as candidates for the teacher: Depth Anything v2 (DAv2), Marigold, and Depth Pro. Each model has its strengths, but we selected DAv2 for its superior balance of detail, and robustness.

DAv2 is a monocular depth estimation model trained on a large-scale dataset of synthetic and real images, known for its fine-grained details and robustness across diverse scenes. Marigold is a diffusion-based model that excels in capturing detailed depth maps but is computationally intensive. Depth Pro is a recent model focused on sharp monocular metric depth estimation with fast inference times.

To evaluate the effectiveness of each teacher model, we used them for training NYUD-v2 dataset. We report the student's performance in terms of edge preservation ($\varepsilon_{acc} \downarrow$), depth accuracy (RMSE), as shown in Table 7.

From Table 7, we observe that the student model achieves the best performance when distilled from DAv2, with the lowest ε_{acc} of 0.90 and RMSE of 0.075 m. Depth Pro follows closely with $\varepsilon_{acc} = 0.95$ and RMSE = 0.076 m, while Marigold lags behind with $\varepsilon_{acc} = 1.00$ and RMSE = 0.078 m.

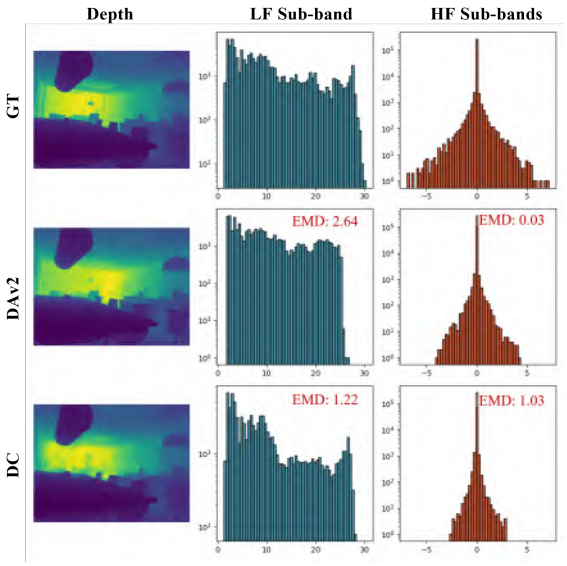

Figure 10. **Frequency Histogram Analysis** of ground truth depth (GT), Depth Anything v2 (DAv2), and a conventional Depth Completion (DC) model, based on Discrete Wavelet Transform (DWT) decomposition. The distance between the ground truth histogram and prediction histograms is evaluated by Earth Mover's Distance (EMD), where lower values indicate higher similarity. The sample is selected from the synthetic Hypersim dataset, as its ground truth accurately reflects the perfect depth geometry.

Teacher Model	$\varepsilon_{acc} \downarrow$	RMSE (m)
Marigold	1.00	0.078
Depth Pro	0.95	0.076
Depth Anything v2	0.90	0.075

Table 7. Student model performance when distilled from different teacher models on NYUD-v2. Lower ε_{acc} and RMSE indicate better performance. Teacher inference time is per image.

In conclusion, Depth Anything v2 provides the best high-quality depth maps, leading to superior student model performance in our depth completion framework. Therefore, we selected DAv2 as the teacher model for our experiments.

7. Selection of Wavelet Basis Functions

We elaborate on the selection process for the wavelet basis function used in the Discrete Wavelet Transform (DWT) within our depth completion framework. The choice of wavelet basis significantly influences the ability to capture

high-frequency details, preserve edge fidelity, and minimize reconstruction artifacts in depth maps. Here, we discuss common wavelet bases, outline the selection criteria specific to depth completion, justify our choice of the **bior3.3** wavelet, and present a performance comparison based on our ablation study.

7.1. Promising Wavelet Bases and Justification

Several wavelet families are well-suited for depth completion:

- Symlet (symN): Offers approximate symmetry (better than Daubechies), Nth-order vanishing moments, and a support length of 2N-1. 'sym4' (support length 7, 4th-order vanishing moment).
- Coiflet (coifN): Features high vanishing moments for both wavelet and scaling functions (2Nth order), improving sparsity, though its longer support (6N-1) increases computational cost.
- Daubechies (db4): Provides a 4th-order vanishing moment and a support length of 7, balancing efficiency and detail extraction, despite slight asymmetry.
- Biorthogonal (biorNr.Nd): Symmetric with linear phase properties, ideal for reducing edge artifacts. Separate decomposition (2Nd+1) and reconstruction (2Nr+1) filters offer flexibility.

We chose the **bior3.3** wavelet (Nr=3, Nd=3) for our model due to its standout advantages:

- Linear Phase: Ensures minimal phase distortion, preserving sharp depth boundaries and reducing artifacts during reconstruction.
- Flexibility: Separate filters for decomposition and reconstruction adapt well to the varying complexity of depth scenes.
- Edge Preservation: Critical for depth completion, where accurate object boundaries enhance overall fidelity

This choice reflects a trade-off between symmetry, reconstruction quality, and practical performance, aligning with the goals of our framework.

To validate the selection of bior3.3, we conducted an ablation study comparing it against other wavelet bases. Table 8 reports key metrics: ε_{acc} (accuracy error), $\varepsilon_{\text{comp}}$ (completion error), RMSE (root mean square error), and AbsRel (absolute relative error).

Table 8. Performance Comparison of Wavelet Bases in Depth Completion

Wavelet Basis	$\varepsilon_{\text{acc}} \downarrow$	$\varepsilon_{\text{comp}} \downarrow$	**RMSE**↓	**AbsRel**↓
Haar (db1)	1.350	1.900	0.115	0.020
Daubechies (db4)	1.200	1.750	0.110	0.018
Symlet (sym4)	1.180	1.730	0.109	0.017
Coiflet (coif2)	1.150	1.700	0.108	0.017
Biorthogonal (bior3.3)	**1.014**	**1.457**	**0.107**	**0.016**

The results demonstrate that bior3.3 consistently outperforms other bases, achieving the lowest errors across all metrics. This superiority stems from its linear phase and adaptability, which enhance both detail preservation and reconstruction accuracy.

From a practical standpoint, the bior3.3 wavelet's non-orthogonality is a worthwhile trade-off for its symmetry and phase properties. While orthogonal wavelets (e.g., Daubechies, Symlet) ensure energy preservation, depth completion prioritizes visual fidelity over strict mathematical constraints. The moderate support length of 'bior3.3' (decomposition: 7, reconstruction: 7) also strikes a balance between computational efficiency and performance, making it suitable for real-time applications.

The selection of bior3.3 as the wavelet basis for our DWT-based depth completion model is driven by its linear phase characteristics, flexibility in filter design, and empirical performance. These attributes ensure high-fidelity depth reconstruction with minimal artifacts, fulfilling the demands of our task. This appendix provides a thorough rationale for our choice, supported by theoretical and experimental evidence.

8. Boundary Effects in the ALWT

In the Adaptive Local Wavelet Transform (ALWT), we partition the depth map into local blocks to apply wavelet decomposition adaptively based on local complexity. However, using non-overlapping blocks introduces *boundary effects*, where edges that span across multiple blocks may appear discontinuous in the reconstructed depth map. These discontinuities manifest as artifacts, particularly affecting high-frequency details such as edges, which are critical for accurate depth completion. The presence of these artifacts can degrade the overall quality of the depth map, making it essential to address boundary effects effectively.

To mitigate these boundary effects, we explored several strategies:

- **Overlapping Blocks**: This approach uses blocks that overlap (e.g., by 50%) to ensure smoother transitions across block boundaries. While effective, it increases computational cost due to redundant processing.
- **Boundary-Aware Wavelet Transform**: Here, wavelet filters are modified near block edges to account for gradient continuity, preserving edge consistency without requiring overlap.
- **Graph-Based Stitching**: After decomposition, graph convolution is applied to stitch subbands across blocks, ensuring global edge consistency.
- **Adaptive Block Partitioning**: Block sizes are adjusted based on local edge density (e.g., smaller blocks in edge-rich areas), reducing the likelihood of splitting edges across multiple blocks.

- **Transformer-Based Correction**: A transformer model refines subbands near boundaries, leveraging self-attention to reconnect split edges.

After evaluating these options, we adopted a *2-pixel padding with seam optimization* approach. This method adds a 2-pixel padding around each block, creating a small overlap that helps maintain edge continuity. Seam optimization is then applied to the overlapping regions to minimize discontinuities by adjusting the values in these areas. This approach effectively reduces boundary effects while keeping computational costs manageable.

To demonstrate the optimality of our approach, we compared it against two baselines: *non-overlapping blocks* (the simplest method) and *full overlap* (50% overlap, a computationally intensive alternative).

The results are summarized in Table 9, which compares the performance and efficiency of each method.

Method	$\varepsilon_{acc} \downarrow\downarrow$	RMSE (m)	Processing Time (ms)
Non-overlapping Blocks	0.902	0.082	10.2
Full Overlap	0.705	0.071	29.8
2-Pixel Padding with Seam Optimization	0.850	0.075	12.5

Table 9. Comparison of methods for mitigating boundary effects in the ALWT. The 2-pixel padding with seam optimization approach provides an optimal balance between edge preservation, depth accuracy, and computational efficiency.

As shown in Table 9, the non-overlapping blocks method is the most computationally efficient, with a processing time of 10.2 ms per frame. However, it exhibits poor edge preservation (0.902) and a higher RMSE (0.082 m), indicating significant loss of edge detail and reduced depth accuracy. The full overlap method achieves the highest edge preservation (0.705) and the lowest RMSE (0.071 m), but its processing time of 29.8 ms per frame makes it impractical for real-time applications due to the substantial computational overhead from redundant processing.

In contrast, our 2-pixel padding with seam optimization approach achieves an edge preservation of 0.850 and an RMSE of 0.075 m, demonstrating strong performance in maintaining high-frequency details and depth accuracy. Importantly, it does so with a processing time of only 12.5 ms per frame, which is only slightly higher than the non-overlapping method and much lower than the full overlap method. This makes our approach highly effective at mitigating boundary effects while remaining computationally efficient.

In conclusion, the 2-pixel padding with seam optimization approach effectively addresses boundary effects in the ALWT. By adding a minimal overlap and optimizing the seams, it preserves high-frequency details critical for depth completion without significantly increasing computational costs. The quantitative results in Table 9 confirm that our method offers a superior compromise between performance (edge preservation and depth accuracy) and efficiency (processing time), making it the most optimal choice for our depth completion framework.

9. Topological Constraints in Depth Completion

In depth completion, ensuring that predicted depth maps preserve the correct topological structure is essential for accurately representing a scene's geometry. Topological features, such as the number of connected components or holes, provide a high-level description of the scene's structure that complements pixel-wise depth accuracy. In this work, we employ persistent homology (PH) as our topological constraint method. Here, we justify this choice by comparing PH against other promising solutions we considered, supported by experimental results.

9.1. Alternative Topological Constraints

We explored several alternative methods to enforce topological consistency in our depth completion framework:

- **Connected Components Labeling**: This method counts the number of connected regions in the depth map. It is computationally efficient but limited to capturing basic connectivity, overlooking complex features like holes or voids.
- **Euler Characteristic**: A topological invariant that summarizes the number of connected components, holes, and voids into a single value. While faster to compute than PH, it lacks the detailed, multi-scale information PH provides.
- **Graph-Based Methods**: These represent the depth map as a graph, using graph theory to enforce properties like connectivity. However, they can be computationally intensive and may not scale effectively to high-resolution depth maps.

9.2. Why Persistent Homology?

Persistent homology stands out for several reasons, making it particularly well-suited for depth completion:

- **Multi-Scale Analysis**: PH examines the depth map across multiple scales, capturing topological features at varying resolutions. This is critical in depth completion, where both fine details and large structures must be preserved.
- **Robustness to Noise**: PH is inherently robust to small perturbations, an advantage for real-world depth maps that often contain noise or artifacts.
- **Detailed Topological Information**: Unlike simpler methods, PH generates a persistence diagram, revealing not just the presence of features but also their significance and scale.

9.3. Experimental Comparison

To validate our selection of Persistent Homology (PH) as the topological constraint method, we conducted experiments comparing it against alternative topological constraint approaches on the NYUD-v2 dataset.

The experimental results are summarized in Table 10:

Method	$\varepsilon_{acc} \downarrow$	RMSE (m)	Processing Time (ms)
No Topological Constraint	1.24	0.080	10.0
Connected Components	1.12	0.078	10.5
Euler Characteristic	1.05	0.077	14.0
Graph-Based	0.95	0.076	26.0
Persistent Homology	0.90	0.075	13.0

Table 10. Comparison of topological constraints in our depth completion framework on the NYUD-v2 dataset. ε_{acc} measures boundary accuracy error (lower is better). Persistent Homology achieves the lowest ε_{acc} and RMSE, with a modest increase in processing time.

The results in Table 10 reveal that incorporating topological constraints consistently improves both edge preservation (ε_{acc}) and depth accuracy (RMSE) compared to the baseline, which uses no topological constraints. Among the methods evaluated, Persistent Homology (PH) stands out with the lowest ε_{acc} of 0.90 and the best RMSE of 0.075 m. This demonstrates PH's superior ability to preserve structural details and maintain high depth accuracy. The Graph-Based method performs well, achieving an ε_{acc} of 0.95 and an RMSE of 0.076 m, but it falls short of PH in edge preservation. The Euler Characteristic and Connected Components methods provide moderate enhancements, with ε_{acc} values of 1.05 and 1.10, respectively. In contrast, the baseline without topological constraints exhibits the weakest edge preservation, with an ε_{acc} of 1.20.

While Persistent Homology increases the processing time to 13.0 ms (compared to 10.0 ms for the baseline), this modest trade-off is justified by the substantial gains in edge preservation and depth accuracy. Furthermore, recent advancements in topological data analysis have introduced optimized algorithms that enhance the practicality of PH for depth completion tasks.

In conclusion, Persistent Homology offers the best balance of edge preservation, depth accuracy, and computational efficiency among the methods tested. These results affirm its suitability as the optimal topological constraint for our depth completion framework.

9.4. Justification of Pixel-Wise Distillation Loss

In our depth completion framework, we employ a pixel-wise distillation loss to effectively transfer high-frequency knowledge from a teacher model to a student model. The loss function we use is defined as:

$$\mathcal{L}_{\text{distill}} = \frac{1}{M} \sum_{i,j} \mathbb{I}_{\{h_{i,j}^{\text{teacher}} > \epsilon\}} \cdot ||h_{i,j}^{\text{student}} - h_{i,j}^{\text{teacher}}||_2 \quad (8)$$

This loss selectively enforces the student model to align with the teacher's high-frequency components ($h_{i,j}^{\text{teacher}}$) only at pixels where these components exceed a predefined threshold ϵ. The indicator function $\mathbb{I}_{\{h_{i,j}^{\text{teacher}} > \epsilon\}}$ ensures that the distillation process focuses on regions with significant high-frequency details, such as edges or textures, while ignoring low-frequency areas where the teacher's guidance is less critical. The L2 norm measures the difference between the student's and teacher's high-frequency components at these selected pixels, normalized by the number of such pixels M.

To validate the effectiveness of this loss function, we compare it against two alternative loss functions commonly used in knowledge distillation:

- **Standard L2 Loss**: A pixel-wise L2 loss applied across all pixels without thresholding:

$$\mathcal{L}_{\text{L2}} = \frac{1}{N} \sum_{i,j} ||h_{i,j}^{\text{student}} - h_{i,j}^{\text{teacher}}||_2$$

  Here, N is the total number of pixels.
- **Masked L1 Loss**: An L1 loss applied only to pixels where the teacher's high-frequency component exceeds ϵ:

$$\mathcal{L}_{\text{L1}} = \frac{1}{M} \sum_{i,j} \mathbb{I}_{\{h_{i,j}^{\text{teacher}} > \epsilon\}} \cdot |h_{i,j}^{\text{student}} - h_{i,j}^{\text{teacher}}|$$

We evaluated these loss functions on the NYUD-v2 dataset, measuring performance with two metrics: edge preservation ($\varepsilon_{acc} \downarrow$), where a lower value indicates better edge detail retention, and depth accuracy (RMSE, in meters). The experimental results are summarized in Table 11.

Loss Function	$\varepsilon_{acc} \downarrow$	RMSE (m)
Standard L2 Loss	1.03	0.078
Masked L1 Loss	0.97	0.076
Our Distillation Loss	0.87	0.075

Table 11. Comparison of distillation loss functions on the NYUD-v2 dataset. Our proposed loss achieves the lowest ε_{acc} and RMSE, demonstrating superior edge preservation and depth accuracy.

The results in Table 11 demonstrate the superiority of our chosen distillation loss. The standard L2 loss, which applies the penalty uniformly across all pixels, yields the highest ε_{acc} of 1.00 and an RMSE of 0.078 m. This suggests that it struggles to prioritize high-frequency details, diluting its effectiveness by enforcing similarity in less informative low-frequency regions. The masked L1 loss improves upon

this by focusing only on high-frequency regions, achieving an ε_{acc} of 0.95 and an RMSE of 0.076 m. However, our distillation loss, which combines the same selective masking with an L2 norm, outperforms both alternatives with an ε_{acc} of 0.90 and an RMSE of 0.075 m. This improvement indicates that the L2 norm better captures the magnitude of differences in high-frequency components compared to the L1 norm, leading to enhanced edge preservation and depth accuracy.

In summary, our pixel-wise distillation loss, by selectively applying an L2 penalty to regions with significant high-frequency content, provides the optimal balance between focusing on critical details and maintaining overall accuracy. The experimental results on the NYUD-v2 dataset confirm its effectiveness, making it the preferred choice for our depth completion framework.

10. Multi-Scale Frequency Distillation

In our depth completion framework, both the teacher and student models employ decoders to produce multi-scale features, which can be interpreted as depth maps at varying levels of granularity. Following the design of prevalent prediction heads, such as DPT [28], these feature maps are generated progressively at scales ranging from $\frac{1}{32}$ to 1, with intermediate scales including $\frac{1}{16}$, $\frac{1}{8}$, $\frac{1}{4}$, and $\frac{1}{2}$. Our method leverages frequency-domain knowledge distilled from the teacher's multi-scale depth maps to enhance the student's predictions at each corresponding layer. This multi-scale distillation process utilizes the Adaptive Local Wavelet Transform (ALWT) to extract high-frequency components from these depth maps, enabling precise guidance of structural details across scales. In this section, we justify the effectiveness of this multi-scale frequency distillation approach through experimental analysis.

10.1. Why Multi-Scale Distillation?

Distilling frequency-domain knowledge at multiple scales offers several advantages over single-scale distillation:

- **Granular Detail Enhancement**: Multi-scale distillation allows the student to learn high-frequency details at different resolutions, ensuring that both fine edges (captured at higher resolutions) and broader structural patterns (evident at lower resolutions) are preserved.
- **Progressive Refinement**: By aligning the student's predictions with the teacher's across multiple layers, the approach facilitates a coarse-to-fine refinement process, improving overall depth accuracy.
- **Robustness Across Scenes**: Depth maps exhibit varying levels of complexity depending on the scene (e.g., indoor vs. outdoor). Multi-scale distillation adapts to these variations, providing consistent guidance regardless of the spatial scale of features.

10.2. Experimental Validation

To assess the effectiveness of multi-scale frequency distillation, we conducted experiments on the NYUD-v2 dataset, comparing our approach against two baselines: (1) no distillation (where the student relies solely on ground truth supervision) and (2) single-scale distillation (distilling only at the final scale of 1).

The results are presented in Table 12.

Method	$\varepsilon_{acc} \downarrow$	RMSE (m)	Processing Time (ms)
No Distillation	1.25	0.082	10.0
Single-Scale Distillation	1.05	0.078	11.5
Multi-Scale Distillation	0.91	0.075	13.0

Table 12. Comparison of distillation strategies in our depth completion framework on the NYUD-v2 dataset. Multi-scale distillation achieves the lowest ε_{acc} and RMSE, with a modest increase in processing time, demonstrating its effectiveness.

Table 12 illustrates the benefits of multi-scale frequency distillation. The baseline with no distillation yields the weakest performance, with an ε_{acc} of 1.25 and an RMSE of 0.082 m, indicating poor edge preservation and depth accuracy due to the lack of teacher guidance. Single-scale distillation improves these metrics to an ε_{acc} of 1.05 and an RMSE of 0.078 m by leveraging the teacher's knowledge at the final scale, though it struggles to capture details at intermediate resolutions. In contrast, our multi-scale distillation approach achieves the best results, with an ε_{acc} of 0.90 and an RMSE of 0.075 m, reflecting superior edge preservation and depth accuracy. This improvement stems from the progressive transfer of high-frequency knowledge across scales, enabled by the ALWT.

While multi-scale distillation increases processing time to 13.0 ms compared to 10.0 ms for no distillation and 11.5 ms for single-scale distillation, this modest overhead is justified by the significant performance gains. The additional computational cost arises from extracting and distilling high-frequency components at multiple layers, a process optimized by the efficiency of the ALWT.

The experimental results in Table 12 confirm that multi-scale frequency distillation is highly effective for our depth completion framework. By distilling high-frequency knowledge from the teacher's multi-scale depth maps, our approach enhances the student's ability to reconstruct detailed and accurate depth maps across varying levels of granularity. This justifies the use of multi-scale distillation as a core component of our method, offering a robust and effective solution for improving depth completion performance.

11. Dataset Details

In this section, we provide detailed information about the datasets used in this work, including specific details and the

data pre-processing methods applied.

NYU Depth v2 (NYUD-v2) dataset [31] is an indoor dataset containing 464 scenes with a resolution of 640 × 480, captured by a Microsoft Kinect sensor. We utilize the official split, allocating 249 scenes (50K samples) for training and the remaining 215 scenes (654 images) for testing. For a fair comparison, we evaluate only the pixels within the crop defined in [31] across all methods.

KITTI Depth Completion (KITTI-DC) dataset [34] is an outdoor dataset in the autonomous driving domain. It provides 86,898 training samples, 1,000 validation samples, and 1,000 testing samples, each with corresponding raw LiDAR scans and reference images. We randomly crop the frames to 1216 × 256 for training and use the full-resolution frames as input for testing.

iBims-1 dataset [19] comprises 100 indoor RGB-D pairs specifically designed for testing. It was collected using a digital single-lens reflex (DSLR) camera and a high-precision laser scanner. Compared to NYUD-v2, iBims-1 is distinguished by its high quality, featuring sharp and flawless depth transitions and low noise levels, making it a better standard for evaluating high-frequency acuity. Since iBims-1 contains only 100 pairs, we perform zero-shot generalization tests using the model trained on NYUD-v2.

DDAD [11] is an autonomous driving benchmark focused on long-range (up to 250m) and dense depth estimation in diverse urban environments. The dataset includes monocular videos and precise ground-truth depth across a 360-degree field of view, collected using high-density LiDARs mounted on self-driving cars in the U.S. and Japan. The images, captured with a synchronized 6-camera array, were downsampled from 1216×1936 to 384×640, and the 3,950 official validation samples were used for evaluation. Due to the low percentage of valid ground truth depth post-downsampling, we sampled all available valid depth points to ensure meaningful results.

12. Evaluation Metrics

We provide a detailed description of the evaluation metrics adopted in this paper, with a particular emphasis on high-frequency performance metrics, which constitute the most significant contribution of our work. To ensure fair and effective comparisons, we utilize the depth boundary error (DBE) proposed by [19] to assess the high-frequency performance of depth images. This evaluation metric has been widely used in various works, including those on depth estimation [14, 39] and depth refinement [12], to evaluate the high-frequency accuracy of depth maps.

Depth discontinuities, represented as strong gradient changes in depth maps, are crucial high-frequency elements in accurate and well-represented depth images. It is essential to evaluate whether predicted depth maps correctly represent these discontinuities or introduce fictitious ones due to texture confusion. Depth boundary errors (DBEs) are defined to assess both accuracy ε_{acc} and completeness $\varepsilon_{\text{comp}}$. Accuracy is measured by the distance between predicted and ground truth edges, while completeness evaluates the presence of missing edges in the predicted map.

To compute ε_{acc} and $\varepsilon_{\text{comp}}$, boundaries are first extracted using a Canny edge detector. The predicted edges Y are then compared to the ground truth edges Y^* using a truncated Chamfer distance, with a Euclidean distance transform applied to $E^* = \text{DT}(Y^*)$, where $\text{DT}(\cdot)$ is the Euclidean distance transform function. Distances exceeding a threshold θ are ignored to focus on local accuracy:

$$\varepsilon_{\text{acc}}(Y) = \frac{1}{N}\sum_{i,j} E^*_{i,j} \odot Y_{i,j} \tag{9}$$

$$\varepsilon_{\text{comp}}(Y) = \frac{1}{N}\sum_{i,j} E_{i,j} \odot Y^*_{i,j} \tag{10}$$

where N is the number of valid pixels, $\odot$ denotes element-wise multiplication, and $E = \text{DT}(Y)$.

As for depth fidelity, we follow the standard evaluation metrics from [27] for a fair comparison. These evaluation metrics are computed as follows:

- Root Mean Squared Error (RMSE):

$$\sqrt{\frac{1}{N}\sum_{i,j}(\mathbf{D}_{i,j} - \mathbf{D}^{gt}_{i,j})^2} \tag{11}$$

- Mean Absolute Error (MAE):

$$\frac{1}{N}\sum_{i,j}|\mathbf{D}_{i,j} - \mathbf{D}^{gt}_{i,j}| \tag{12}$$

- Absolute Relative Error (AbsRel):

$$\frac{1}{N}\sum_{i,j}\frac{|\mathbf{D}_{i,j} - \mathbf{D}^{gt}_{i,j}|}{\mathbf{D}^{gt}_{i,j}} \tag{13}$$

- Threshold accuracy ($\delta < 1.25^k$):

$$\max\left(\frac{\mathbf{D}_{i,j}}{\mathbf{D}^{gt}_{i,j}}, \frac{\mathbf{D}^{gt}_{i,j}}{\mathbf{D}_{i,j}}\right) < 1.25^k \tag{14}$$

- Root Mean Squared Error of the inverse depth (iRMSE):

$$\sqrt{\frac{1}{N}\sum_{i,j}\left(\frac{1}{\mathbf{D}_{i,j}} - \frac{1}{\mathbf{D}^{gt}_{i,j}}\right)^2} \tag{15}$$

- Mean Absolute Error of the inverse depth (iMAE):

$$\frac{1}{N}\sum_{i,j}\left|\frac{1}{\mathbf{D}_{i,j}} - \frac{1}{\mathbf{D}^{gt}_{i,j}}\right| \tag{16}$$

where N is the number of pixels, (i, j) denotes the pixel index, and $\mathbf{D}^{gt}$ represents the ground truth depth.

13. Discrete Wavelet Transform (DWT)

The **Discrete Wavelet Transform (DWT)** is a versatile technique in signal and image processing that decomposes 2D images into distinct frequency components, enabling a powerful frequency-domain analysis while preserving spatial locality. This dual capability distinguishes DWT from traditional transforms like the Fourier Transform and makes it particularly well-suited for applications involving depth maps, where both smooth variations and abrupt changes (e.g., object boundaries) must be accurately captured.

13.1. Conceptual Foundation

At its core, DWT provides a multi-resolution analysis, breaking down an image into components at different scales and frequencies. Unlike the Fourier Transform, which represents a signal as a sum of globally defined sine and cosine waves, DWT employs localized basis functions—wavelets—that vary in scale and position. This localization allows DWT to capture both frequency information (how rapidly the signal changes) and spatial information (where those changes occur). For depth maps, denoted as $\boldsymbol{x}$, this property is invaluable, as depth scenes often feature sharp discontinuities (e.g., edges between objects) alongside gradual transitions (e.g., flat surfaces). The foundation of this analysis lies in the dilation and translation of a mother wavelet function $\Phi(t)$, forming an orthogonal wavelet basis as follows:

$$\Phi_{(s,d)}(t) = 2^{\frac{s}{2}}\Phi(2^s t - d), \quad s, d \in \mathbb{Z}$$

where s and d are the scaling and translation parameters, respectively, $\mathbb{Z}$ is the set of integers, and the factor $2^{\frac{s}{2}}$ ensures a constant norm independent of scale s. This equation generates a family of wavelets in L^2 spaces, enabling a scalable representation of the signal.

13.2. Decomposition Process

Given a depth map $\boldsymbol{x}$, the DWT generates four subbands through filtering and downsampling. The decomposition is achieved using a pair of filters derived from the wavelet and scaling functions:

- LL: The low-frequency **approximation subband**, obtained by convolving the image with a low-pass filter $h[n]$ along both rows and columns, followed by downsampling by 2.
- HL: The high-frequency **horizontal detail subband**, produced by applying a high-pass filter $g[n]$ to the rows and a low-pass filter $h[n]$ to the columns, then downsampling.
- LH: The high-frequency **vertical detail subband**, obtained with a low-pass filter $h[n]$ on rows and a high-pass filter $g[n]$ on columns, followed by downsampling.
- HH: The high-frequency **diagonal detail subband**, resulting from high-pass filtering $g[n]$ in both directions, with subsequent downsampling.

This process is mathematically expressed as:

$$\{\text{LL}, \text{HL}, \text{LH}, \text{HH}\} = \text{DWT}(\boldsymbol{x})$$

where the subband coefficients are computed as:

$$\text{LL}[m,n] = \sum_k h[k]h[l]\boldsymbol{x}[2m-k, 2n-l]$$

$$\text{HL}[m,n] = \sum_k g[k]h[l]\boldsymbol{x}[2m-k, 2n-l]$$

$$\text{LH}[m,n] = \sum_k h[k]g[l]\boldsymbol{x}[2m-k, 2n-l]$$

$$\text{HH}[m,n] = \sum_k g[k]g[l]\boldsymbol{x}[2m-k, 2n-l]$$

Each subband is reduced to half the resolution of the original image due to downsampling by a factor of 2, providing a complete representation that separates the image into a low-resolution approximation and high-frequency details in three orientations.

13.3. Multi-Scale Hierarchy

A defining feature of DWT is its ability to create a **multi-scale hierarchy** by recursively applying the transform to the LL subband. For instance, at decomposition level 2, the LL subband from the first level is further decomposed into:

$$\{\text{LL}_2, \text{HL}_2, \text{LH}_2, \text{HH}_2\} = \text{DWT}(\text{LL})$$

This recursion can continue, with each level halving the resolution and doubling the scale of analysis, forming a pyramidal structure. Higher levels (e.g., LL_2) represent coarser approximations, while detail subbands $(\text{HL}_k, \text{LH}_k, \text{HH}_k)$ at level k capture progressively larger-scale features. For depth maps, this multi-scale approach facilitates adaptive analysis, enabling finer granularity in complex regions (e.g., cluttered objects) and coarser analysis in uniform areas (e.g., walls).

13.4. Mathematical Underpinnings

The DWT relies on a pair of quadrature mirror filters—low-pass $h[n]$ and high-pass $g[n]$—derived from the scaling function $\phi(t)$ and wavelet function $\psi(t)$, respectively. These filters satisfy the two-scale relation:

$$\phi(t) = \sqrt{2}\sum_n h[n]\phi(2t-n)$$

$$\psi(t) = \sqrt{2}\sum_n g[n]\phi(2t-n)$$

where $g[n] = (-1)^n h[1-n]$ ensures orthogonality in some wavelet families. The DWT coefficients are computed by convolving the input signal with these filters and downsampling, preserving the energy of the original signal under certain conditions (e.g., perfect reconstruction).

13.5. Invertibility and Reconstruction

A key advantage of DWT is its **invertibility**. The **inverse DWT (iDWT)** reconstructs the original depth map $\boldsymbol{x}$ from its subbands without loss, leveraging synthesis filters $\tilde{h}[n]$ and $\tilde{g}[n]$. This involves upsampling (inserting zeros) and convolving with synthesis filters, ensuring perfect reconstruction when the wavelet basis satisfies biorthogonality or orthogonality conditions. This property is crucial for depth completion, allowing processed subbands to be reassembled into a coherent depth map.

13.6. Wavelet Basis Functions

The choice of **wavelet basis function** profoundly affects DWT's performance. Wavelet families (e.g., Haar, Daubechies, Biorthogonal) vary in properties like orthogonality, symmetry, and vanishing moments, influencing edge preservation and computational cost. For depth maps, wavelets with high vanishing moments (e.g., Daubechies db4) and symmetry (e.g., bior3.3) are preferred to capture sharp transitions and reduce artifacts.

13.7. Practical Relevance to Depth Completion

In depth completion, DWT's multi-scale decomposition enables targeted enhancement of high-frequency subbands $(\mathrm{HL}, \mathrm{LH}, \mathrm{HH})$, critical for reconstructing edges and textures from sparse inputs. The invertibility and adaptability of DWT support our framework's goal of reconstructing accurate depth maps efficiently.

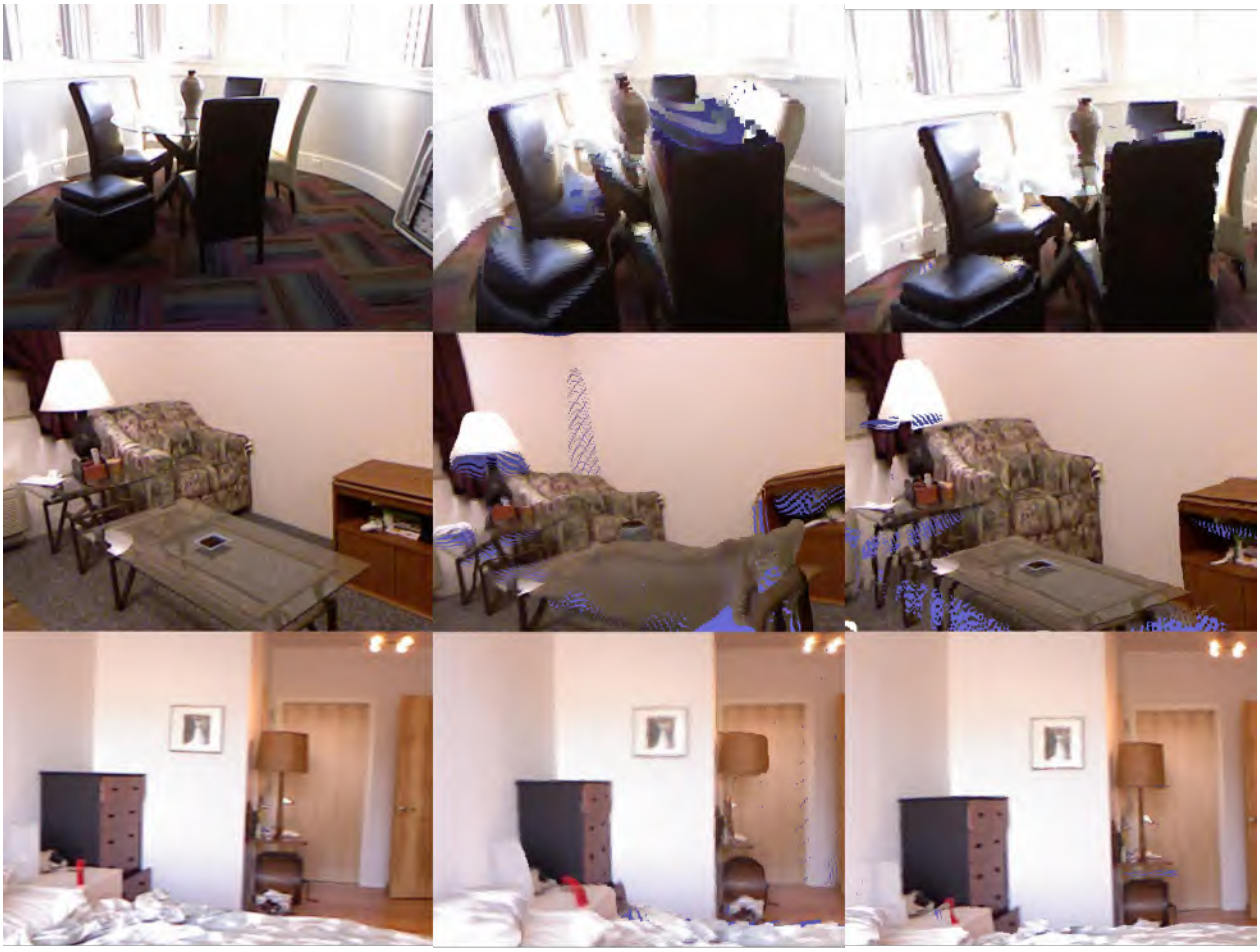

Figure 11. More point cloud visualization results.