PE3R: Perception-Efficient 3D Reconstruction

Supplementary Material

6. Proof of Interpolation Properties

This section provides the complete proofs for the two key properties of our area-weighted interpolation strategy, which underpin the effectiveness of the Pixel Embedding Disambiguation described in Sec. 3.3 of the main text.

Proposition 1. *Vector Normalization:* The interpolated vector $\hat{\mathbf{F}}_B$ preserves unit norm, ensuring it remains within the original semantic embedding space.

Proof. The norm of $\hat{\mathbf{F}}_B$ is given by:

$$\|\hat{\mathbf{F}}_B\|^2 = \|a\mathbf{F}_A + b\mathbf{F}_B\|^2 . \tag{8}$$

Expanding this expression:

$$\begin{aligned}\|\hat{\mathbf{F}}_B\|^2 &= (a\mathbf{F}_A + b\mathbf{F}_B)\cdot(a\mathbf{F}_A + b\mathbf{F}_B) \\ &= a^2\|\mathbf{F}_A\|^2 + 2ab\mathbf{F}_A\cdot\mathbf{F}_B + b^2\|\mathbf{F}_B\|^2.\end{aligned} \tag{9}$$

Since $\mathbf{F}_A$ and $\mathbf{F}_B$ are unit vectors:

$$\|\mathbf{F}_A\| = 1, \|\mathbf{F}_B\| = 1, \mathbf{F}_A\cdot\mathbf{F}_B = \cos(\theta). \tag{10}$$

Substituting these values, we get:

$$\begin{aligned}\|\hat{\mathbf{F}}_B\|^2 = \frac{1}{\sin^2(\theta)}&(\sin^2((1-t)\theta) + \sin^2(t\theta) \\ &+ 2\sin((1-t)\theta)\sin(t\theta)\cos(\theta)).\end{aligned} \tag{11}$$

Using trigonometric identities:

$$\begin{aligned}\sin^2(\theta) = \sin^2((1-t)\theta) &+ \sin^2(t\theta) \\ &+ 2\sin((1-t)\theta)\sin(t\theta)\cos(\theta),\end{aligned} \tag{12}$$

we find that:

$$\|\hat{\mathbf{F}}_B\|^2 = \sin^2(\theta)/\sin^2(\theta) = 1. \tag{13}$$

Thus, $\hat{\mathbf{F}}_B$ is confirmed to be a unit vector. □

Proposition 2. *Semantic Guidance:* If $\mathbf{F}_A$ has a higher similarity to a reference semantic vector $\mathbf{F}_C$ than $\mathbf{F}_B$ does, then $\hat{\mathbf{F}}_B$ will also exhibit a higher similarity to $\mathbf{F}_C$ than $\mathbf{F}_B$ does. This steers the aggregated representation toward more semantically meaningful directions.

Proof. The cosine similarity between $\hat{\mathbf{F}}_B$ and $\mathbf{F}_C$ is:

$$\hat{\mathbf{F}}_B\cdot\mathbf{F}_C = a(\mathbf{F}_A\cdot\mathbf{F}_C) + b(\mathbf{F}_B\cdot\mathbf{F}_C). \tag{14}$$

Since $\mathbf{F}_A\cdot\mathbf{F}_C > \mathbf{F}_B\cdot\mathbf{F}_C$, we have:

$$a(\mathbf{F}_A\cdot\mathbf{F}_C) + b(\mathbf{F}_B\cdot\mathbf{F}_C) > (a+b)(\mathbf{F}_B\cdot\mathbf{F}_C). \tag{15}$$

Threshold	rel↓	τ ↑
0.001	6.1	49.3
0.002	5.9	50.3
0.003	**5.5**	**55.1**
0.004	5.6	53.6
0.005	5.8	52.8

Table 8. Anomaly point selection with varying thresholds on ScanNet. The optimal value of 0.003 is used as the default in all main experiments.

Using trigonometric properties:

$$a + b = \frac{\sin((1-t)\theta)}{\sin(\theta)} + \frac{\sin(t\theta)}{\sin(\theta)} = 1, \tag{16}$$

we conclude:

$$\hat{\mathbf{F}}_B\cdot\mathbf{F}_C = a(\mathbf{F}_A\cdot\mathbf{F}_C) + b(\mathbf{F}_B\cdot\mathbf{F}_C) > \mathbf{F}_B\cdot\mathbf{F}_C. \tag{17}$$

This confirms that $\hat{\mathbf{F}}_B$ semantically integrates information from both $\mathbf{F}_A$ and $\mathbf{F}_B$, steering the aggregated representation toward more meaningful directions. □

7. Extended Ablation Studies

This section provides extended ablation studies that complement the analysis in Sec. 4.3 of the main paper, offering further insights into the design choices and parameter sensitivity of PE3R.

7.1. Anomaly Point Selection

The accuracy of our anomaly point detection mechanism is validated through both empirical and theoretical analysis.

Empirical Validation. We determine the anomaly threshold by statistically analyzing the mean 3D distance between spatially consistent points across semantic regions. This threshold effectively separates normal points from anomalies. Table 8 presents results under different thresholds on ScanNet, with the optimal performance observed at a threshold of 0.003.

Theoretical Validation. While anomaly detection could theoretically be approached through parametric modeling of intra-object distributions (*e.g.*, using Gaussian estimation or least-squares fitting), such methods become computationally prohibitive in complex, large-scale scenes. Our empirical thresholding strategy provides a practical and scalable alternative that maintains high performance without introducing significant computational overhead.

Window Size	avg. rel↓	avg. τ ↑
3×3	4.9	65.5
5×5	**4.5**	**68.0**
7×7	4.8	66.1

Table 9. Effect of sliding window size on anomaly detection performance across all depth estimation benchmarks.

a	b	avg. rel↓	avg. τ ↑
0.00	1.00	5.3	60.2
0.10	0.90	**4.5**	**68.0**
0.20	0.80	4.7	62.3
0.50	0.50	4.9	61.6
0.80	0.20	6.1	59.4
0.90	0.10	6.5	57.5
1.00	0.00	10.2	50.2

Table 10. Effect of semantic-aware smoothing parameter α on reconstruction performance. Moderate smoothing ($\alpha = 0.1$) achieves the best results.

7.2. Sliding Window Size Analysis

The size of the sliding window used for local semantic-aware distance computation significantly impacts both reconstruction quality and computational efficiency. As demonstrated in Table 9, we evaluate three window sizes on the multi-view depth estimation task. A 5×5 window achieves the optimal trade-off, capturing sufficient local context for robust anomaly detection without introducing excessive computational cost or over-smoothing fine geometric details.

7.3. RGB Image Smoothing Analysis

To enhance robustness in challenging regions characterized by reflections, transparency, or complex textures, we apply semantic-aware smoothing to the input images prior to pointmap estimation. The smoothing operation is defined as $\hat{\mathbf{y}} = \alpha \cdot \mathbf{x} + (1 - \alpha) \cdot \mathbf{y}$, where $\mathbf{x}$ is the mean RGB value of the semantic region, $\mathbf{y}$ is the original pixel value, and α controls the smoothing strength.

As shown in Table 10, moderate smoothing ($\alpha = 0.1$) provides the optimal balance between noise suppression and detail preservation. Lower values ($\alpha \leq 0.1$) insufficiently address visual artifacts, while higher values ($\alpha \geq 0.2$) over-smooth genuine structural details, ultimately degrading geometric accuracy.

7.4. Iterative Refinement Analysis

We further investigate the effect of repeated refinement iterations in the semantic point cloud reconstruction module. As shown in Table 11, while a single iteration provides substantial improvement, additional iterations yield diminishing returns with increased computational cost. This suggests that our method achieves effective refinement in a single pass, maintaining the efficiency advantages of the overall feed-forward pipeline.

Iterations	1	2	3	4
avg. rel↓	**4.5**	4.6	4.6	4.6
avg. τ ↑	**68.0**	67.9	67.9	67.9

Table 11. Effect of iterative refinement cycles on reconstruction performance (Mip-NeRF360). A single iteration achieves optimal performance.