Dual-S3D: Hierarchical Dual-Path Selective SSM-CNN for High-Fidelity Implicit Reconstruction

Supplementary Material

I. Expanded Experimental Comparisons

I.1. Detailed Comparisons

This section presents additional results to further illustrate the superiority of our method in reconstructing smooth and detailed 3D surfaces. Figure S1 provides zoomed-in comparisons across four examples (A, B, C, D), showcasing our method's consistent outperformance of SSR and GT in terms of cleaner surfaces, reduced artefacts, and better handling of occluded or damaged regions.

In Examples A and C, our method demonstrates strong geometric inference capabilities. In Example A, the cabinet reconstructed by our method exhibits smoother surfaces compared to SSR [36] and GT, which both suffer from noticeable noise unrelated to the input image. Moreover, in the occluded region where the sofa partially covers the cabinet, our method infers a more accurate geometry than SSR, successfully reconstructing the hidden surface. Similarly, in Example C, where SSR and GT results display visible damage and holes in the chair seat, our method reconstructs a complete, undamaged surface. This indicates that our model effectively learns harmonious geometry, even in challenging cases with incomplete input.

In Examples B and D, our method excels in capturing subtle surface details. For instance, in Example B, the reconstructed sofa armrest produced by our method is smoother and more rounded compared to SSR. Likewise, in Example D, the office chair cushion reconstructed by our method achieves a noticeably smoother and more realistic surface, enhancing the overall visual quality. These results highlight our model's ability to recover both global consistency and fine-grained surface details effectively.

The superior performance of our method can be attributed to two critical components. First, the depth-driven features provide robust geometric guidance, particularly in occluded or incomplete regions. Second, the Selective State-Space Model (SSM) [7] dynamically prioritizes feature alignment and fusion, balancing geometric stability and enhanced surface quality. Together, these innovations enable our method to produce smoother reconstructions and handle complex scenarios more effectively than existing approaches.

I.2. Normal Visualization Across Angles

Figure S2 presents a comparative analysis of 3D reconstruction results produced by our method alongside state-of-the-art (SOTA) approaches, including XTC [42], Omnidata [8], and ground truth (GT). The visualizations encompass input images, depth maps, and normal maps for various objects. The XTC [42], Omnidata [8], and GT data are sourced directly from [36], ensuring consistent and fair comparisons. XTC [42] emphasizes cross-modal feature learning, while Omnidata [8] utilizes pre-trained depth priors for single-image 3D reconstruction. These methods serve as baselines to evaluate our model's performance across diverse and challenging scenarios.

Our method exhibits several notable advantages. While baseline methods like XTC [42] and Omnidata [8] excel at enhancing local surface details, particularly in regions with sharp textures (e.g., the wrinkles on the bed in the first row), they often amplify noise and exhibit geometric inconsistencies in occluded or incomplete areas. In contrast, our method prioritizes global structure, resulting in smoother transitions and fewer noise artefacts across all examples. This balance between global consistency and local refinement underscores our model's robustness in handling complex scenes.

Although specific fine details, such as subtle wrinkles, may appear less pronounced in our reconstructions than XTC or Omnidata, the significant noise reduction and improved structural consistency make our method well-suited for real-world applications where smoothness and accuracy are paramount.

I.3. Robustness Under Occlusion and Multi-Angle Visualization

Figure S3 highlights the robustness of our method in handling occlusions and generating normal maps across a wide range of viewing angles, from -90° to 90°. The examples span diverse object categories, including chairs, tables, sofas, and beds, with input images often exhibiting partial occlusions or noisy elements. Despite these challenges, our method consistently produces high-quality and coherent normal maps, demonstrating adaptability to complex scenes and diverse object types.

Our method accurately reconstructs the complete geometry in the second row, where the table is partially occluded, preserving its flat surfaces and clean edges. Similarly, in the sixth row, the cabinet example illustrates our model's capacity to infer missing details caused by occlusions, such as occluded side panels, while maintaining global structural consistency. In the eighth row, the sofa reconstruction showcases our model's ability to achieve seamless transitions and realistic surface details even under severe occlusion.

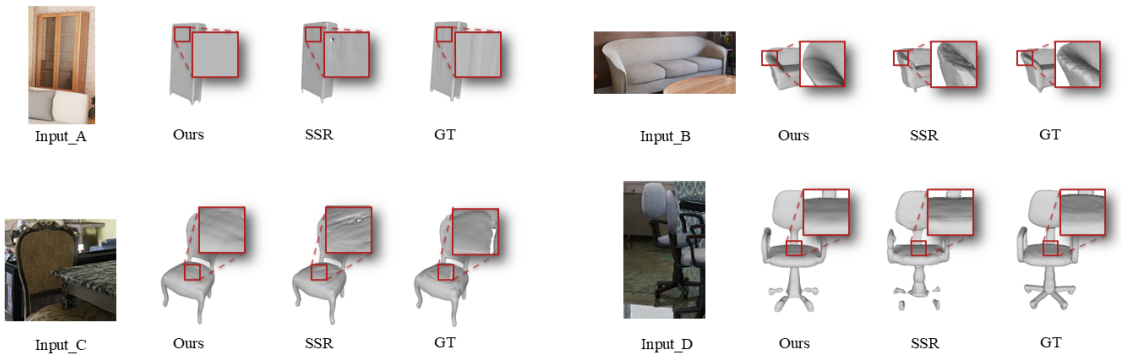

Figure S1. Zoomed-in comparisons of reconstructed model details. From left to right: Input images, results from our method (Ours), SSR, and GT. Examples include (A) a cabinet partially occluded by a sofa, (B) a sofa armrest, (C) a chair with a damaged seat, and (D) an office chair cushion.

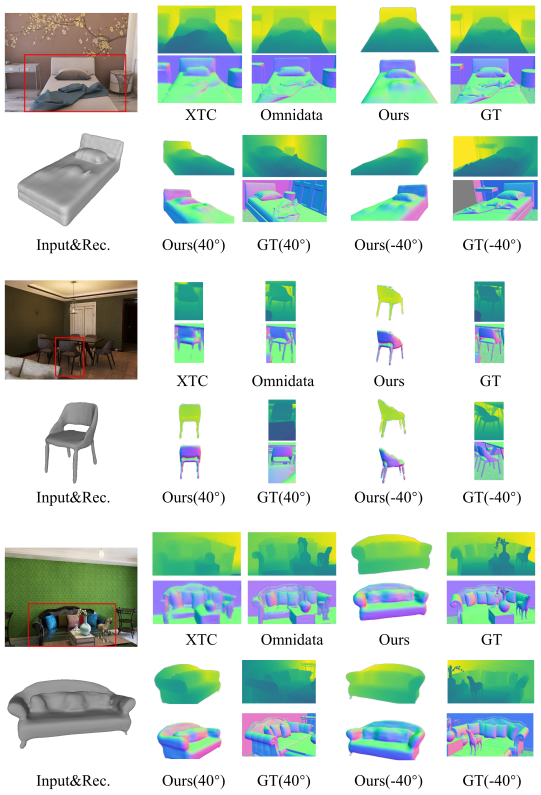

Figure S2. Comparative analysis of reconstruction results, including depth and normal maps, for various objects. Results for XTC [42], Omnidata [8], and GT are sourced from [36], while ours represents the proposed method.

These results underscore the robustness of our method in challenging scenarios. Moreover, the ability to generate consistent normal maps across multiple viewpoints highlights its suitability for applications requiring multi-angle analysis and high-fidelity 3D reconstructions.

II. Point Sampling

This supplementary material provides additional details on our point sampling strategy. In our framework, robust point sampling is achieved through an adaptive approach using the BoxBound. This module refines the sampling of points along rays by iteratively adjusting the sampling distribution based on error bounds computed from Signed Distance Function (SDF) evaluations. The refined samples are then used in the reconstruction inference pipeline to generate high-fidelity 3D surfaces.

II.1. Adaptive Point Sampling with BoxBoundSampler

The BoxBoundSampler begins with uniform sampling along each camera ray, then iteratively refines the sample set by:

1. **Initialization:**
   - Use a InitSampler to generate initial z-values (z_vals) along each ray between near and far clipping planes.
   - Compute initial error bounds based on the differences between consecutive samples.
2. **Iterative Refinement:**
   - For each iteration, compute 3D sample points along the rays using the current z_vals.
   - Evaluate the SDF at these sample points via the implicit network.
   - Calculate each interval's error bound d^* using geometric relations (e.g., via Heron's formula) and derive a local error estimate.
   - Update the beta parameter through a binary (line) search to satisfy the error threshold.
   - Upsample additional points by inverting the cumula-

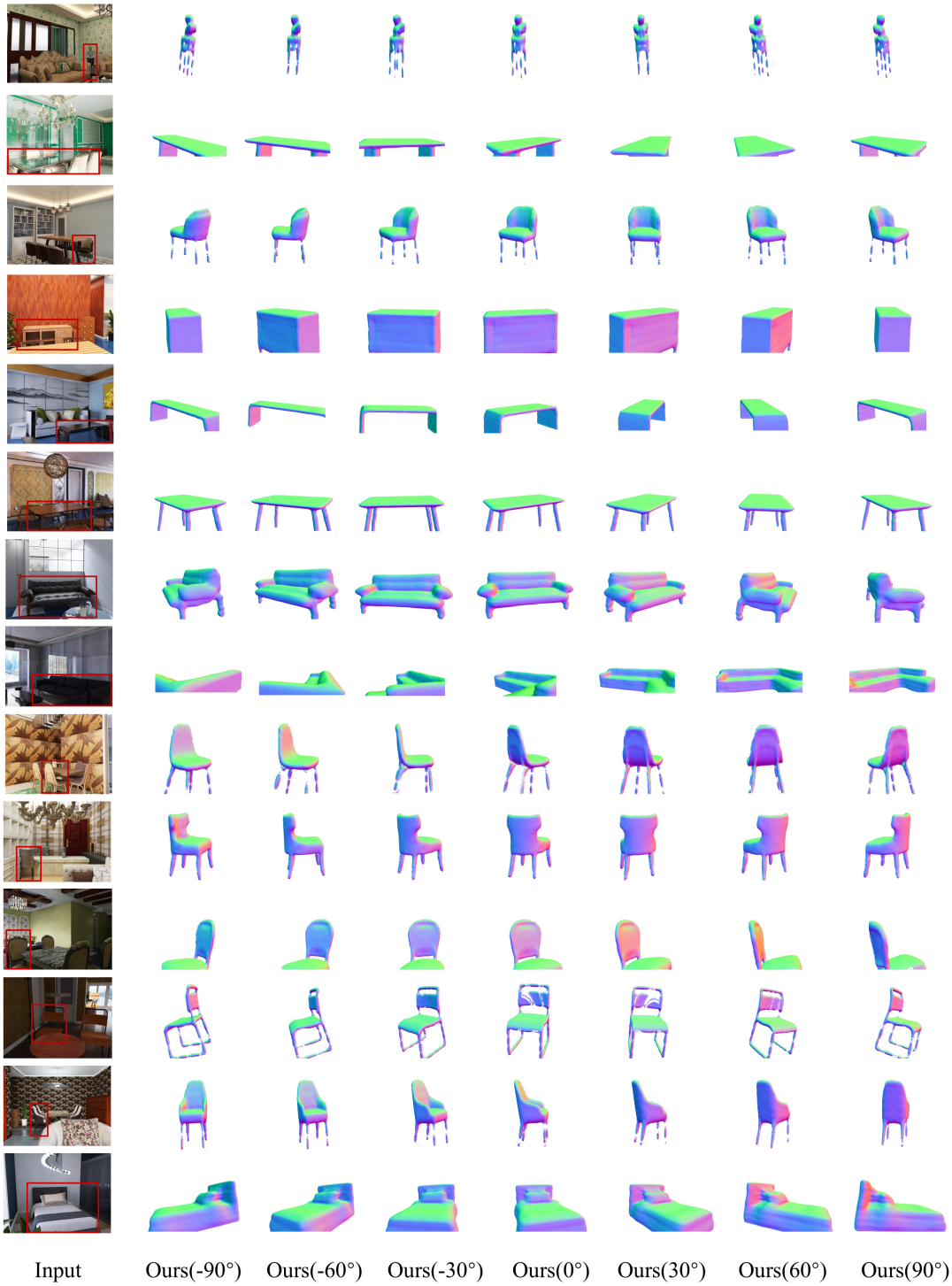

Figure S3. Multi-angle visualization of reconstructed normal maps from -90° to 90°. Rows showcase results for various objects, including chairs, tables, sofas, and beds. Despite significant occlusions and noisy inputs, our method produces consistent, high-quality reconstructions across all angles.

tive distribution function (CDF) computed from the estimated weights (derived from density and free energy).
- Merge and sort the new samples with the existing z_vals.

3. **Finalization:**
   - Optionally, extra samples from near and far boundaries are added.
   - The final set of z-values is returned, along with a randomly selected sample for eikonal loss computation.

The Algorithm S1 summarizes the key steps of the BoxBoundSampler.

II.2. Reconstruction Inference Pipeline

During reconstruction inference, the system performs the following steps:

- **Data Preparation:** Load test images, depth maps, and camera parameters (intrinsics, extrinsics, and pose).
- **Point Sampling:** Use the BoxBoundSampler (as detailed above) to generate refined ray samples.
- **Volume Rendering:** Integrate the sampled points into a volume rendering framework that leverages the SDF predictions to compute surface probabilities.
- **Mesh Extraction:** Apply a surface sliding algorithm to extract 3D meshes from the rendered volume. Both color and noncolor meshes can be exported for evaluation.

III. Scene Composition and Visualization

Reconstructing and composing complex 3D scenes remains a significant challenge in single-image 3D reconstruction. By leveraging datasets that provide detailed 3D bounding box annotations and camera pose information, our framework computes the world coordinates for individual objects and seamlessly integrates them into coherent scene representations. This enables the assembly of more miniature object reconstructions into more significant, unified scenes, paving the way for tackling even more challenging environments.

To simulate multi-view observations, our pipeline rotates the camera pose about the y-axis by preset angles (-90°, -45°, 0°, 45°, and 90°) using the rotation utility (rend_util.rot_camera_pose). For each rotation angle, the inference process splits the input image into manageable pixel blocks, processes them through the network, and then merges the outputs. This produces high-quality renderings—including RGB images, depth maps, and normal maps—that capture fine details and global structure.

Figures S4, S5, and S6 summarize our scene composition results:

- **Figure S4:** Reconstructed scenes based on depth maps, highlighting geometric precision and spatial relationships.

Algorithm S1 Adaptive Point Sampling with BoxBoundSampler

Require: Ray directions $R \in \mathbb{R}^{N_r \times 3}$, camera position $C \in \mathbb{R}^3$, initial uniform depth samples $z\_vals \in \mathbb{R}^{N_z}$, near/far plane bounds $(near, far)$, pretrained SDF model SDF_Model, maximum iterations T, convergence threshold ϵ

Ensure: Refined depth samples z_vals, randomly selected samples for Eikonal regularization z_sample_eik

$iter \leftarrow 0$, $converged \leftarrow false$
Initialize variance parameter β from initial z_vals
while $\neg converged \wedge iter < T$ **do**
  $P \leftarrow C + z\_vals \cdot R$ {Sample points along rays}
  $sdf \leftarrow$ SDF_Model(P) {Evaluate SDF at sampled points}
  for $i = 1$ **to** $|z\_vals| - 1$ **do**
    $dists[i] \leftarrow z\_vals[i+1] - z\_vals[i]$ {Compute inter-sample distances}
  end for
  $d^* \leftarrow$ ComputeBoxBound$(dists, sdf)$ {Estimate optimal distance bound}
  $\beta \leftarrow$ BinarySearchBeta(d^*, ϵ) {Optimize β via binary search}
  $density \leftarrow$ ComputeDensity(sdf, β) {Compute sampling density}
  $weights \leftarrow$ ComputeWeights$(density, dists)$ {Convert density to sampling weights}
  $z\_new \leftarrow$ InverseCDF$(weights)$ {Sample new depths via inverse transform sampling}
  $z\_vals \leftarrow$ SORT$(z_vals \cup z_new)$ {Merge and sort depth samples}
  $converged \leftarrow$ CheckConvergence(β, ϵ) {Evaluate convergence criteria}
  $iter \leftarrow iter + 1$
end while
if Extra sampling is enabled **then**
  $z\_vals \leftarrow$ MergeExtraSamples$(z_vals, near, far)$ {Refine boundary sampling (optional)}
end if
$z\_sample\_eik \leftarrow$ RANDOM_CHOICE(z_vals) {Random sampling for Eikonal loss}
return z_vals, z_sample_eik

- **Figure S5:** Scenes rendered with normal maps, emphasizing surface orientations and structural coherence.
- **Figure S6:** Fully rendered scenes that integrate geometry and texture for visually realistic results.

Each figure provides visualizations from -90° to 90°, offering a comprehensive 180° view of the scene. These results demonstrate the robustness of our approach in generating unified 3D representations that maintain high fidelity in geometry and texture, even in multi-object environments.

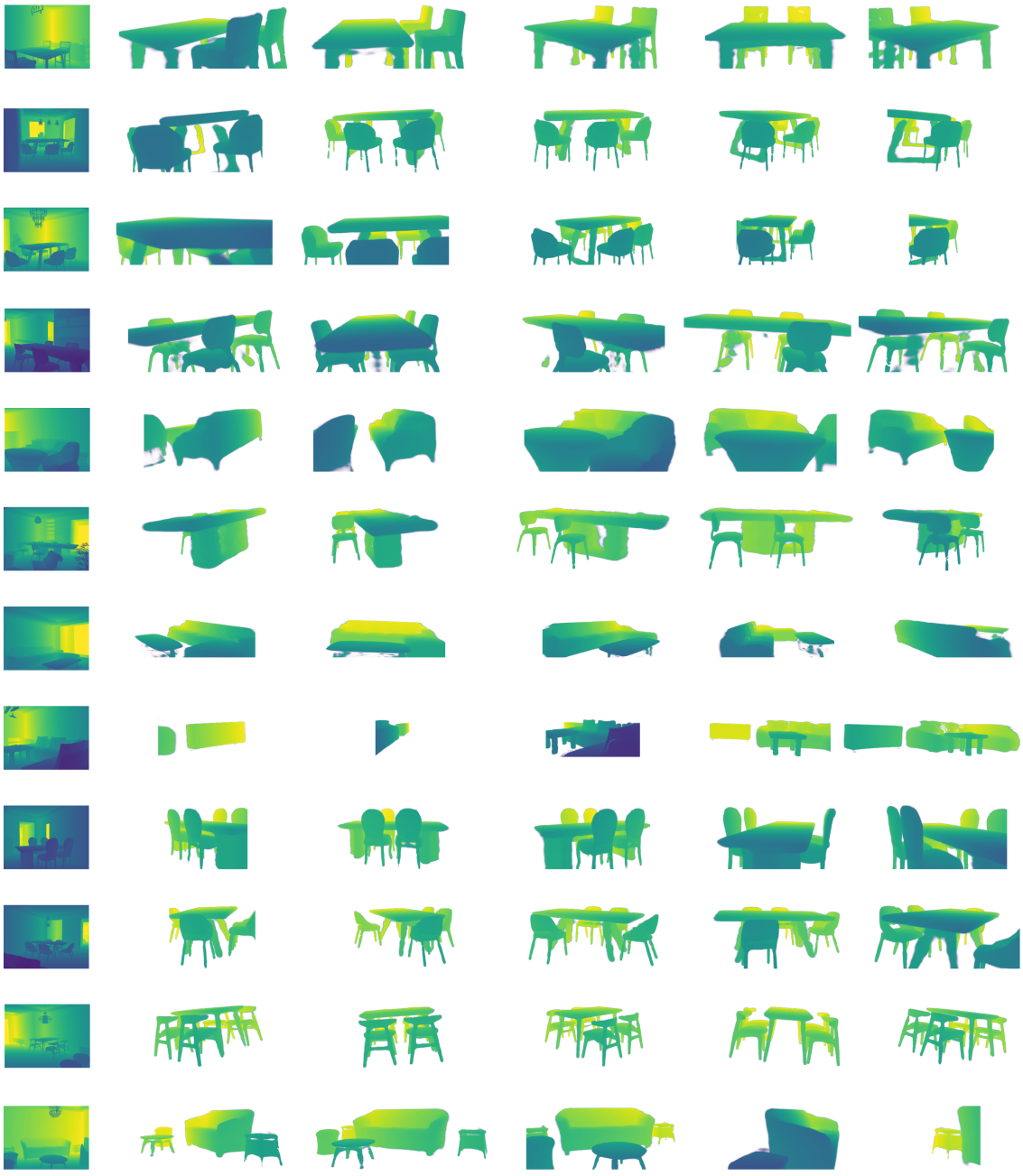

Figure S4. Scene composition visualized using depth information. Reconstructed scenes are shown with -90° to 90° views, emphasizing geometric accuracy and spatial relationships.

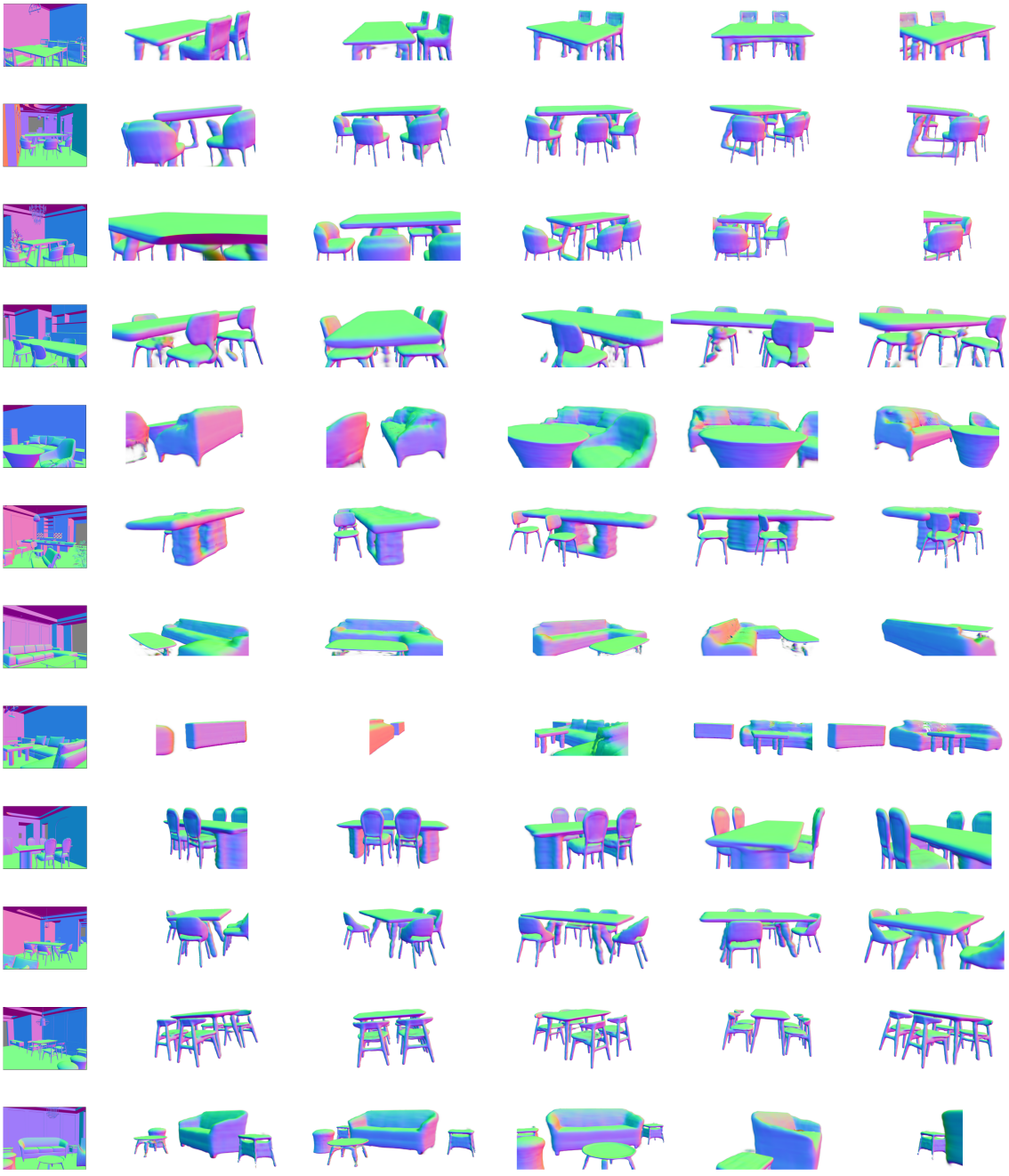

Figure S5. Scene composition visualized using normal maps. The visualization highlights surface orientations and structural coherence across the reconstructed scenes.

Figure S6. Fully rendered scene compositions integrating both geometry and texture. The results showcase visually realistic reconstructions with high fidelity in texture and geometry.