Occlusion-Aware Temporally Consistent Amodal Completion for 3D Human-Object Interaction Reconstruction

Hyungjun Doh*
Elmore Family School of Electrical and Computer Engineering
Purdue University
West Lafayette, IN, USA
hdoh@purdue.edu

Dong In Lee*†
Department of Artificial Intelligence
Korea University
Seoul, Republic of Korea
dilee99@korea.ac.kr

Seunggeun Chi*
Elmore Family School of Electrical and Computer Engineering
Purdue University
West Lafayette, IN, USA
chi65@purdue.edu

Pin-Hao Huang
Honda Research Institute USA
San Jose, CA, USA
pin-hao_huang@honda-ri.com

Kwonjoon Lee
Honda Research Institute USA
San Jose, CA, USA
kwonjoon_lee@honda-ri.com

Sangpil Kim
Department of Artificial Intelligence
Korea University
Seoul, Republic of Korea
spk7@korea.ac.kr

Karthik Ramani
Elmore Family School of Electrical and Computer Engineering
School of Mechanical Engineering
Purdue University
West Lafayette, IN, USA
ramani@purdue.edu

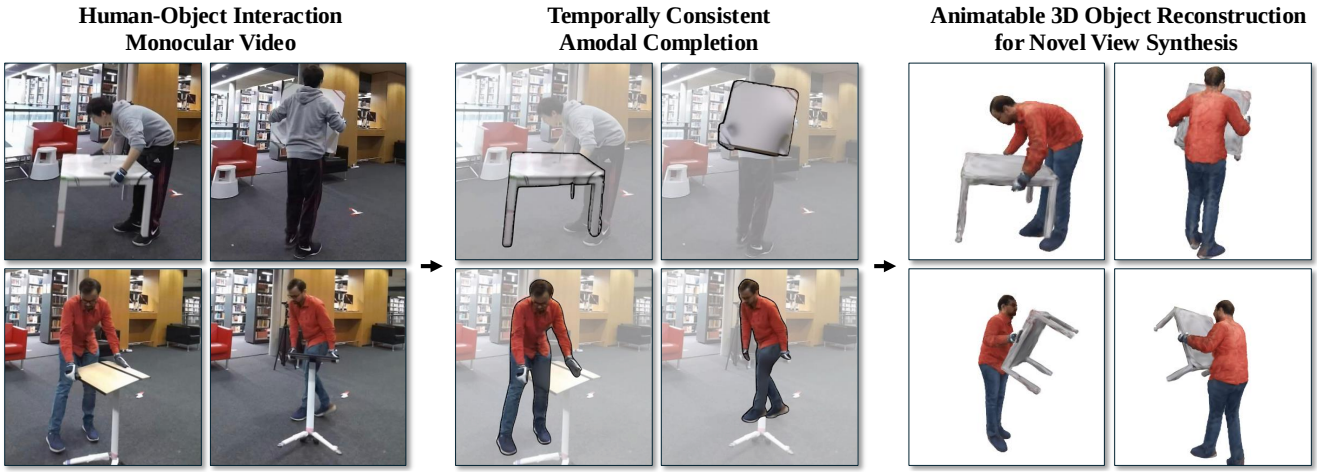

Figure 1: (*left*) In human–object interaction (HOI) scenarios, occlusions frequently affect both the human and the object. (*middle*) We inpaint the occluded regions while preserving temporal consistency for both entities across frames. (*right*) Leveraging the temporally consistent image sequences, we reconstruct the human and object using a 3D Gaussian splatting representation, enabling animatable 3D HOI applications.

Abstract

We introduce a novel framework for reconstructing dynamic human–object interactions from monocular video that overcomes challenges associated with occlusions and temporal inconsistencies. Traditional 3D reconstruction methods typically assume static objects or full visibility of dynamic subjects, leading to degraded performance when these assumptions are violated—particularly in scenarios where mutual occlusions occur. To address this, our framework leverages amodal completion to infer the complete structure

*Three authors contributed equally to this research.
†Work done at Purdue University while a visiting scholar.

MM '25, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3754769

of partially obscured regions. Unlike conventional approaches that operate on individual frames, our method integrates temporal context, enforcing coherence across video sequences to incrementally refine and stabilize reconstructions. This template-free strategy adapts to varying conditions without relying on predefined models, significantly enhancing the recovery of intricate details in dynamic scenes. We validate our approach using 3D Gaussian Splatting on challenging monocular videos, demonstrating superior precision in handling occlusions and maintaining temporal stability compared to existing techniques.

CCS Concepts

• **Computing methodologies** → **Shape inference**; Reconstruction.

Keywords

Human-Object Interaction, Amodal Completion, Temporal Consistency

ACM Reference Format:

Hyungjun Doh, Dong In Lee, Seunggeun Chi, Pin-Hao Huang, Kwonjoon Lee, Sangpil Kim, and Karthik Ramani. 2025. Occlusion-Aware Temporally Consistent Amodal Completion for 3D Human-Object Interaction Reconstruction. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3746027.3754769

1 Introduction

Understanding the complex interplay between humans and objects is a fundamental challenge in computer vision and robotics, with broad implications for both theory and real-world applications. This understanding is critical for enabling technologies such as autonomous robots that navigate complex human environments and augmented or virtual reality systems that rely on precise 3D reconstructions. Despite significant progress, reliably interpreting human-object interactions (HOI) in dynamic, real-world settings remains an open problem. Visual complexities—especially mutual occlusions—often obscure critical details, underscoring the need for enhanced 3D reconstruction techniques that capture and animate these interactions effectively.

Multi-view 3D reconstruction [15, 26, 33] methods have proven effective in generating detailed models from multiple images. Object reconstruction techniques [12, 35, 43, 50] are generally designed for static scenes, where the object remains stationary. In contrast, human reconstruction methods [9, 14, 18] typically handle dynamic subjects assuming full visibility. However, in HOI scenarios, these assumptions break down. Both humans and objects are in motion, leading to frequent and mutual occlusions as illustrated in Figure 1, resulting in ambiguous visual cues that complicate the reconstruction process. The presence of dynamic and mutual occlusions introduces complexities beyond the scope of static object or dynamic full-visibility human reconstruction methods.

Amodal completion [3, 4, 21, 28, 29, 40, 42, 49] has emerged as a promising strategy to mitigate these occlusion challenges by inferring the complete structure of partially hidden regions. However, most existing approaches are designed for static scenes. Although they perform well in controlled environments, they often falter in dynamic, real-world settings. A key limitation is their reliance on image-level completion, which overlooks the rich temporal context provided by adjacent sequences in HOI cases. Without leveraging temporal information, reconstructions derived from amodal-completed images can suffer from inconsistencies, ultimately compromising the overall quality and stability of the models.

Recent research [5, 13, 37, 38, 44, 46, 48] suggests that incorporating temporal consistency can effectively address these challenges. By enforcing coherence across consecutive video frames, models can integrate information from multiple viewpoints more effectively, leading to more accurate recovery of occluded regions. This strategy also enables the incremental refinement of hidden details over time, yielding reconstructions that more faithfully capture the dynamic nature of human–object interactions.

Building on these insights, we propose a novel, temporally consistent amodal completion framework tailored for monocular video input. By integrating temporal context into the amodal completion process, our approach overcomes the limitations of traditional frame-by-frame methods, significantly enhancing both stability and realism. Specifically, our framework consists of three key components. First, we introduce *Bidirectional Temporal Feature (BTF) Warping*, which leverages optical flow to warp latent features from both past and future frames into the current frame's latent space, enabling effective propagation of temporal cues and aligns spatially meaningful information across time. Second, we present a *Temporal Fusion Attention* mechanism that adaptively aggregates these temporally aligned features, producing a coherent and semantically enriched latent representation. Third, we propose a *Template-free Occlusion Identification* strategy that combines 2D segmentation and 3D projections to localize occlusions without relying on predefined templates. Finally, we apply amodal completion, guided by the temporal-aware latent features and precise occlusion masks, to complete missing regions with high fidelity. This unified pipeline enables robust recovery of occluded structures across time, ensuring temporal consistency and accurate amodal completion. Unlike prior approaches, our method generalizes effectively to dynamic, real-world human–object interaction scenarios.

We validate the effectiveness of our method through extensive experiments on two public datasets, BEHAVE [2] and InterCap [10, 11], which present severe occlusions and diverse human–object interactions. Our framework demonstrates consistently superior performance in both quantitative metrics and qualitative comparisons. Moreover, we extend our evaluation to reconstructing 3D human–object interactions from monocular video sequences using 3D Gaussian Splatting [15] (3DGS), enabling various 3D application of animatable HOI including novel-view synthesis.

In summary, our contributions include:

- **Temporally Consistent Completion Method:** We carefully design a Bidirectional Temporal Feature (BTF) Warping and a Temporal Fusion Attention mechanism that incorporate temporal context to amodal completion by adaptive fusion of aligned features across past and future frames.
- **Template-free Occlusion Identification Strategy:** We propose a hybrid approach that integrates 2D segmentation with 3D point cloud projection to accurately localize occluded regions. This method introduces a novel masking technique that identifies occlusions efficiently, without the need for predefined templates.

- **Application to 3D Reconstruction:** To the best of our knowledge, our work is the first approach to reconstruct photo-realistic and animatable 3D human-object interactions from monocular videos. By leveraging 3D Gaussian Splatting [15], we demonstrate that our pipeline is applicable to various subtasks, including novel-view and novel-pose synthesis for HOI.

2 Related Work

2.1 3D Reconstruction

3D reconstruction [15, 26, 33] from multi-view images has demonstrated versatility across various fields. In particular, 3D reconstruction using explicit representations [15] has emerged as an effective method for capturing 3D information. This representation has been successfully employed in 3D scene reconstruction [7, 22, 24], object reconstruction [12, 35, 43, 50], and human reconstruction [9, 14, 18]. However, these methods often face challenges when occlusions occur. To mitigate the occlusion problem in human reconstruction, few research proposed [20, 34] a diffusion-based method. However, these methods implicitly secure temporal consistency through a 2D diffusion prior, not guaranteeing temporal consistency. In contrast to existing approaches that primarily address either human or static objects, our method specifically targets cases involving human-object interactions via template-free amodal completion.

2.2 Amodal Completion

Amodal completion focuses on inferring the hidden shapes and appearances of occluded objects. Recently, various works leverage generative models to reconstruct missing regions. For instance, Chi et al. [4] completes occlusion using contact points while pix2gestalt [29] synthesizes "wholes" from partial observations and Xu et al. [42] expands occluded areas with context-awareness. Other studies propose transformer-based amodal segmentation [49], leverage 3D shape priors to guide amodal masks [3], or learn compositional scene representations to handle complex occluders [21, 28]. Self-supervised frameworks alsowemerged for amodal completion by aligning visible parts with plausible hidden geometry [16, 40]. However, these single-image methods often struggle with dynamic human-object interactions, where severe occlusions may shift from frame to frame—leading to temporally inconsistent reconstructions. As an alternatives, video inpainting methods [23] has also been introduced, however these methods are usually for removing subjects well fitted to background and environment, showing limitation on HOI cases. Our approach addresses these issues by integrating temporal cues, ensuring stable amodal completions across sequences.

2.3 Temporal Consistency

Ensuring temporal consistency across image sequences is essential for stable and coherent video processing. Traditional approaches often rely on optical flow to propagate features between frames [13, 37, 38]. Recent advancements in video diffusion have incorporated explicit temporal modeling to improve alignment and consistency across frames. Several state-of-the-art techniques [5, 44, 46, 48] adopt flow-guided recurrent architectures, enabling temporally aware representations for applications such as video super-resolution, inpainting, video-to-video translation, and text-driven video editing. Complementary strategies include transformer-based frame propagation [47] and cross-frame feature matching through diffusion-based tokens [8], both of which contribute to enhanced coherence in complex temporal dynamics. While these methods primarily focus on video generation and restoration tasks, our approach uniquely leverages optical flow and feature warping mechanisms to specifically address occlusions and ensure temporally consistent reconstructions in 3D HOI scenarios.

3 Method

Our goal is to achieve temporal consistency while completely recovering occluded regions for accurate 3D reconstruction of Human–Object Interactions (HOI) from monocular video. This is challenging due to frequent and severe occlusions in HOI scenarios. To address this challenge, we propose a novel pipeline that integrates temporal context into the amodal completion process.

We begin by formulating the temporally consistent amodal completion problem, clearly defining the inputs, outputs, and objectives in Section 3.1. Based on this formulation, our method consists of the following four main components, as illustrated in Figure 2. First, in Section 3.2, we introduce Bidirectional Temporal Feature (BTF) warping, which aligns features across neighboring frames using optical flow and latent feature warping. Second, in Section 3.3, we present Temporal Fusion Attention, a mechanism that dynamically fuses the temporally aligned features into a coherent representation. Third, in Section 3.4, we describe a template-free occlusion identification strategy that accurately localizes occluded regions. The last component, explained in Section 3.5, is temporally-aware amodal completion mechanism. Consequently, our method enables reliable 3D reconstruction through temporally consistent amodal completion, as detailed in the supplementary materials.

3.1 Problem Definition

We address the problem of filling in occluded regions in a sequence of video frames such that the completed regions are both semantically plausible and temporally coherent. For each frame, let $I_{in} \in \mathbb{R}^{H\times W\times 3}$ be the segmented image containing only the visible regions, and let $M_{in} \in \{0,1\}^{H\times W}$ be a binary mask, where 1 indicates the occluded regions to be inpainted. In addition, we provide a temporally-aware latent feature z_{in}, aggregated from temporally aligned features across neighboring frames, and a text prompt P as textual guidance. Adapting the problem definition of [42], we formalize the inpainting process with the diffusion model [32] guided by both latent and mask inputs as:

$$I_{out} = \mathbf{F}_{s\to e}\left(I_{in},\, M_{in},\, z_{in},\, P\right), \tag{1}$$

where $\mathbf{F}_{s\to e}$ denotes the diffusion process from denoising time step s to e. The effectiveness of our framework critically depends on two factors: (1) the quality of the latent feature z_{in}, which provides temporally consistent and semantically rich context, and (2) the accuracy of the occlusion mask M_{in}, which precisely localizes the region to be completed. In the following sections, we describe how each component is constructed to ensure effective amodal completion under complex human–object interaction scenarios.

3.2 Bidirectional Temporal Feature Warping

To ensure temporal consistency across consecutive frames, we introduce a Bidirectional Temporal Feature (BTF) warping method that aligns latent features from both past and future frames to

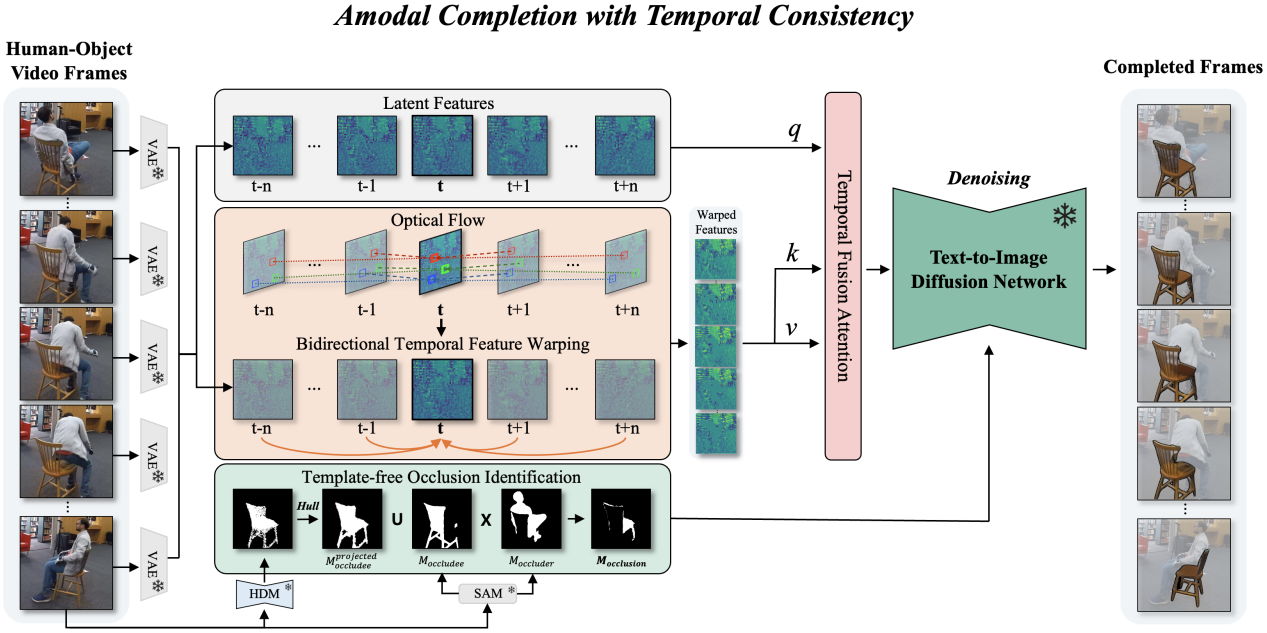

Figure 2: Overview of Our Framework: Given a human–object interaction (HOI) monocular video, our framework performs amodal completion through (1) Bidirectional Temporal Feature Warping via optical flow, (2) Temporal Fusion Attention for aggregating multi-frame context, (3) Template-free Occlusion Identification using 2D and 3D cues, and (4) temporally-aware amodal completion. This design enables temporally consistent and accurate amodal completion in complex HOI scenarios.

the current frame. As part of this approach, we define a temporal support set $\mathcal{S}(t)$ at each time step t, consisting of past and future neighboring frames:

$$\mathcal{S}(t) = \{t-n, \ldots, t-1, t+1, \ldots, t+n\}. \tag{2}$$

We empirically choose $n = 7$ aggregating total 14 neighboring features. For each element in the support set $\tau \in \mathcal{S}(t)$, we estimate optical flow from frame I^τ to the reference frame I^t to spatially align their latent representations. The optical flow is computed as:

$$o_\tau^t = \text{Flow}(I^\tau, I^t), \quad \forall\, \tau \in \mathcal{S}(t), \tag{3}$$

where $\text{Flow}(\cdot,\cdot)$ denotes an off-the-shelf optical flow estimation network [38].

We then encode latent features using a pretrained Variational Autoencoder (VAE) [17] $\mathcal{E}$ as utilized in diffusion models [32], defined as $z^\tau = \mathcal{E}(I^\tau)$. To spatially align the latent representations across frames, the encoded feature z^τ is subsequently warped into the coordinate space of frame t using the estimated optical flow o_τ^t which is bilinearly scaled to match the resolution of z^τ:

$$\hat{z}_\tau^t = \text{Warp}(z^\tau, o_\tau^t), \quad \forall\, \tau \in \mathcal{S}(t). \tag{4}$$

The warped features $\hat{z}_\tau^t$ serve as temporally aligned observations centered at frame I^t, enabling the aggregation of multi-frame context from both past and future frames. By referencing complementary information from adjacent frames, our framework can recover regions that are occluded or missing in the current view in complex human-object interaction scenarios.

3.3 Temporal Fusion Attention

While the warped features are temporally aligned, it is crucial to effectively incorporate temporal context from each warped feature into a unified latent representation. Therefore, we propose a Temporal Fusion Attention mechanism that selectively integrates complementary cues from neighboring frames. Specifically, we employ a cross-attention strategy to aggregate temporally aligned latent features within the support set $\mathcal{S}(t)$. At each time step t, the queries, keys, and values are defined as:

$$Q^t = z^t, \quad K^t = V^t = \text{concat}(\hat{z}_{t-n}^t, \cdots, \hat{z}_{t+n}^t), \tag{5}$$

where $Q^t \in \mathbb{R}^{1\times d}$ represents the latent feature from the current frame, and $K^t \in \mathbb{R}^{2n\times d}$, $V^t \in \mathbb{R}^{2n\times d}$ represent the warped latent features aligned from adjacent frames. The attention-weighted representation is computed via scaled dot-product attention:

$$z_{fused}^t = \text{Attn}(Q^t, K^t, V^t) = \text{Softmax}\left(\frac{Q^t \cdot K^t}{\sqrt{d}}\right) V^t, \tag{6}$$

where d denotes the feature dimensionality. By leveraging temporally warped features from both past and future frames, this mechanism selectively integrates complementary information that is not visible in the current view, thereby improving occlusion completion. This yields z_{fused}^t, a semantically rich context that captures occluded structures and interaction context from neighboring frames. By emphasizing temporally relevant and contextually informative features, this attention mechanism significantly enhances coherence and robustness across frames, thereby facilitating high-quality and stable amodal completion.

3.4 Template-free Occlusion Identification

To accurately localize the occluded area, it is essential to estimate the complete shape of the target—referred to as the occludee—which is the object of interest partially hidden in the scene. Therefore, We employ the Hierarchical Diffusion Model (HDM) [41], a template-free approach specifically designed for reconstructing human-object

interactions. HDM facilitates the inference of a comprehensive 3D point cloud representing the complete geometry of human and object. We then project this 3D point cloud onto the 2D image plane using a projection function $\pi : \mathbb{R}^3 \rightarrow \mathbb{R}^2$. The resulting set of projected points is denoted as $C = \{\mathbf{u}_i \in \mathbb{R}^2 \mid \mathbf{u}_i = \pi(\mathbf{U}_i), \mathbf{U}_i \in \mathbb{R}^3,\ i = 1, 2, \ldots, n\}$, where each $\mathbf{u}_i = (x_i, y_i)$ corresponds to the 2D coordinates of the projection of the 3D point $\mathbf{U}_i$. This set C thus represents the 2D image plane positions of all projected points from the original 3D geometry. To extract a coherent shape from these points, we apply a concave hull algorithm [1] that constructs a tight boundary around the set. Unlike a convex hull, which encloses the outermost points with the smallest convex polygon, the concave hull denoted as ConcaveHull$(\cdot)$ allows for inward indentations, enabling the capture of non-convexities and finer geometric details. This leads to a more accurate approximation of the object's shape. The projected full shape mask of the target is then defined as:

$$M_{occludee}^{projected} = \text{ConcaveHull}(C). \tag{7}$$

To ensure a more robust estimation of the target's shape, we construct a fused mask, M_{union}, by combining the segmented visible mask $M_{occludee}^{visible}$, obtained by SAM2 [31], with the projected full shape mask $M_{occludee}^{projected}$:

$$M_{union} = M_{occludee}^{visible} \cup M_{occludee}^{projected}, \tag{8}$$

which allows the two masks to effectively complement one another, integrating both observed and inferred regions of the target. The Figure 2 demonstrates how the visible and projected masks are combined to yield a more complete representation.

Finally, we introduce an occluder mask $M_{occluder}$, also segmented using SAM2, representing the region that occludes the target. The occluder can be either a human or an object, depending on the target of interest. The actual occlusion mask $M_{occlusion}$, which delineates the occluded area of the target, is determined by intersecting the fused mask M_{union} with the occluder mask $M_{occluder}$:

$$M_{occlusion} = M_{union} \cap M_{occluder}. \tag{9}$$

With this approach, we precisely identify the occluded region of the target object without relying on any predefined templates, significantly enhancing the generalizability and robustness.

3.5 Temporally-aware Amodal Completion

With temporally fused features z_{fused}^t and the occlusion regions $M_{occlusion}^t$ identified in previous stages, we inpaint the occluded region of the target using the diffusion inpainting pipeline. To achieve temporally consistent amodal completion, we first apply occludee mask $M_{occludee}^t$ to the input image I thereby removing background distractions and encouraging the model to focus on the target:

$$I_{occludee}^t = M_{occludee}^t \odot I^t. \tag{10}$$

Next, we use the occlusion mask $M_{occlusion}^t$ to specify the region that should be completed. To incorporate temporal context, we leverage a temporally-aware latent feature z_{fused}^t, applying mask $\hat{M}_{union}^t$ to ensure that latent representation only affects the target. Note that $\hat{M}_{union}^t$ is bilinearly downsampled mask to match the dimension of z_{fused}^t. A corresponding text prompt P is also provided

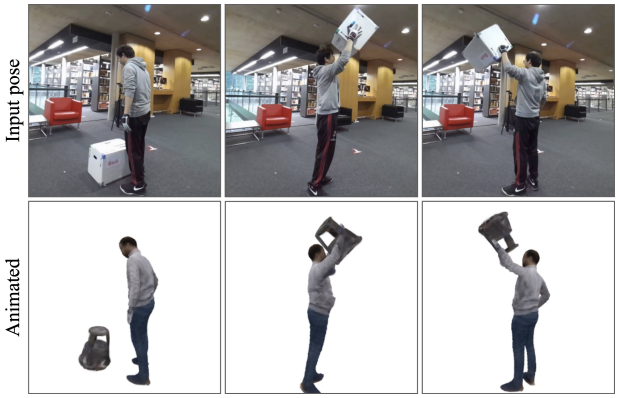

Figure 3: Animatable 3D Reconstruction of Human–Object Interactions: we train 3DGS [15] using temporally consistent amodal completions from our pipeline, sampled at 1 fps (15–47 frames per video). Conditioned on motion trajectories from the input video, our method enables realistic animation of novel human–object pairs while preserving geometry, appearance, and temporal coherence.

to guide the inpainting process. This process can be written as:

$$I_{out}^t = \mathbf{F}_{s \rightarrow e}(I_{occludee}^t,\ M_{occlusion}^t,\ z_{fused}^t \odot \hat{M}_{union}^t,\ P). \tag{11}$$

Through the integration of temporal cues, the diffusion inpainting pipeline reliably fills in the occluded regions while maintaining temporal consistency across frames. Consequently, our framework effectively resolves ambiguities due to occlusions and complex dynamics inherent in realistic human–object interactions, improving the quality and stability of amodal completion outcomes.

3.6 Application in 3D Reconstruction

To demonstrate the utility of our temporally consistent amodal completion in downstream tasks, we reconstruct photo-realistic and animatable 3D human–object interaction scenes from monocular video using 3D Gaussian Splatting (3DGS) [15]. Our pipeline produces temporally inpainted frames I_{out}^t, which serve as dense appearance supervision even under severe occlusions.

For object reconstruction, we follow a GS-Pose [25], using provided 6-DoF object and camera poses. To ensure consistent scale and visibility, each object in image is cropped and re-centered in every frame $\tilde{I}_{out}^t$. A 3D Gaussian representation $\mathcal{G}_{obj}$ is then optimized by minimizing a photometric loss between rendered images and inpainted frames:

$$\mathcal{L}_{obj} = \sum_t \mathcal{L}_{photo}\left(R(\mathcal{G}_{obj}, H_t \circ X_t),\ \tilde{I}_{out}^t\right), \tag{12}$$

where $R(\cdot)$ denotes the Gaussian renderer, and H_t, X_t are the camera and object poses. The $\mathcal{L}_{photo}$ is a combination of L1 and SSIM terms:

$$\mathcal{L}_{photo}(I_r, I_{gt}) = \|I_r - I_{gt}\|_1 + \lambda \cdot \left(1 - \text{SSIM}(I_r, I_{gt})\right), \tag{13}$$

where I_r is the rendered image, I_{gt} is the ground truth image, and λ controls the balance between pixel-wise and perceptual similarity. In our experiments, we set $\lambda = 0.2$ to balance these terms effectively.

For human reconstruction, we adopt GaussianAvatar [9], which learns a canonical 3D Gaussian representation and a pose-conditioned deformation field based on SMPL parameters. The model is trained with the same inpainted sequences, allowing robust geometry and appearance recovery even under partial occlusions.

Method	BEHAVE [2]				InterCap [10, 11]			
	Amodal Completion		Temporal Consistency		Amodal Completion		Temporal Consistency	
	IoU ↑	CLIP ↑	Warp-err (x10^{-3}) ↓	TC Score ↑	IoU ↑	CLIP ↑	Warp-err (x10^{-3}) ↓	TC Score ↑
Pix2gestalt [29]	61.29%	26.77	141.08	95.97	64.91%	25.32	112.75	95.83
SD Inpainting [32]	60.81%	27.63	9.87	96.78	45.16%	27.23	7.04	96.71
LaMa [36]	43.54%	26.08	6.34	96.96	56.42%	26.02	3.78	**98.01**
VDT [23]	55.69%	26.48	**5.84**	96.58	59.96%	26.11	3.42	96.39
Ours (HDM Mask)	**61.75**%	**27.64**	<u>6.26</u>	**97.19**	**70.18**%	**27.65**	**3.22**	<u>96.95</u>
Ours (Ground Truth Mask)	70.90%	28.01	6.09	98.15	74.79%	27.66	2.93	97.98

Table 1: Quantitative Comparison on BEHAVE [2] and InterCap [10, 11] for Amodal Completion and Temporal Consistency: Our method achieves consistently strong performance, highlighting its robustness to occlusion and temporal challenges. Bold and underline denote the best and second-best scores.

Together, as visualized in Figure 3, these reconstructions validate that our temporally consistent amodal completion provides a reliable and expressive foundation for photo-realistic, animatable 3D human-object modeling from monocular video.

4 Experiment

4.1 Datasets

BEHAVE The BEHAVE dataset [2] consists of 321 RGB-D sequences capturing indoor human–object interactions, recorded using 4 Kinect cameras. The test set includes 3 subjects interacts with 20 objects. Among these, we select 18 videos for 18 objects excluding keyboard and basketball due to the lacking of ground truth pose annotation in 30 fps video sequence. For human, we selected 3 videos for 3 subjects in the test set. As a results, we applied our pipeline to approximately 27,000 frames and evaluate these results.

InterCap InterCap [10, 11] contains 223 RGB-D videos of human-object interactions, captured from 6 distinct viewpoints with 10 subjects and 10 objects. From this dataset, we extract 10 videos that collectively represent all 10 objects.

4.2 Evaluation Metrics

Amodal Completion We evaluate the performance of our amodal completion using two metrics: the CLIP score [30] and Intersection over Union (IoU). The CLIP score evaluates the alignment between the generated images and the corresponding category prompts, while IoU measures the overlap between the predicted and groundtruth amodal masks. Specifically, we calculate the CLIP score for each inpainted frame within tight bounding boxes to minimize the influence of background pixels. For the IoU evaluation, we segment the inpainted frame using SAM2 [31] to obtain masks and then calculate the overlap with the object's groundtruth masks.

Temporal Consistency We assess temporal consistency using two metrics: the Temporal Consistency score (TC score) [6] and the Flow Warping Error [19]. The TC score is computed by extracting CLIP image embeddings for all output video frames and calculating the average cosine similarity between pairs of consecutive frames. Flow Warping Error, commonly used in optical flow estimation and video processing, measures the discrepancy between a warped image and its corresponding target frame. In our approach, we compute optical flows between consecutive ground-truth object masks using SEA-RAFT [38]. We then warp the previously inpainted frame to align with the current frame, and measure the discrepancy between this warped previous frame and the actual current frame. A higher Temporal Consistency Score and a lower Flow Warping Error both indicate better temporal consistency.

3D Reconstruction Since our method reconstructs only the object or the human—excluding the background—we evaluate reconstruction quality using masked versions of standard metrics: Peak Signal-to-Noise Ratio (PSNR-M), Structural Similarity Index Measure (SSIM-M) [39], and Learned Perceptual Image Patch Similarity (LPIPS-M) [45]. Inspired by [27], PSNR-M, SSIM-M and LPIPS-M are calculated within tight bounding boxes around the reconstructed regions to minimize the influence of background pixels.

4.3 Temporally Consistent Amodal Completion

Quantitative Results We quantitatively evaluate our approach against several state-of-the-art baselines, including Pix2gestalt [29], Stable Diffusion Inpainting [32], LaMa [36], and VDT [23], initialized with random noise. The results are summarized in Table 1, covering both human and object categories, as the human is treated as one of 19 object classes in our experiments, and only frames with an object occlusion ratio between 15% and 70% are considered. Our method consistently achieves the best performance across amodal completion metrics, IoU and CLIP, on both BEHAVE and InterCap. This indicates a higher fidelity in reconstructing occluded regions, validating the effectiveness of our template-free regioning strategy in complex HOI cases. To assess ideal condition performance, we report results using ground-truth masks in the last row.

Regarding temporal consistency, our method achieves the lowest warping error among all methods, reflecting superior temporal alignment between consecutive frames. This highlights our model's strength in preserving temporal coherence throughout the video sequence. Interestingly, although LaMa reports the highest TC score, it ranks lowest in IoU and performs poorly in terms of visual quality. This may be attributed to its tendency to overwrite masked regions with background-matching content, which is especially beneficial in static-camera datasets like BEHAVE and InterCap. In contrast, our method reconstructs occlusions by explicitly leveraging spatial-temporal information from adjacent frames, leading to more semantically faithful and temporally robust completions.

Qualitative Results Figure 4 presents a qualitative comparison across various methods on three representative HOI categories:

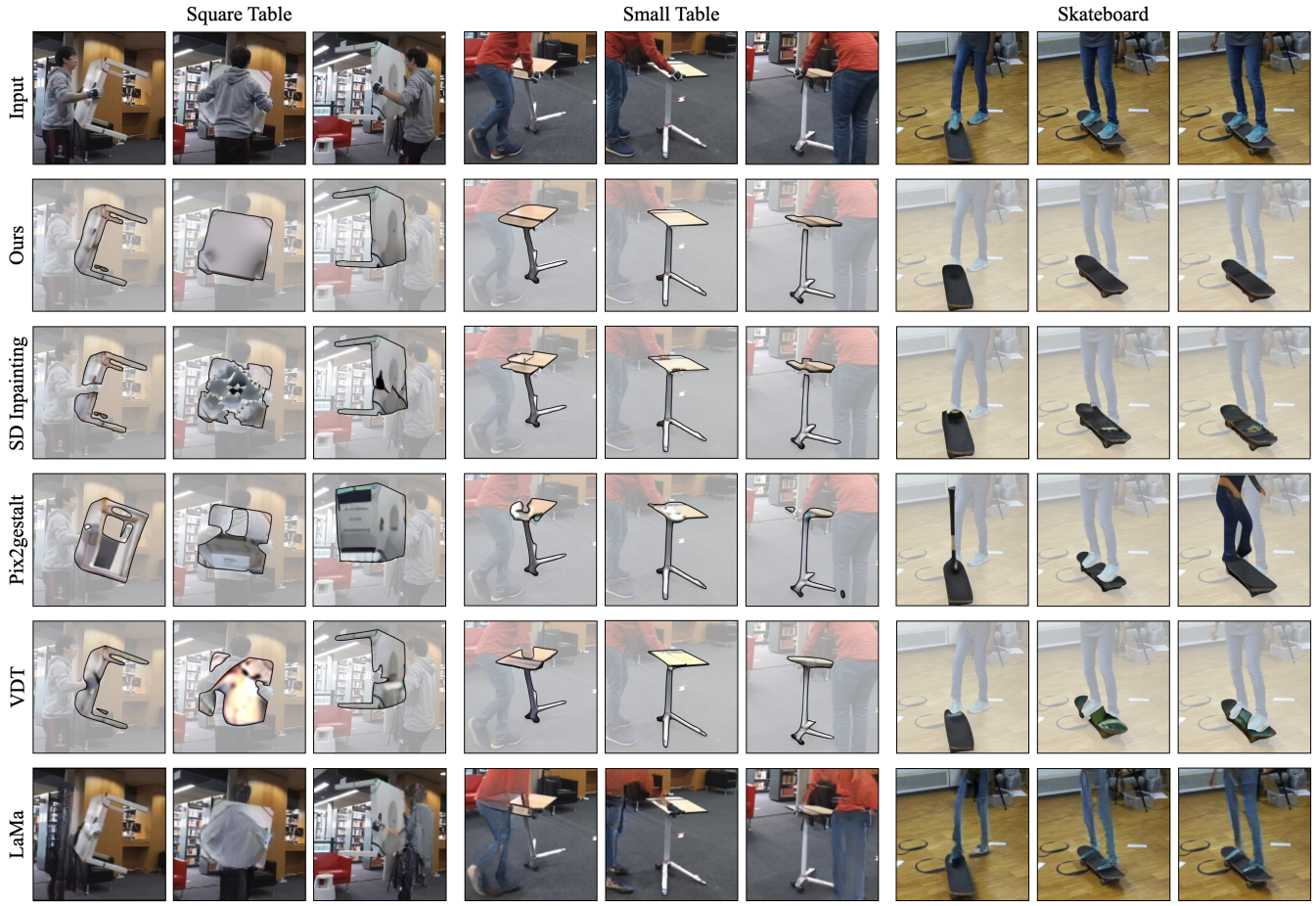

Figure 4: Qualitative comparison on BEHAVE [2] (Square Table, Small Table) and InterCap [10, 11] (Skateboard). Our method produces accurate and temporally consistent completions of occluded regions.

Square Table, *Small Table*, and *Skateboard*. Compared to baseline methods, our approach produces semantically faithful and visually coherent reconstructions, especially in the occluded regions. For *Square Table*, Pix2gestalt generates the implausible shape and VDT fails to complete the proper region. In contrast, our method preserves object integrity with precise geometry and sharper texture completion. In the *Skateboard* scenario, our method excels in preserving object shape and continuity under occlusion. Competing methods often hallucinate incorrect geometry or texture, especially under relatively simple occlusions caused by human limbs. Overall, the visual results confirm that our method outperforms prior approaches in generating amodally completed outputs that are both spatially and temporally coherent, reinforcing the strength of our template-free, temporally consistent completion framework.

4.4 3D Reconstruction

To the best of our knowledge, this is the first method to enable animatable and photo-realistic 3D reconstruction in human–object interaction (HOI) scenarios through temporally consistent amodal completion. We use the completed monocular video frames generated by our pipeline to train 3D Gaussian Splatting [15] (3DGS), enabling high-quality reconstruction from monocular inputs. For training, we extract 40 to 47 frames per sequence at 1 fps, a sparse set of inpainted frames.

We compare three settings: (1) our full pipeline with temporally-aware amodal completion, (2) Stable Diffusion inpainting without our temporal modules, and (3) the original input sequence without any completion. As shown in Table 2, our method achieves the highest PSNR-M, SSIM-M and LPIPS-M, outperforming all comparison methods. As illustrated in Figure 5, our approach improves both geometric accuracy and temporal consistency, which are crucial for smooth and realistic novel-view rendering. Furthermore, when conditioned on SMPL body poses or 6D object trajectories, the trained model produces photo-realistic animations of both humans and objects, highlighting the robustness and completeness of our pipeline. In contrast, without our temporally consistent amodal completion, occluded regions remain unresolved and frame-to-frame inconsistencies persist, leading to degraded 3D reconstruction quality.

4.5 Ablation Study

To assess the contributions of each component in our temporal consistency framework, we conduct ablation studies on the BEHAVE dataset, as shown in Table 3. We analyze the impact of two key components: feature warping and cross-frame attention. Without

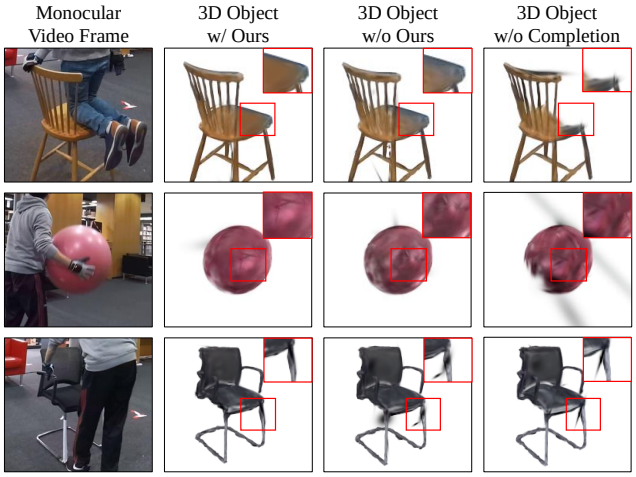

Figure 5: Qualitative results of 3D Reconstruction: Our method (left) produces clearer geometry and fewer artifacts than without our method (middle) or no completion (right).

Method	PSNR-M ↑	SSIM-M ↑	LPIPS-M ↓
w/o Completion	13.48	0.5679	0.3604
SD Inpainting [32]	15.38	0.6186	0.2455
Ours	**15.43**	**0.6240**	**0.2383**

Table 2: Quantitative results on 3D reconstruction from the BEHAVE dataset [2]: We report the metrics, computed within tight bounding boxes around humans or objects.

Feature Warp.	Cross Attn.	IoU ↑	CLIP ↑	Warp-err (x10^{-3}) ↓	TC ↑
✗	✗	60.81	27.63	9.87	96.78
✗	✓	60.82	27.64	6.56	97.06
✓	✗	58.35	27.60	**6.23**	97.16
✓	✓	**61.75**	**27.64**	<u>6.26</u>	**97.19**

Table 3: Ablation Studies on temporal consistency strategy for on BEHAVE dataset [2]. Bold and underline denote the best and second-best scores.

either feature warping or cross attention (first row), the model performs reasonably well, but shows higher warping error and lower temporal consistency. Introducing cross attention alone (second row) slightly improves all metrics, particularly reducing the warping error. Interestingly, feature warping alone (third row) leads to the lowest warping error (6.23) and improved TC score, demonstrating its effectiveness in aligning features across time. However, it degrades amodal completion accuracy (IoU). Combining both feature warping and cross attention (last row) yields the best overall performance, achieving the highest IoU (61.75), CLIP score (27.64), and TC score (97.19). This confirms that both components are complementary and essential for generating temporally consistent and semantically meaningful completions.

We conduct an additional ablation study to evaluate the effectiveness of our template-free occlusion identification strategy, as shown in Table 3. Specifically, we compare our method against a baseline that relies on predefined human masks to guide the inpainting process. Our approach derives the occlusion mask $M_{\text{occlusion}}$ by combining 2D segmentation with 3D point cloud projection, leading to significantly better performance in terms of IoU (61.75 vs. 53.77) and warping error (6.26 vs. 8.90). These results highlight our method's superior ability to accurately localize occluded regions and maintain temporal coherence. Although the human-mask baseline yields a slightly higher CLIP score and TC Score, this is likely due to its tendency to inpaint overly large regions, which degrades spatial accuracy and temporal consistency, as reflected in the lower IoU and higher warping error. Qualitative results of the ablation study are provided in the supplementary material.

Mask	IoU ↑	CLIP ↑	Warp-err (x10^{-3}) ↓	TC ↑
Human Mask	53.77	**28.31**	8.90	**97.60**
Ours ($M_{occlusion}$)	**61.75**	27.64	**6.26**	97.19

Table 4: Ablation Studies on Template-free Occlusion Identification strategy on BEHAVE dataset [2].

5 Discussion and Limitations

Dynamic human–object interactions present significant challenges for optical flow-based shape estimation, particularly due to nonlinear motion and frequent occlusions. Our approach leverages HDM for amodal completion, enabling the reconstruction of full object shapes—including occluded regions. However, the method may not always accurately infer the geometry of unseen parts. Future directions include integrating multi-view geometry techniques, such as Structure-from-Motion (SfM) and Multi-View Stereo (MVS), to densify and refine reconstructions, thereby enhancing the robustness of our pipeline.

Currently, our method assumes that each input video contains only a single human and a single object. While effective for controlled scenarios, this assumption limits applicability in more complex settings, such as multi-person interactions or interactions involving multiple objects. Extending the framework to handle such cases is a promising direction for future work.

Additionally, our method relies heavily on the inpainting capabilities of the Stable Diffusion model, which may struggle to accurately reconstruct objects or appearances that were not well represented during its training. We also observed that the quality of inpainting is highly sensitive to the accuracy of the occlusion masks. In particular, when the masks fail to fully exclude occluding regions—leaving behind small residual artifacts—the inpainting output is noticeably degraded. These limitations highlight the critical importance of precise occlusion localization, especially in scenes involving fine-grained interactions or partial visibility.

6 Conclusion

We proposed a novel framework for reconstructing dynamic human–object interactions from monocular video, with a focus on addressing occlusions and temporal inconsistency. Our method achieves state-of-the-art performance across multiple benchmarks, demonstrating both quantitative superiority and qualitative robustness. It also generalizes well to real-world scenarios. Notably, by integrating our approach with 3D Gaussian Splatting, we enable the generation of photorealistic and animatable 3D reconstructions from monocular input, supporting advanced applications such as novel-view and novel-pose synthesis in complex HOI scenes.

Acknowledgments

We wish to thank all the reviewers for their invaluable feedback. This work is partially supported by the NSF under the Future of Work at the Human-Technology Frontier (FW-HTF) 1839971 and Partnership for Innovation: Technology Transfer (PFI-TT) 2329804. Additional support for this work is provided by the Culture, Sports, and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2024 (International Collaborative Research and Global Talent Development for the Development of Copyright Management and Protection Technologies for Generative AI, RS-2024-00345025), Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2019-II190079, Artificial Intelligence Graduate School Program (Korea University). We also acknowledge the Feddersen Distinguished Professorship Funds and a gift from Thomas J. Malott. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the funding agency.

References

[1] K. Bellock. [n. d.]. alphashape. https://github.com/bellockk/alphashape. GitHub repository, accessed 2025-04-11.

[2] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. 2022. Behave: Dataset and method for tracking human object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15935–15946.

[3] A. Chen, B. Smith, and C. Lee. 2023. Amodal 3D Shape from Partial Views. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 4567–4576.

[4] Seunggeun Chi, Enna Sachdeva, Pin-Hao Huang, and Kwonjoon Lee. 2025. Contact-Aware Amodal Completion for Human-Object Interaction via Multi-Regional Inpainting. arXiv:2508.00427 [cs.CV] https://arxiv.org/abs/2508.00427

[5] Yuren Cong, Mengmeng Xu, Christian Simon, Shoufa Chen, Jiawei Ren, Yanping Xie, Juan-Manuel Perez-Rua, Bodo Rosenhahn, Tao Xiang, and Sen He. 2023. Flatten: optical flow-guided attention for consistent text-to-video editing. *arXiv preprint arXiv:2310.05922* (2023).

[6] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. 2023. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*. 7346–7356.

[7] Yang Fu, Sifei Liu, Amey Kulkarni, Jan Kautz, Alexei A. Efros, and Xiaolong Wang. 2024. COLMAP-Free 3D Gaussian Splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 20796–20805.

[8] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. 2023. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373* (2023).

[9] Liangxiao Hu, Hongwen Zhang, Yuxiang Zhang, Boyao Zhou, Boning Liu, Shengping Zhang, and Liqiang Nie. 2024. Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 634–644.

[10] Yinghao Huang, Omid Taheri, Michael J Black, and Dimitrios Tzionas. 2022. InterCap: Joint markerless 3D tracking of humans and objects in interaction. In *DAGM German Conference on Pattern Recognition*. Springer, 281–299.

[11] Yinghao Huang, Omid Taheri, Michael J Black, and Dimitrios Tzionas. 2024. InterCap: Joint Markerless 3D Tracking of Humans and Objects in Interaction from Multi-view RGB-D Images. *International Journal of Computer Vision* (2024), 1–16.

[12] Ajay Jain, Matthew Tancik, and Pieter Abbeel. 2021. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5885–5894.

[13] Jisoo Jeong, Jamie Menjay Lin, Fatih Porikli, and Nojun Kwak. 2022. Imposing consistency for optical flow estimation. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*. 3181–3191.

[14] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. 2022. Neuman: Neural human radiance field from a single video. In *European Conference on Computer Vision*. Springer, 402–418.

[15] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics* 42, 4 (July 2023). https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/

[16] M. Kim, J. Park, and K. Lee. 2023. Monocular Differentiable Rendering for Self-Supervised 3D Amodal Masks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 789–798.

[17] Diederik P Kingma, Max Welling, et al. 2013. Auto-encoding variational bayes.

[18] Muhammed Kocabas, Jen-Hao Rick Chang, James Gabriel, Oncel Tuzel, and Anurag Ranjan. 2024. Hugs: Human gaussian splats. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 505–515.

[19] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. 2018. Learning blind video temporal consistency. In *Proceedings of the European conference on computer vision (ECCV)*. 170–185.

[20] Inhee Lee, Byungjun Kim, and Hanbyul Joo. 2024. Guess the unseen: Dynamic 3d scene reconstruction from partial 2d glimpses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1062–1071.

[21] P. Li, Q. Zhang, and R. Others. 2022. Compositional Models for Amodal Layout Completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2345–2354.

[22] Jiaqi Lin, Zhihao Li, Xiao Tang, Jianzhuang Liu, Shiyong Liu, Jiayue Liu, Yangdi Lu, Xiaofei Wu, Songcen Xu, Youliang Yan, et al. 2024. Vastgaussian: Vast 3d gaussians for large scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5166–5175.

[23] Haoyu Lu, Guoxing Yang, Nanyi Fei, Yuqi Huo, Zhiwu Lu, Ping Luo, and Mingyu Ding. 2023. Vdt: General-purpose video diffusion transformers via mask modeling. *arXiv preprint arXiv:2305.13311* (2023).

[24] Tao Lu, Mulin Yu, Linning Xu, Yuanbo Xiangli, Limin Wang, Dahua Lin, and Bo Dai. 2024. Scaffold-gs: Structured 3d gaussians for view-adaptive rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20654–20664.

[25] Jilan Mei, Junbo Li, and Cai Meng. 2024. GS2Pose: Tow-stage 6D Object Pose Estimation Guided by Gaussian Splatting. *arXiv preprint arXiv:2411.03807* (2024).

[26] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* 65, 1 (2021), 99–106.

[27] Michal Nazarczuk, Thomas Tanay, Sibi Catley-Chandar, Richard Shaw, Radu Timofte, and Eduardo Pérez-Pellitero. 2024. AIM 2024 sparse neural rendering challenge: Dataset and benchmark. *arXiv preprint arXiv:2409.15041* (2024).

[28] H. Nguyen, T. Davis, and X. Xu. 2022. Learning Disentangled Shape-Texture for Amodal Completion. In *Advances in Neural Information Processing Systems (NeurIPS)*. 1–12.

[29] Ege Ozguroglu, Ruoshi Liu, Dídac Surís, Dian Chen, Achal Dave, Pavel Tokmakov, and Carl Vondrick. 2024. pix2gestalt: Amodal segmentation by synthesizing wholes. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, 3931–3940.

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR, 8748–8763.

[31] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. 2024. SAM 2: Segment Anything in Images and Videos. *arXiv preprint arXiv:2408.00714* (2024). https://arxiv.org/abs/2408.00714

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

[33] Johannes L Schonberger and Jan-Michael Frahm. 2016. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4104–4113.

[34] Adam Sun, Tiange Xiang, Scott Delp, Fei-Fei Li, and Ehsan Adeli. 2024. Occfusion: Rendering occluded humans with generative diffusion priors. *Advances in Neural Information Processing Systems* 37 (2024), 92184–92209.

[35] Cheng Sun, Min Sun, and Hwann-Tzong Chen. 2022. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5459–5469.

[36] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. 2022. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2149–2159.

[37] Z. Teed and J. Deng. 2020. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. In *European Conference on Computer Vision (ECCV)*. 402–419.

[38] Yihan Wang, Lahav Lipson, and Jia Deng. 2024. Sea-raft: Simple, efficient, accurate raft for optical flow. In *European Conference on Computer Vision*. Springer, 36–54.

[39] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.

[40] J. Wu, Z. Yang, and H. Kim. 2022. Self-Supervised Amodal Reconstruction from Single Images. In *European Conference on Computer Vision (ECCV)*. 341–356.
[41] Xianghui Xie, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. 2024. Template Free Reconstruction of Human-object Interaction with Procedural Interaction Generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
[42] Katherine Xu, Lingzhi Zhang, and Jianbo Shi. 2024. Amodal completion via progressive mixed context diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9099–9109.
[43] Chen Yang, Sikuang Li, Jiemin Fang, Ruofan Liang, Lingxi Xie, Xiaopeng Zhang, Wei Shen, and Qi Tian. 2024. GaussianObject: High-Quality 3D Object Reconstruction from Four Views with Gaussian Splatting. *ACM Transactions on Graphics (TOG)* 43, 6 (2024), 1–13.
[44] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. 2023. Rerender a video: Zero-shot text-guided video-to-video translation. In *SIGGRAPH Asia 2023 Conference Papers*. 1–11.
[45] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 586–595.
[46] Zhixing Zhang, Bichen Wu, Xiaoyan Wang, Yaqiao Luo, Luxin Zhang, Yinan Zhao, Peter Vajda, Dimitris Metaxas, and Licheng Yu. 2024. Avid: Any-length video inpainting with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7162–7172.
[47] Shangchen Zhou, Chongyi Li, Kelvin CK Chan, and Chen Change Loy. 2023. Propainter: Improving propagation and transformer for video inpainting. In *Proceedings of the IEEE/CVF international conference on computer vision*. 10477–10486.
[48] Shangchen Zhou, Peiqing Yang, Jianyi Wang, Yihang Luo, and Chen Change Loy. 2024. Upscale-a-video: Temporal-consistent diffusion model for real-world video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2535–2545.
[49] X. Zhou, Y. Li, Z. Wang, and T. Others. 2023. Amodal Instance Segmentation with Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1234–1243.
[50] Zehao Zhu, Zhiwen Fan, Yifan Jiang, and Zhangyang Wang. 2024. Fsgs: Real-time few-shot view synthesis using gaussian splatting. In *European conference on computer vision*. Springer, 145–163.