

SHED Light on Segmentation for Depth Estimation

Supplementary Material

1. Implementation Details

1.1. Training details

We train our model on the NYUv2 dataset [2] using the official training split. Each RGB image is first cropped to remove invalid boundaries (coordinates: 43, 45, 608, 472), then resized to 416×416 resolution. The corresponding depth maps undergo the same spatial preprocessing and are normalized by dividing raw depth values by 1000.

For data augmentation, we apply horizontal flipping with a probability of 0.5, gamma correction with $\gamma \in [0.9, 1.1]$, brightness scaling using a random factor from $[0.75, 1.25]$, and per-channel color jittering with multiplicative factors in $[0.9, 1.1]$. After augmentation, random spatial crops of size 416×416 are applied to both images and depth maps.

For tokenization, we generate superpixels using OpenCV's SEEDS [6] algorithm. Each image is segmented into 676 superpixels using a single-level hierarchy (`num_levels=1`) and a histogram bin size of 5. The algorithm is run for 50 iterations to refine superpixel boundaries.

1.2. Evaluation details

We evaluate on the official NYUv2 [2] test split, which contains 654 images. All evaluations use an input resolution of 416×416 pixels, with depth values clamped to the range $[10^{-3}, 10.0]$.

Per-pixel depth metrics and efficiency. We compute standard depth estimation metrics over valid pixels where ground truth depth is available. Depth error metrics include AbsRel (mean absolute relative error), RMSE (root mean squared error), and Log10 (mean absolute logarithmic error). Accuracy is measured using threshold metrics δ^1 , δ^2 , and δ^3 , which denote the percentage of pixels where the predicted-to-ground-truth depth ratio is below 1.25 , 1.25^2 , and 1.25^3 , respectively. All metrics are computed with numerical safeguards, including epsilon clamping at $1e-6$ to prevent division by zero and log-domain errors. For efficiency, we report FLOPs using PyTorch's `FlopCounterMode` for a single forward pass on a 416×416 input.

Occlusion boundary. We follow the evaluation protocol of the NYUv2-OC++ dataset [3] to assess occlusion boundary accuracy. Each predicted depth map is first min-max normalized, followed by the application of the OpenCV Canny edge detector [1] with low and high thresholds of 100 and 200 to produce a binary mask of predicted boundary pixels. Using the ground-truth boundary labels from NYUv2-OC++, we compute two metrics: ε_a (accuracy), the average distance from each predicted edge pixel to the nearest ground-truth edge; and ε_c (consistency), the average dis-

tance from each ground-truth edge pixel to the nearest predicted edge. Both are reported in squared pixels, where lower values indicate better alignment, and 0 denotes perfect correspondence.

3D scene reconstruction. We reconstruct 3D point clouds by back-projecting each pixel of the predicted depth maps into 3D space using the known camera intrinsics from the NYUv2 dataset, following the standard protocol for semantic scene completion [5]. To evaluate reconstruction quality, we compute the Chamfer distance between the predicted and ground truth point clouds. Recall quantifies how well the predicted points cover the ground truth surface, while precision measures how accurately the predicted points align with the true geometry.

Detailed evaluation protocols for **intra-segment coherence** and **layout-aware retrieval** are provided in Secs. 2 and 3, respectively.

2. Intra-segment Coherence

We provide a detailed explanation of our proposed metric, *intra-segment coherence*, which evaluates the structural quality of depth maps by measuring how closely the depth distribution within each segment matches the corresponding reference ground-truth distribution.

Fig. 1 illustrates the evaluation of intra-segment coherence between predicted and ground-truth depth maps. For each ground-truth object segment, we first discard predicted pixels with depths outside the valid range $[d_{\min}, d_{\max}] = [0.10, 10.0]$. We then compute the one-dimensional first-order Wasserstein distance W between the predicted and reference depth distributions within each valid segment. An object is classified as a true positive (TP) if $W \leq \tau$, and as a false positive (FP) otherwise, given a tolerance threshold τ . By sweeping τ from 0 to 0.50 in uniform steps, we compute average precision and recall to provide a comprehensive evaluation across varying tolerance levels.

3. Layout-aware Retrieval

We provide a detailed explanation of our proposed metric, *layout-aware retrieval*, which evaluates the structural quality of learned representations by measuring how well the model retrieves frames from the same 3D scene in a video, either across the full sequence or within nearby frames.

To compute this metric, we extract the [CLS] token from the final layer of the vision transformer in DPT and SHED for each image, denoted as $\mathbf{z}_{\text{cls}} \in \mathbb{R}^D$, and apply l_2 normalization. We then construct a full pairwise similarity matrix, where each entry is computed as the cosine sim-

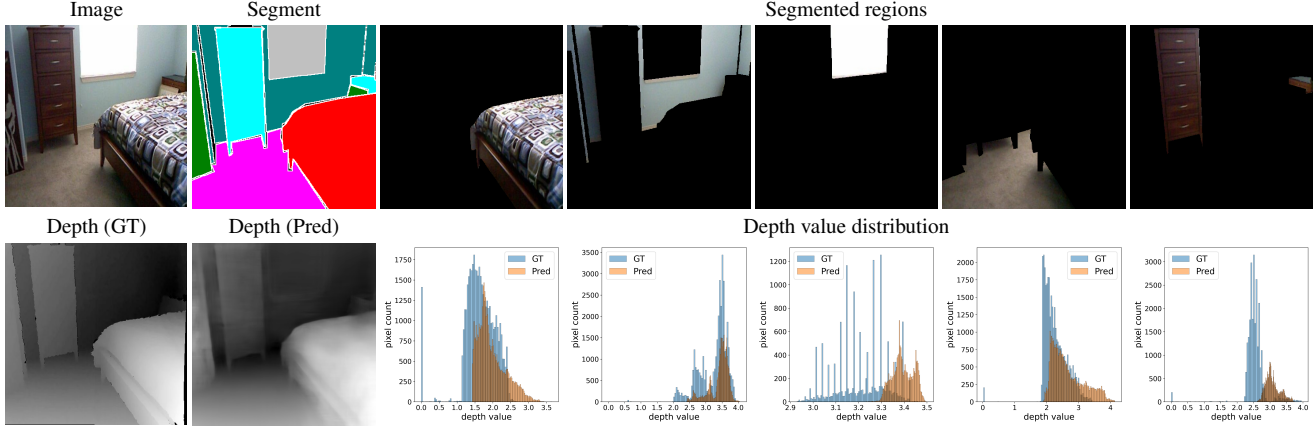


Figure 1. **Visualization of intra-segment coherence computation.** The image (top left) shows a bedroom scene with a bed, dresser, and window. The segment map (top middle) highlights different segmented regions using distinct colors: the bed in red, the dresser in blue, and the floor in pink. The segmented regions (top right) are shown overlaid on the original RGB image, illustrating how the image is partitioned. The histograms (bottom right) compare the depth value distributions of ground truth (GT, blue) and prediction (Pred, orange) within each segmented region across the range $[0.1, 10.0]$. The one-dimensional first-order Wasserstein distance W is computed between GT and predicted depth values for each region. A region is classified as a true positive if $W \leq \tau$, and τ is swept from 0 to 0.50 to compute average precision and recall.

ilarity between l_2 -normalized image embeddings: $S_{ij} = \mathbf{z}_{\text{cls}}^{(i)} \cdot \mathbf{z}_{\text{cls}}^{(j)}$.

Retrieval performance is evaluated in two settings: *scene-level* and *frame-level*. In the scene-level setting, the goal is to retrieve other frames from the same annotated scene, testing the model’s ability to maintain structural consistency under varying viewpoints. In the frame-level setting, each frame is treated as an independent query, focusing on retrieving visually similar frames regardless of scene membership. We additionally define a frame- k variant, where retrieval is restricted to the k temporally adjacent frames, allowing us to assess the model’s sensitivity to local layout changes. We report top- K nearest neighbor retrieval accuracy on the NYUv2 [2] test set.

To further evaluate the robustness of layout-aware retrieval, we vary both the candidate set size (k) and the top- K threshold. A smaller k imposes a stricter constraint, requiring the model to identify the most similar frame from a limited pool. As shown in Fig. 2, SHED consistently outperforms DPT across all settings. Performance improves for both methods as k increases, with the largest gap observed at low top- K values. These results suggest that SHED captures layout similarity more precisely and excels at retrieving the most relevant match.

4. Additional Visualizations

We provide additional visualizations of depth maps by SHED and DPT in Fig. 3 and more visualization of image retrieval in Fig. 4 and 3D reconstruction in Fig. 5.

5. Limitations and Broader Impacts

5.1. Limitations and future works

While our experiments focus on monocular depth estimation, the SHED framework is also applicable to other dense prediction tasks such as segmentation, optical flow, and image generation. We show that depth supervision induces emergent structural representations, including geometry-aware features and segmentations. Extending this approach to other objectives may uncover new forms of structure. For example, training on optical flow could yield motion-aware segmentations.

Depth estimation is fundamental to 3D scene understanding. Although we present initial results on 3D reconstruction, further investigation is needed to assess the utility of SHED in downstream tasks involving 3D reasoning and robotics. In particular, our framework enables unsupervised discovery of 3D object parts, which may serve as building blocks for modeling interactions, dynamics, and part affordances in physical environments and embodied systems.

Finally, while we compare SHED fairly against DPT under matched training conditions, state-of-the-art models such as Depth Anything benefit from large-scale pretraining and extensive engineering. Scaling up SHED with broader datasets and more compute would be a valuable next step. Notably, by jointly modeling segmentation and depth, SHED has the potential to evolve into a unified foundation model that combines the capabilities of Segment Anything and Depth Anything.

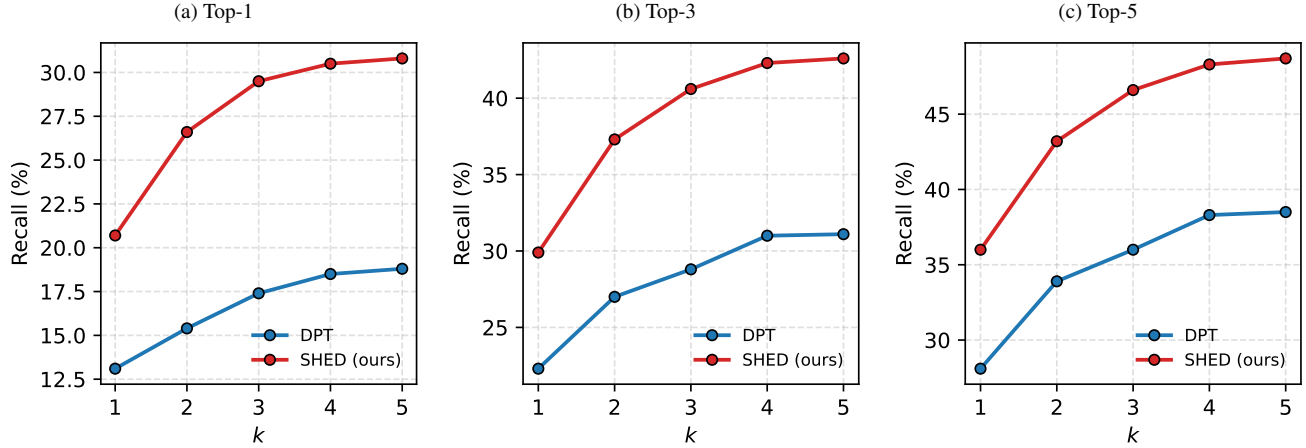


Figure 2. **Frame- k recall with varying temporal range k .** We report Top- K retrieval recall ($K \in \{1, 3, 5\}$) for DPT and SHED across temporal ranges $k \in [1, 5]$. The consistent gains highlight the robustness of our depth-supervised representation to spatial and viewpoint changes.

5.2. Broader impacts

Structured understanding of the 3D world and accurate depth estimation are central challenges in AI, with direct impact on safety-critical applications such as autonomous driving, augmented reality, and robotics. In practice, failures in these systems often result not from a lack of data but from insufficient structured reasoning, making predictions vulnerable to occlusion, unusual viewpoints, and dynamic environments. Our framework promotes geometry-aware perception by producing robust and interpretable depth estimates that align with scene structure. This can improve reliability in complex real-world settings and lead to systems with more transparent failure modes.



Figure 3. **Comparison of depth maps and boundaries.** SHED captures more accurate shapes with sharper boundaries, whereas DPT produces blurrier results. As a result, SHED correctly identifies boundaries (green), while DPT fails to detect them (red).

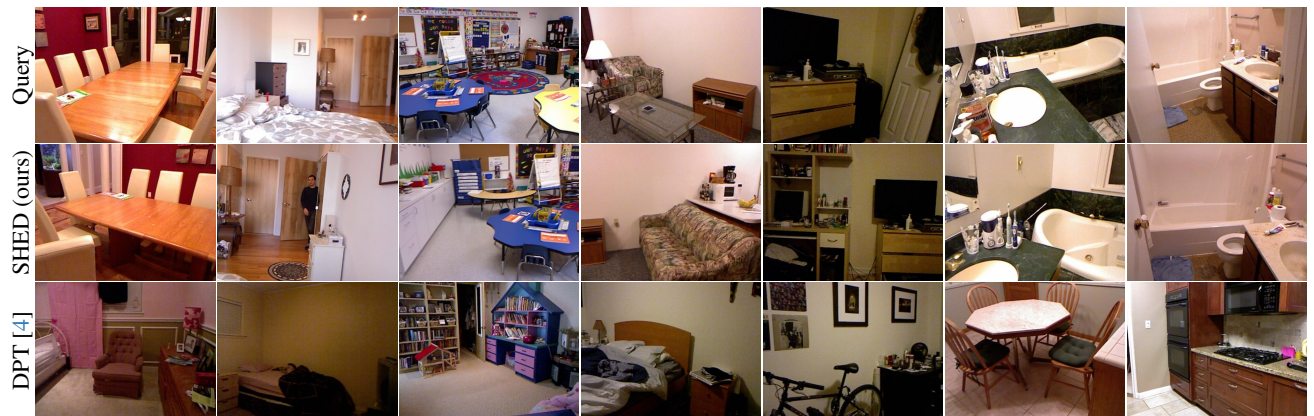


Figure 4. **Comparison of top-1 image retrieval results.** SHED retrieves samples that are more structurally similar to the query, indicating that its global embedding effectively captures scene layouts. In contrast, DPT focuses more on visual appearance, as shown in column 3, where it retrieves an image from a different scene that shares a similar color of blue.

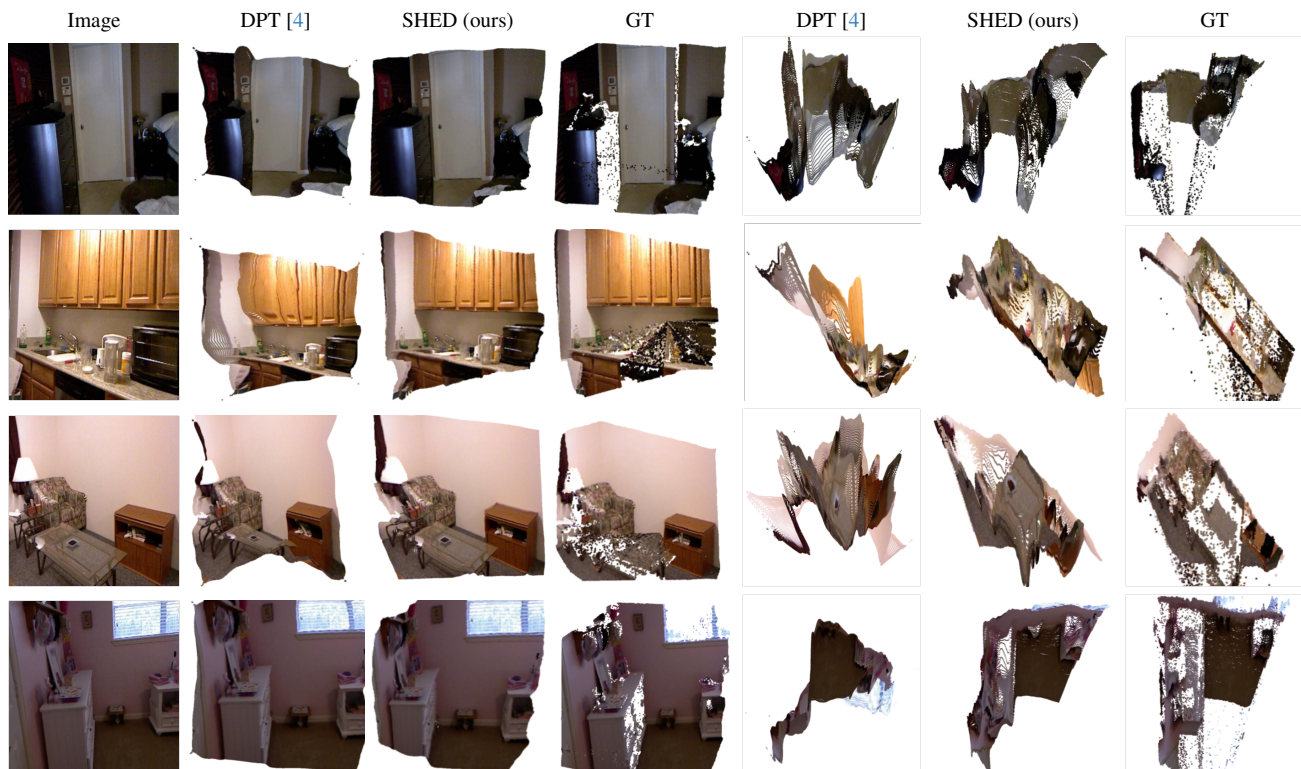


Figure 5. **Comparison of 3D reconstruction results.** Frontal views (cols 2–4) and bird’s-eye views (cols 5–7). DPT yields curved wall boundaries in the frontal views, which lead to distorted 3D reconstructions visible in the bird’s-eye views. In contrast, SHED produces sharp depth edges that preserve straight object contours and more faithfully represent the ground-truth 3D geometry.

163

References

164

[1] John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 1986. [1](#)

165

166

167

168

169

[2] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. [1](#), [2](#)

170

171

172

173

174

[3] Michael Ramamonjisoa, Yuming Du, and Vincent Lepetit. Predicting sharp and accurate occlusion boundaries in monocular depth estimation using displacement fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14648–14657, 2020. [1](#)

175

176

177

178

[4] 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. [4](#), [5](#)

179

180

181

182

183

[5] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1746–1754, 2017. [1](#)

184

185

186

187

188

189

[6] Michael Van den Bergh, Xavier Boix, Gemma Roig, Benjamin De Capitani, and Luc Van Gool. Seeds: Superpixels extracted via energy-driven sampling. In *Computer Vision—ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7–13, 2012, Proceedings, Part VII 12*, pages 13–26. Springer, 2012. [1](#)