Virtual Reality & Intelligent Hardware

•Article•

October 2026 Vol. 8 No. 5: 596–618
DOI: 10.1016/j.vrih.2026.08.002

Learning novel view synthesis for human from image pairs

Kaijin ZHAO, Chuyue ZHAO, Zhiheng JIANG, Guoqing ZHOU, Qing WANG*

School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China

Received 18 March 2026; **Revised** 16 June 2026; **Accepted** 2 August 2026

Abstract: This paper endeavors to address the challenging task of robust novel view synthesis (NVS) for human subjects using sparse image pairs. Previous generalizable NVS methods often struggle to mitigate accumulated errors from depth estimation. Furthermore, approaches relying on parametric human priors are typically ill-equipped to handle human-object interactions (e.g., playing instruments) and falter when presented with partial-body observations. To overcome these limitations, we propose a robust and generalizable method for human-centric novel view synthesis using only a pair of images as input. By introducing an explicit error correction mechanism, our approach simultaneously enhances both the stability and fidelity of the synthesized views. Specifically, our method first employs a hybrid depth estimation strategy, integrating monocular depth features with stereo matching information to generate a more accurate and robust depth map. Subsequently, we design an attention-based feature fusion module to extract valid information and filter out artifacts caused by depth inaccuracies, effectively fusing the warped input image pairs into a high-quality initial rendering. Finally, we introduce a dual-branch compositing mechanism that leverages a coarse yet highly stable plane-sweep volume (PSV) representation as a structural correction cue to blend the synthesized results. This mechanism effectively eliminates artifacts, such as ghosting and broken limbs in occluded regions, while preserving the clarity of the background.

Keywords: Human novel view synthesis; Sparse-view reconstruction; Neural rendering

Supported by the National Natural Science Foundation of China (Grant No. 62031023).

Citation: Kaijin ZHAO, Chuyue ZHAO, Zhiheng JIANG, Guoqing ZHOU, Qing WANG. Learning novel view synthesis for human from image pairs. Virtual Reality & Intelligent Hardware, 2026, 8(5): 596–618.

1 Introduction

Novel view synthesis (NVS) aims to synthesize photorealistic images of a target view from a given set of reference images, standing as a core task in computer vision and graphics. As a crucial sub-domain of NVS, human-centric novel view synthesis plays a vital role in practical applications such as augmented/virtual reality (AR/VR), free-viewpoint video, and immersive telepresence. To meet the demands for low cost and

Corresponding author, qwang@nwpu.edu.cn

Peer review under the responsibility of Beijing Zhongke Journal Publishing Co. Ltd.
2096-5796/©Copyright 2026 Beijing Zhongke Journal Publishing Co. Ltd., Publishing services by Elsevier B.V. on behalf of KeAi Communication Co. Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by/4.0/).

easy deployment in real-world scenarios, these applications typically require systems to generate high-fidelity and geometrically accurate human renderings using only sparse input views (e.g., stereo image pairs). Throughout its evolution, NVS technology has made a quantum leap from classic light field-based methods [1, 2] to a new era represented by neural rendering [3 – 11]. Despite the remarkable progress of existing techniques, robust generalization under sparse-view conditions remains a formidable challenge, particularly in recovering fine-grained geometric details that are highly perceptually sensitive, such as facial features and limbs, and when handling complex human-object interactions. As shown in Figure 1, these errors become particularly evident in challenging regions such as facial details and limbs, where existing methods often produce distortions or incomplete structures.

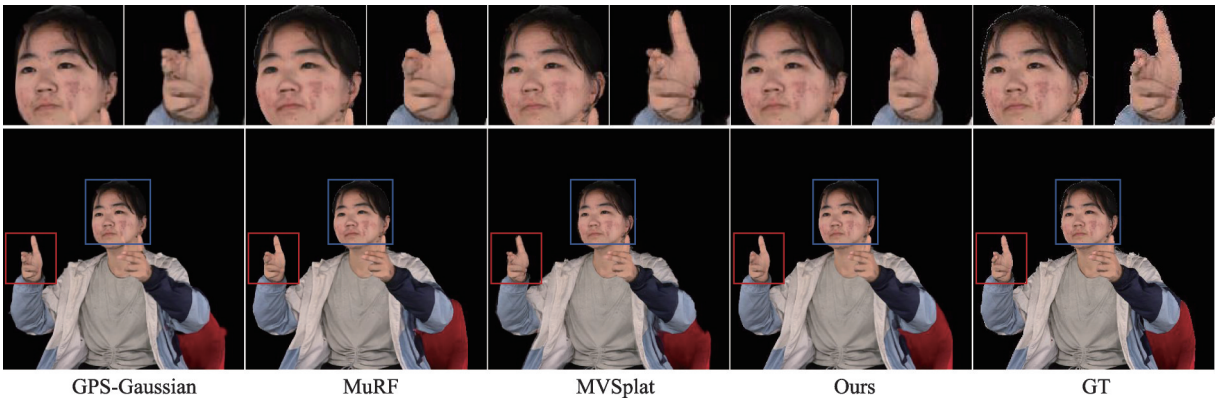


Figure 1 Our method outperforms state-of-the-art approaches by delivering superior visual quality, even in ill-posed regions prone to depth estimation errors. It also excels in reconstructing the geometric structures and texture details of human figures, as indicated by the red and blue boxes in the figure.

In recent years, methods based on neural radiance fields (NeRF) [8,12] and 3D Gaussian splatting (3DGS) [13,14] have significantly advanced the field of NVS. To achieve generalization under sparse-view settings, existing generalizable NeRF and 3DGS approaches typically rely on depth estimation or point cloud priors, back-projecting 2D image features into 3D space for alignment and rendering. For human-centric NVS tasks, another mainstream strategy is to incorporate human parametric models. These methods estimate a base mesh containing human pose and shape from the input images, which is subsequently used to initialize 3DGS or to align image features via densification.

However, these existing methods still face critical bottlenecks in practical applications. On the one hand, for methods relying on depth priors, current depth estimation techniques are highly prone to errors when handling textureless regions and complex human self-occlusions. Without effective error-correction mechanisms, these initial geometric flaws are directly propagated and amplified in the final rendered views, leading to floaters around body boundaries, broken limbs, and facial distortions, severely degrading visual fidelity. On the other hand, although incorporating human parametric models ensures the structural coherence of the human body, these methods struggle to represent loose clothing and complex hairstyles. More critically, they typically rely on full-body images as input, rendering them incapable of handling partial-body inputs or modeling human-object interactions. These limitations severely restrict the deployment of existing human-centric NVS methods in complex, real-world scenarios.

To address these issues, we propose a robust and generalizable method for human-centric novel view synthesis using only a pair of images as input. Building upon the enhancement of depth estimation accuracy via monocular depth features, our approach explicitly introduces a robust error correction mechanism. Specifically, we first integrate features from a monocular depth estimation network into an iterative depth

estimation framework to obtain the depth map of the target view. Subsequently, we design an attention-based fusion module to extract valid features and filter out erroneous information caused by depth inaccuracies. Moreover, we propose a dual-branch compositing mechanism to further enhance the stability of the final rendered results. By leveraging a coarse yet stable plane-sweep volume (PSV) representation for the foreground, this mechanism effectively eliminates artifacts such as ghosting and broken limbs in occluded regions while strictly preserving the clarity of the background. Extensive quantitative and qualitative evaluations on challenging datasets, including ThumanSit [15], Dual-GS [16], and real-world captures, validate the effectiveness and superiority of our approach.

In a word, our contributions are summarized as follows.

•We propose a novel error-correction framework for robust novel view synthesis from an image pair, which alleviates artifacts caused by depth ambiguities.

•We design an attention-based feature fusion module coupled with a dual-branch compositing mechanism, which significantly improves rendering quality in occluded regions and preserves high-fidelity facial details.

•Our method achieves superior performance compared to state-of-the-art approaches in experiments on multiple human datasets and practical scenarios.

2 Related work

Our work on robust human novel view synthesis is situated at the intersection of generalizable 3D scene representation and human-centric modeling. In this section, we first review the recent advancements and limitations of generalizable novel view synthesis methods. Subsequently, we further discuss the challenges faced by existing human-centric novel view synthesis approaches.

2.1 Generalizable 3D scene representations

Explicit representations. Explicit methods model scenes with concrete geometric primitives like meshes, point clouds, or volumes. Early works like MPIs [17–20] achieved stability by discretizing the scene into depth planes, but were fundamentally limited by sampling density, leading to rendering distortions. Diverging from discretization, another line of work leverages explicit depth maps to warp and fuse features from input views. FWD [21] pioneers this by forward-warping point cloud features and fusing them with a Transformer for real-time synthesis. Building on this, Li et al. [22] proposed a spatio-temporal feature warping (STFW) mechanism that further enforces temporal consistency using optical flow and learns adaptive fusion weights from camera poses. However, the performance of these methods remains inherently sensitive to the accuracy of the estimated or captured depth maps. More recently, 3DGS [3] has achieved remarkable quality by optimizing a set of Gaussian primitives. Despite achieving impressive visual quality, 3DGS relies on iteratively updating the attributes and quantity of Gaussian primitives via gradient descent. This requirement for per-scene optimization severely restricts its generalization capabilities to unseen environments.

To overcome this limitation, subsequent research often acquires explicit geometric priors via depth estimation, thereby circumventing the iterative optimization of Gaussian positions. For instance, pixelSplat [4] predicts a dense probability distribution over the 3D space and samples from it to determine the centers of Gaussian primitives. MVSplat [5], building upon multi-view stereo (MVS) [23], constructs cost volumes to estimate depth maps, which are then used to unproject 3D Gaussians. DepthSplat [24] and TranSplat [25] further incorporate monocular depth priors (e.g., ref. [26] or ref. [27]) and monocular geometric features to enhance depth accuracy in textureless regions. However, these approaches generally lack a robust error-handling mechanism for depth mispredictions. Consequently, any inaccuracies or failures in depth estimation

directly translate into severe artifacts or structural collapse in the final rendered views.

Concurrently, several 3DGS-based methods have introduced Gaussian pruning mechanisms. However, these techniques are primarily designed for dense input views to prevent excessive Gaussian primitives from causing severe memory overhead and ghosting artifacts. For instance, GGN [28] introduces a Gaussian graph network to generate efficient and generalizable Gaussian representations, employing a Gaussian pooling layer to aggregate multiple primitives. Nevertheless, it still fundamentally relies on depth estimation for unprojection, making it difficult to prevent the accumulation of depth errors. Alternatively, FreeSplat [29] and FreeSplat+ [30] determine the validity of Gaussian primitives by evaluating their consistency across different input views within a unified coordinate system. While this strategy effectively mitigates depth errors under multi-view settings, it falters when given only an image pair. In such sparse-view scenarios, the network cannot reliably distinguish between valid geometry and noisy interference among inconsistent Gaussians, leaving the elimination of depth estimation errors an unresolved challenge.

On the other hand, some methods estimate the attributes of Gaussian primitives through feature matching [31–33]. This purely data-driven strategy inherently ties the quality of the final synthesis to the capacity of the underlying network. Although these approaches have demonstrated promising results, they typically demand substantial computational resources and extensive training times, thereby limiting their practical applicability.

In summary, existing generalizable 3DGS methods heavily rely either on explicit geometric priors or on massive network capacity. Consequently, the fidelity of the final rendering is directly influenced by the accuracy of the initial depth estimates or feature matches. When these intermediate predictions fail, they not only degrade the visual quality but also produce noticeable artifacts, such as misaligned geometry or floating structures, ultimately leading to incomplete or distorted scene representations.

Implicit representations. In contrast, implicit methods like NeRF [6] represent scenes as continuous functions parameterized by neural networks. Although they achieve impressive photo-realistic results, the volumetric rendering process requires dense sampling along each ray, making optimization computationally intensive and time-consuming. To address these bottlenecks, subsequent works have enhanced rendering quality [34–38] and accelerated training by leveraging explicit data structures or lightweight architectures [39–41]. Beyond quality and efficiency, reducing the reliance on dense multi-view captures has also been a major focus, leading to methods optimized for sparse inputs [42–44]. Furthermore, to enhance generalizability across diverse scenes, rather than directly relying on spatial coordinates, recent methods leverage geometric priors to sample features from input images. Some approaches employ pre-estimated depth to guide ray sampling [12,45–47], whereas others exploit MVS features to inform the rendering process. For instance, MVSNeRF [48] and IBRNet [7] utilize feature matching to predict color and density, while NeuRay [49] further incorporates these features for visibility estimation. Similarly, MuRF [8] introduces a multi-stage paradigm where coarse predictions guide a finer sampling distribution. Despite these advancements, their heavy reliance on intermediate geometric proxies—such as estimated depth maps or learned feature fields—makes them highly vulnerable to estimation errors. When provided with flawed geometric cues, existing methods lack effective mechanisms to correct these errors, inevitably leading to blurry regions, noisy surfaces, and ghosting artifacts that severely degrade synthesis quality.

2.2 Human-centric novel view synthesis

Existing methods for modeling human avatars often rely on strong geometric priors as constraints. A common strategy involves leveraging parametric human body models (e.g., SMPL [50] or SMPL-X [51]), with GHG [52] constructing multi-level scaffolds by dilating the base template, and Animatable Gaussians [53] extending SMPL-X to directly incorporate clothing. Furthermore, unified volumetric avatar [54]

enhances its comprehension of 3D space by constructing a signed height indicator relative to the SMPL template. GM-NeRF [55] employs SMPL as a coarse geometric prior and refines details using a geometry-guided attention mechanism.

Building upon SMPL model, Transhuman [56] constructs a grouped representation to capture more precise geometric details. An alternative to explicit body models is to infer geometry from image features. Such methods guide the final rendering process by jointly optimizing a surface field with the radiance field [57] or by incorporating an auxiliary visibility field [58].

However, whether relying on parametric models or feature matching, these generalizable methods share a critical limitation: their underlying geometric representations are often overly simplified, and they frequently impose strict assumptions on the scene content. Consequently, they severely struggle to handle complex human-object interactions or incomplete human inputs (e.g., partial-body observations). Meanwhile, another line of work focuses on dynamic scene reconstruction [59–63]. These approaches effectively capture motion by leveraging temporal information, learning deformation models, or in some cases leveraging reliable flow priors for regularization. However, they typically require extensive per-subject optimization, which restricts their ability to generalize to unseen individuals.

To bypass the input restrictions imposed by parametric human models, some recent methods directly leverage depth priors to guide the novel view synthesis of human subjects [64,65]. However, these approaches remain highly susceptible to the inevitable errors introduced during depth estimation. As a notable exploration to address this issue, GPS-Gaussian+ [66] introduces a deformation field applied to the Gaussian primitives unprojected from the estimated depth. Yet, because this deformation mechanism operates without any additional auxiliary cues, it often exacerbates artifacts and further degrades the synthesis quality in complex scenarios.

3 Method

Our method aims to achieve high-quality, robust, and generalizable novel view rendering of the human body using a pair of input images. The overview of our method is illustrated in Figure 2, which can be divided into three parts. First, we perform depth estimation on the input image pair and transform the resulting depth maps to the target view, enabling the matching and integration of input information. Since depth estimation errors can severely degrade rendering quality, we incorporate monocular depth features into the stereo depth estimation framework and fuse them to obtain more stable results in ill-posed regions such as occlusions. Subsequently, we construct a fusion network that integrates self-attention and soft attention mechanisms to effectively discriminate and aggregate the aligned input view information, yielding an initial fused result. Finally, we introduce a dual-branch compositing module. By leveraging the stable geometric cues provided by the PSV as the foreground, this module effectively eliminates potential artifacts in the fused background while strictly preserving its clarity, ultimately rendering a visually pleasing and high-fidelity result.

3.1 Depth estimation

We first perform depth estimation on the input image pair. Given the input images $\boldsymbol{I}_L$, $\boldsymbol{I}_R$, their corresponding intrinsic parameters $\boldsymbol{K}_L$, $\boldsymbol{K}_R$, and extrinsic parameters $\boldsymbol{P}_L$, $\boldsymbol{P}_R$, as well as the target view camera parameters $\boldsymbol{K}_t$, $\boldsymbol{P}_t$, we initially apply stereo rectification to the input images and camera parameters, yielding rectified images $\boldsymbol{I}'_L$, $\boldsymbol{I}'_R$ and corresponding parameters $\boldsymbol{K}'_L$, $\boldsymbol{K}'_R$, $\boldsymbol{P}'_L$, $\boldsymbol{P}'_R$. Subsequently, following the design of Lipson et al. [67], we construct a cascaded gated recurrent unit (GRU) network. Specifically, we first employ two non-shared encoders named a feature encoder and a context encoder to extract features from the rectified images $\boldsymbol{I}'_L$, $\boldsymbol{I}'_R$. The feature encoder yields features $\boldsymbol{f}$ and $\boldsymbol{g}$, while the context encoder produces $\boldsymbol{C}$, where $\boldsymbol{f}$

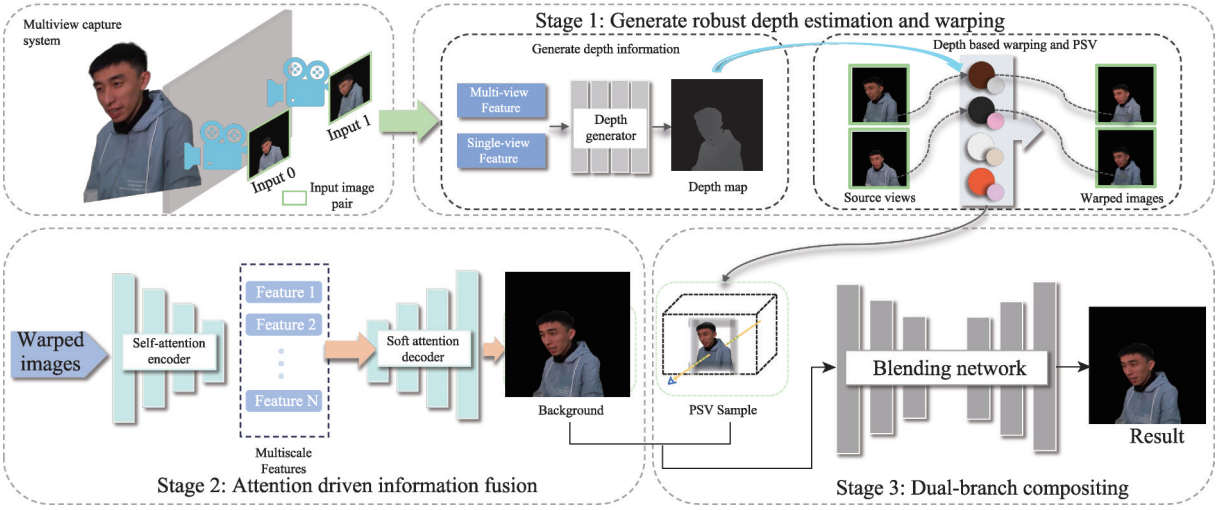


Figure 2 Method overview. Our method takes a pair of input images as input. First, we design a depth estimation network that integrates monocular and binocular features to generate robust depth maps for the input views. These depth maps are then transformed and fused into the target view's camera coordinate system, yielding the depth map under the target view. Subsequently, using the depth map, we warp the images from the two input views to the target view and construct a fusion network that combines self-attention and soft attention mechanisms to integrate the input information, producing an initial fused result. However, simple fusion struggles to address artifacts and distortions caused by depth estimation errors. To this end, we introduce a plane-sweep volume (PSV) refinement module by treating the initial fusion result as the background and the PSV sampling result as the foreground. A network estimates the valid regions of the background and fuses them with the foreground information, ultimately rendering a high-quality image under the target view.

and $\boldsymbol{g}$ are used to build the correlation volume, and $\boldsymbol{C}$ initializes the hidden states and is injected into each GRU unit. The structure of the encoders is identical except that the feature encoder utilizes instance normalization [68] and the context encoder uses batch normalization [69].

Next, we construct the correlation volume using $\boldsymbol{f}$ and $\boldsymbol{g}$ ($\boldsymbol{f},\ \boldsymbol{g} \in \mathbb{R}^{H\times W\times D}$). Due to stereo rectification, we compute correlations only along the x-coordinate. The correlation volume $\boldsymbol{C} \in \mathbb{R}^{H\times W\times W}$ is built as follows:

$$\boldsymbol{C}_{ijk} = \sum_{h} \boldsymbol{f}_{ijh} \cdot \boldsymbol{g}_{ikh},\ \boldsymbol{C} \in \mathbb{R}^{H\times W\times W}. \tag{1}$$

During each GRU iteration, we sample from the correlation volume using the current disparity estimate to obtain binocular stereo information. Furthermore, we incorporate monocular features to augment the contextual features $\boldsymbol{C}$. Given the exceptional performance and generalization capabilities of the depth anything model [26] in various monocular depth estimation tasks, we leverage its pre-trained model as another encoder. Specifically, we freeze the parameters of depth anything and extract its deep features. These features are then added as a residual to the contextual features $\boldsymbol{C}$ that are iteratively processed by each GRU cell, which yields the monocular feature. Subsequently, the monocular feature, along with the stereo information, is input into the GRU network for iterative refinement, producing the final disparity $\boldsymbol{D}_i$. Since disparity and depth are interconvertible using camera parameters, for notational convenience, we hereafter refer to both as $\boldsymbol{D}_i$.

Upon obtaining the depth map for the input view, we reproject it to the target view's camera coordinate system using the rectified intrinsic and extrinsic camera parameters ($\boldsymbol{K}_L'$, $\boldsymbol{K}_R'$, $\boldsymbol{P}_L'$, $\boldsymbol{P}_R'$). This reprojection yields the corresponding depth maps $\boldsymbol{D}_L'$ and $\boldsymbol{D}_R'$. To synthesize the final depth map for the target view $\boldsymbol{D}_t$, we take the element-wise minimum of the two reprojected depth maps, effectively selecting the foreground depth. This operation is defined as

$$\boldsymbol{D}_t = \min\left\{\boldsymbol{D}_L',\ \boldsymbol{D}_R'\right\}. \tag{2}$$

Finally, we warp the rectified input images, $\boldsymbol{I}_L'$ and $\boldsymbol{I}_R'$, to the target view based on the synthesized depth map

$\boldsymbol{D}_t$, thereby generating the rendered images $\boldsymbol{W}_L$ and $\boldsymbol{W}_R$.

3.2 Attention-driven information fusion

After obtaining the warped images $\{\boldsymbol{W}_L, \boldsymbol{W}_R\}$, depth inaccuracies and occlusions often cause the warped features from $\boldsymbol{W}_L$ and $\boldsymbol{W}_R$ to be spatially misaligned in certain regions. Although a plain convolutional network can detect such inconsistencies, its limited receptive field prevents it from reliably determining which view contains valid information and which carries artifacts. To overcome this, we equip the fusion network with a self-attention encoder that explicitly compares each local token with the global context, enabling it to discriminate error-prone regions from trustworthy ones. In the decoder, soft attention further adaptively selects features across scales, preserving fine textures while leveraging high-level semantics to produce robust fusion weights.

Specifically, our approach first combines the strengths of different architectures to achieve optimal results. The Swin Transformer [70] excels at extracting global features but performs less effectively in handling local information. In contrast, CNN architecture network excels at capturing local details, but often struggles to capture global context. Therefore, we combine the Swin Transformer with a CNN network featuring residual connections to extract image features. Then, we adopt the feature modulation block in Dehamer [71] to modulate and integrate these complementary features, which is formulated as

$$\begin{aligned} \boldsymbol{F}_c^s &= C(\boldsymbol{W}_L | \boldsymbol{W}_R), \\ \boldsymbol{F}_t^s &= T(\boldsymbol{W}_L | \boldsymbol{W}_R), \\ \boldsymbol{F}_m^s &= G_\gamma^s(\boldsymbol{F}_t^s) \otimes IN(\boldsymbol{F}_c^s) \oplus G_\beta^s(\boldsymbol{F}_t^s), \end{aligned} \tag{3}$$

where $C(\cdot)$ denotes the CNN network, $T(\cdot)$ denotes the Swin Transformer network. | denotes a concatenation of the warped image pair $\{\boldsymbol{W}_L, \boldsymbol{W}_R\}$. $\boldsymbol{F}_c^s$ and $\boldsymbol{F}_t^s$ denote the features extracted by the CNN network and the Swin Transformer, respectively. $\boldsymbol{F}_m^s$ denotes the modulated features. $s \in \{1, 2, 3, 4\}$ denotes the stage level, and we set 4 stages in all our experiments. Specifically, the spatial resolution is downsampled by half at each stage. *IN* represents the instance normalization operation. $G_\gamma^s(\cdot)$ and $G_\beta^s(\cdot)$ denote a single convolution layer, and the parameters are not shared. $\otimes$ and $\oplus$ denote the element-wise multiplication and addition, respectively.

In order to capture the information embedded within the features and enhance the network's capacity, inspired by MIRNet [72], we introduce the dual attention unit (DAU) and selective kernel feature fusion (SKFF) modules into our network. By integrating these modules into our decoder, we enhance the extraction of relevant information while minimizing irrelevant details, thereby improving the quality of the decoded results, which is denoted as $\boldsymbol{I}_f$. The detailed network architecture of this module is shown in Figure 3.

3.3 Dual-branch compositing

Although the fused image $\boldsymbol{I}_f$ offers a reasonably complete reconstruction of the target view, it remains susceptible to artifacts caused by accumulated errors from the depth estimation stage. To mitigate these distortions, we introduce a dual-branch compositing strategy based on alpha compositing. In this design, the detail-rich but potentially artifact-prone $\boldsymbol{I}_f$ serves as the logical background (the bottommost layer) in the alpha compositing process, while the geometrically stable PSV acts as an overlaid rendering layer above it. It is worth noting that in human-centric novel view synthesis tasks, it is common to separate the foreground human from the background scene, or to decouple dynamic and static regions. Although we also use the term "background" here, it does not refer to such scene-level foreground-background separation, but merely serves as a convenient description of the alpha compositing process. PSV operates by warping the source image to the target view via a series of homography transformations, each corresponding to a discrete depth

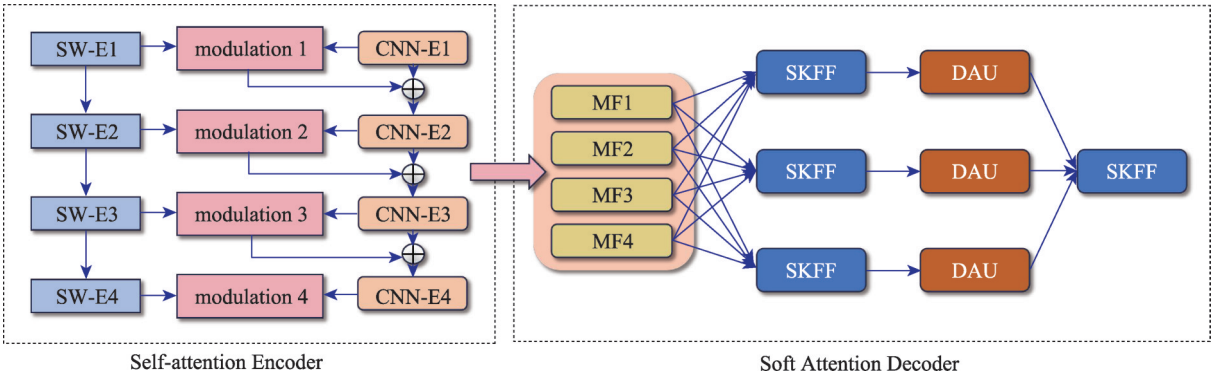


Figure 3 We employ a self-attention encoder that utilizes a feature modulation module to modulate multi-scale features obtained from both Swin Transformer and CNN, resulting in more information-rich features. The modulated results serve as residuals feeding into the next CNN layer, and the encoder outputs are sent to the decoder. In the soft attention decoder, we concatenate the multi-scale features and pass them through a selective kernel feature fusion (SKFF) module, followed by a dual attention unit (DAU) module. This process ultimately yields the fused result at the target view.

plane within a predefined range. This method ensures geometric consistency, thereby preventing artifacts like limb discontinuities or texture distortions that often plague direct depth-based projection methods. However, the PSV relies on discrete depth sampling and struggles to recover precise depth values. Since the true scene depth is continuous, the warped features from the two input views can rarely achieve exact pixel-wise alignment on any one sampled plane. When this misalignment occurs, the per-pixel fusion network becomes ambiguous and lacks a way to correct the offset; it tends to degenerate into a simple average of the misaligned features, discarding high-frequency details and resulting in the characteristic blur. The problem is further aggravated in occluded regions, where the network cannot distinguish a genuine occlusion from a mismatch caused by insufficient sampling density, again leading to averaging artifacts. To this end, we design a PSV-based blending module to fuse the sharp textural details of $\boldsymbol{I}_f$ with the structural completeness from the PSV, yielding a high-quality final result. By supplying $\boldsymbol{I}_f$ as an additional prior, we give the blending network a clear reference for what a sharp, artifact-free rendering should look like. With this guidance, the network no longer needs to resolve the depth ambiguity from the PSV alone; it only has to decide where $\boldsymbol{I}_f$ is already correct and where structural correction from PSV is needed.

First, we construct the PSV. The homography for each plane is computed as follows:

$$\begin{bmatrix} u_s \\ v_s \\ 1 \end{bmatrix} \sim \boldsymbol{K}_s \left(\boldsymbol{R}^{\top} + \frac{\boldsymbol{R}^{\top} \boldsymbol{t} \boldsymbol{n} \boldsymbol{R}^{\top}}{a - \boldsymbol{n} \boldsymbol{R}^{\top} \boldsymbol{t}} \right) \boldsymbol{K}_t^{-1} \begin{bmatrix} u_t \\ v_t \\ 1 \end{bmatrix}, \tag{4}$$

where $[u_s, v_s, 1]^{\top}$ and $[u_t, v_t, 1]^{\top}$ represent the homogeneous coordinates of the source and target views, respectively. $\boldsymbol{n}$ denotes the normal of the plane, and a is the plane's bias term. When $\boldsymbol{x}$ is specified as $[u_s, v_s, 1]^{\top}$, we set $\boldsymbol{n}^{\top}\boldsymbol{x} + a = 0$. Specifically, $\boldsymbol{n} = [0, 0, 1]^{\top}$ and $a = -d_s$, where d_s represents the depth of the plane. $\boldsymbol{K}_s' \in \{\boldsymbol{K}_L', \boldsymbol{K}_R'\}$ and $\boldsymbol{K}_t$ denote the intrinsic camera parameters of the source and target cameras, respectively. $\boldsymbol{R}$ and $\boldsymbol{t}$ represent the relative pose between the cameras.

Then, we concatenate the PSV and $\boldsymbol{I}_f$ and feed them into a U-Net architecture network. Thereby, we can obtain a set of blending weights $\boldsymbol{w}$ and opacity $\boldsymbol{\sigma}$, which can be formulated as

$$(\boldsymbol{w}, \boldsymbol{\sigma}) = U(\boldsymbol{I}_{psv}, \boldsymbol{I}_f), \tag{5}$$

where $U(\cdot)$ represents a U-Net network, and $\boldsymbol{I}_{psv}$ represents the PSV results.

Finally, we composite $\boldsymbol{I}_f$ and $\boldsymbol{I}_{psv}$ to produce the final rendered image. In our alpha compositing pipeline, the bottommost layer is directly taken as $\boldsymbol{I}_f$. The PSV is processed through N discrete depth planes ($N = 32$), each with a blending weight map $\boldsymbol{w}_i$ and an opacity map $\boldsymbol{\sigma}_i$. The per-plane blended color is obtained as $\boldsymbol{L}_i =$

$\boldsymbol{w}_i \otimes \boldsymbol{I}_{psv,i}$. Using the standard over operator, the final image $\boldsymbol{I}_t$ is computed by recursively blending these layers over $\boldsymbol{I}_f$:

$$\boldsymbol{I}^{(0)} = \boldsymbol{I}_f,\ \boldsymbol{I}^{(i)} = \boldsymbol{L}_i \odot \boldsymbol{\sigma}_i + (1 - \boldsymbol{\sigma}_i) \odot \boldsymbol{I}^{(i-1)},\ i = 1, \cdots, N, \boldsymbol{I}_t = \boldsymbol{I}^{(N)}, \tag{6}$$

where $\odot$ denotes element-wise multiplication. Equivalently, the closed-form expression is

$$\boldsymbol{I}_t = \sum_{i=1}^{N} \left(\boldsymbol{L}_i \odot \boldsymbol{\sigma}_i \odot \prod_{j=i+1}^{N} (1 - \boldsymbol{\sigma}_j) \right) + \boldsymbol{I}_f \odot \prod_{i=1}^{N} (1 - \boldsymbol{\sigma}_i). \tag{7}$$

In effect, the predicted opacity $\boldsymbol{\sigma}$ learns to preserve the sharp background $\boldsymbol{I}_f$ in well-reconstructed areas (where $\boldsymbol{\sigma}$ is low) and to fall back on the geometrically stable PSV foreground only in regions affected by depth-induced artifacts or occlusions (where $\boldsymbol{\sigma}$ is high). This mechanism cleanly decouples texture fidelity from structural correction, allowing the network to retain the superior clarity of the depth-based fusion while patching inevitable errors.

3.4 Optimization

Our method jointly supervises the estimated depth map $\boldsymbol{D}_t$ and the outputs of the two rendering stages, $\boldsymbol{I}_f$ and $\boldsymbol{I}_t$.

Depth loss. For depth supervision, we compute the L1 loss between the ground-truth depth $\boldsymbol{D}_{gt}$ and the full sequence of predictions $\{\boldsymbol{D}_1, \boldsymbol{D}_2, \cdots \boldsymbol{D}_T\}$. The disparity loss is defined as

$$\mathcal{L}_{disp} = \sum_{i=1}^{T} \gamma^{T-t} \left\| \boldsymbol{D}_{gt} - \boldsymbol{D}_i \right\|_1, \tag{8}$$

where T denotes the length of the full sequence, and γ is set to 0.9.

Image losses. For the rendered images, we apply an L1 loss, denoted as $\mathcal{L}_f$, between the initial fused image $\boldsymbol{I}_f$ and the ground-truth image $\boldsymbol{I}_{gt}$. The final rendered output $\boldsymbol{I}_t$ is supervised by both an L1 loss $\mathcal{L}_t$ and a structural similarity index measure (SSIM) loss $\mathcal{L}_{ssim}$ [73].

Total loss. The final loss is a weighted sum of these components:

$$\mathcal{L} = w_1 \mathcal{L}_{disp} + w_2 \mathcal{L}_f + w_3 \mathcal{L}_t + w_4 \mathcal{L}_{ssim}. \tag{9}$$

In our experiments, we set the weights to $w_1 = 1$, $w_2 = 0.1$, $w_3 = 0.8$, and $w_4 = 0.1$.

4 Experiment

4.1 Implementation details

Our model is implemented in PyTorch and trained on a single NVIDIA L40 GPU. Specifically, we adopt a two-stage training strategy. Firstly, the depth estimation module is pre-trained for 50000 iterations. Following this, the entire model is jointly trained in an end-to-end manner for an additional 100000 iterations. We use a batch size of 1 for all training stages. The model is optimized using the AdamW optimizer [74] with an initial learning rate set to 1×10^{-5}.

4.2 Dataset and metrics

To comprehensively evaluate the performance and generalization capability of our method, we conduct experiments on three diverse datasets: the THumanSit and Dual-GS datasets, and a newly captured real-world dataset. To ensure a fair comparison, all test samples are unseen by any of the compared methods during training. Critically, to assess zero-shot generalization, no methods (including our own) were trained or fine-tuned on the Dual-GS and our real-world datasets.

•**THumanSit:** A synthetic dataset featuring various human-chair interactions and sitting poses. To align with practical application scenarios, we render only the upper body and evaluate on 1600 test samples.

•**Dual-GS:** A dataset containing scenes of complex interactions between humans and musical instruments. We select four representative scenes and use 200 frames from each for testing (800 frames in total).

•**Real-world dataset:** A new dataset captured to test performance in practical settings. It consists of 1050 frames from 7 subjects, recorded using four parallel Azure Kinect cameras. To further evaluate the quality of the rendered results, we employ the off-the-shelf matting model [75] to mask out human regions in all captured images.

Following standard practice, we evaluate the quality of rendered images using peak signal-to-noise ratio (PSNR), SSIM [73], and learned perceptual image patch similarity (LPIPS) [76].

4.3 Baselines

We compare our method with several generalizable methods that synthesize novel views from image pairs, including (1) GPS-Gaussian [64] and GPS-Gaussian+ [66], 3DGS-based methods tailored for human subjects, (2) MuRF [8], a NeRF-based method featuring promising visual quality and robust generalization capabilities, (3) MVSplat [5] and DepthSplat [24], 3DGS-based methods for general-purpose novel view synthesis. For fair comparison, all baselines were retrained on our training data with their original hyper-parameters preserved; only the number of iterations and batch size were unified to match our setup, allowing a comparison under equivalent training budgets.

4.4 Comparison with state-of-the-art methods

Quantitatively, our method achieves the best results on most evaluation metrics. It is worth noting that none of the evaluated methods were trained on the Dual-GS and real-world datasets; nevertheless, our approach still delivers the highest performance, highlighting its robust generalization capability. Furthermore, on the THumanSit dataset, our method significantly outperforms all competing baselines. These performance gains can be attributed to the dual-branch compositing module's capacity to rectify geometric inconsistencies and reduce floating and ghosting artifacts resulting from depth estimation errors. Table 1 reports the specific evaluation metrics of the compared methods on each dataset.

Table 1 Quantitative comparison of novel view synthesis. We report the average PSNR, SSIM, and LPIPS results on THumanSit, Dual-GS, and real-world datasets. The best performance is marked in bold

Method	THumanSit			Dual-GS			Real-world		
	PSNR (dB) ↑	SSIM ↑	LPIPS ↓	PSNR (dB) ↑	SSIM ↑	LPIPS ↓	PSNR (dB) ↑	SSIM ↑	LPIPS ↓
GPS-Gaussian [64]	26.56	0.947	0.0853	24.53	0.887	0.1280	24.52	0.890	0.1287
GPS-Gaussian+ [66]	24.18	0.914	0.1104	24.10	0.880	0.1367	25.18	0.891	0.1589
MVSplat [5]	25.83	0.898	0.0619	24.62	0.874	**0.0696**	21.21	0.810	0.1889
DepthSplat [24]	23.77	0.852	0.0932	24.17	0.866	0.1204	19.70	0.805	0.2410
MuRF [8]	27.92	0.940	0.0578	25.04	0.890	0.1042	29.85	0.937	0.0850
Ours	**28.20**	**0.963**	**0.0568**	**25.56**	**0.915**	0.1052	**30.16**	**0.940**	**0.0767**

Qualitatively, our method demonstrates superior performance across all three datasets, as illustrated in Figures 4–6. It effectively mitigates ghosting artifacts and floaters in the synthesized results while preventing the facial distortions common in competing approaches. This superiority is attributable to our explicit error correction mechanism against depth estimation inaccuracies. By integrating attention-driven fusion with PSV-guided dual-branch compositing, our approach systematically corrects local artifacts and resolves ambiguities in ill-posed regions such as occlusions, thereby preventing large-scale ghosting and structural discontinuities. In contrast, GPS-Gaussian directly back-projects Gaussian primitives using the estimated depth, thereby inevitably propagating depth inaccuracies into the final rendering. Even after retraining, it

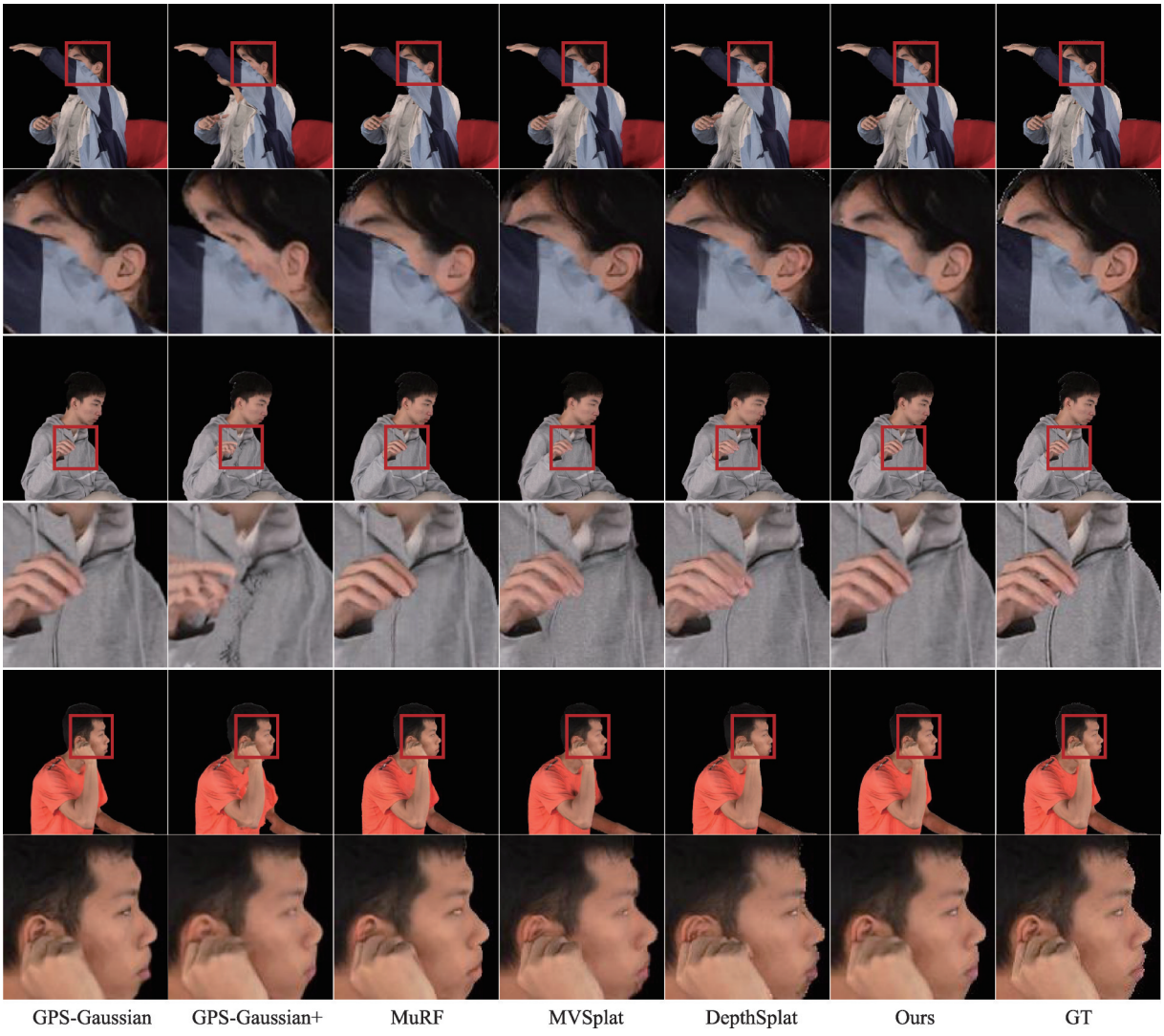


Figure 4 Qualitative comparison on THumanSit dataset. Compared to prior works, our method yields superior visual results, effectively eliminating common rendering artifacts such as floaters, limb distortions, and ghosting, as highlighted in the red rectangles.

still exhibits visible floating artifacts and facial distortions on out-of-domain datasets due to its high sensitivity to depth scale variations, a limitation that persists because the method directly feeds raw depth into the network to enforce geometric accuracy.

Building upon its predecessor, GPS-Gaussian+ refines depth estimation and introduces a deformation field to mitigate the inevitable errors in the depth estimation process. However, this deformation is predicted solely from the fused image-depth features without incorporating any auxiliary information. Consequently, the deformation mechanism frequently fails in complex scenes, introducing additional artifacts and struggling to recover accurate geometry. Furthermore, both GPS-Gaussian and GPS-Gaussian+ directly feed the estimated depth into the network to regress Gaussian primitive attributes, making both methods sensitive to depth distribution. A shift in depth distribution can substantially degrade the rendered results. For example, in Figure 6, the depth distribution of the real-world dataset deviates considerably from that of the training set. This causes GPS-Gaussian to regress Gaussian primitives with underestimated scales, allowing the background to bleed through the foreground. In contrast, GPS-Gaussian+ tends to produce Gaussian primitives with overestimated scales, ultimately resulting in blurry and distorted outputs.

In contrast, MuRF achieves relatively promising visual quality by performing dense sampling within a



Figure 5 Qualitative comparison on Dual-GS dataset. Under a zero-shot setting where no methods were trained on this dataset, our method demonstrates superior generalization capabilities. As highlighted by the red rectangles and their magnified views, our method not only reconstructs the finest facial geometry but also remains free of floaters and ghosting artifacts in complex human-instrument interaction zones and on intricate clothing, which are prominent in the results of GPS-Gaussian and MuRF, respectively.

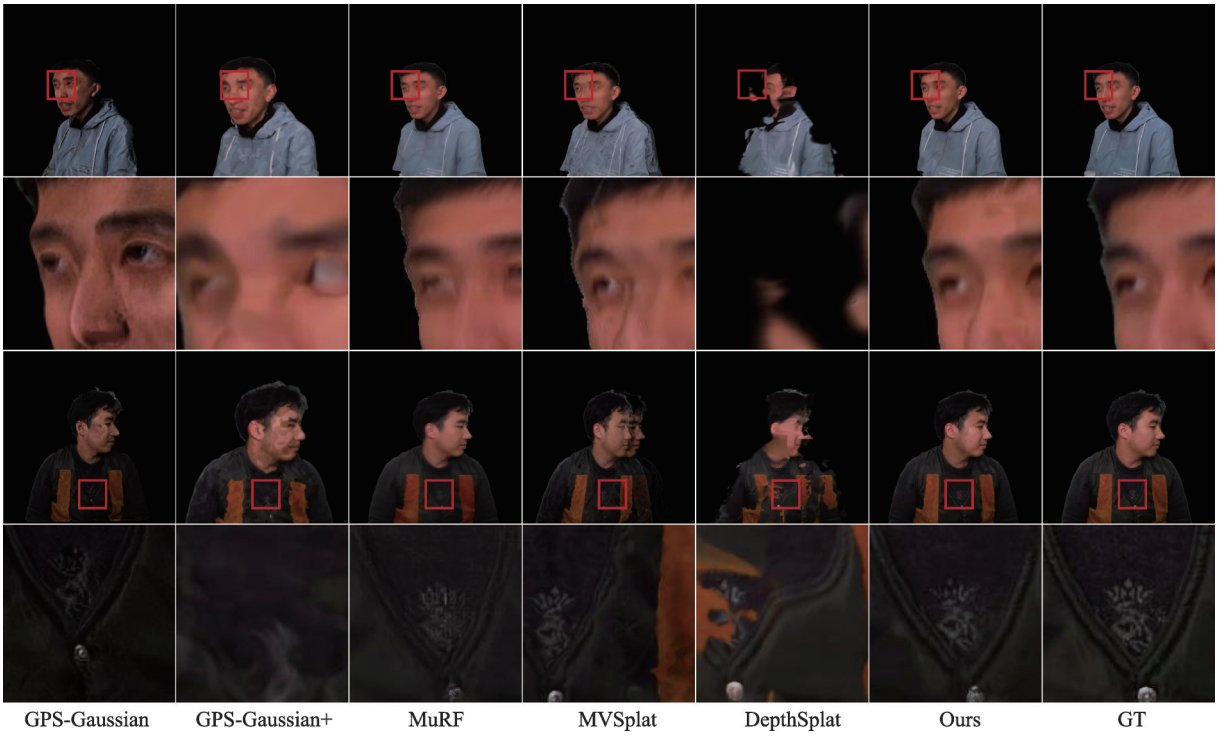


Figure 6 Qualitative comparison on real-world dataset. In challenging real-world scenarios, our method maintains superior visual fidelity. It successfully avoids the severe facial distortions and feature misalignments common in other methods, while maximally preserving clothing textures. Key areas of comparison are highlighted with red rectangles.

predefined depth range. However, it employs a two-stage cascaded sampling strategy, where the sampling density in the second stage relies entirely on the predictions from the first stage. Consequently, any

inaccurate predictions in the initial stage are directly propagated, leading to severe artifacts in the final rendering. For instance, on the unseen real-world and Dual-GS datasets, the reduced sampling accuracy in the first stage directly compromises the second-stage rendering, resulting in more pronounced artifacts and distortions on these two datasets. Furthermore, when fusing features from input views, MuRF lacks the capability to discriminate between informative signals and interfering noise. As a result, the network tends to average uncertain regions, which causes the blue color of the sleeve to bleed into the occluded face in the first example of Figure 4, producing a soft blue gradient instead of a clear boundary between the sleeve and the face. Conversely, our approach effectively circumvents this issue by incorporating an attention mechanism during the fusion process to adaptively aggregate valid information.

Both MVSplat and DepthSplat first estimate depth and then directly back-project Gaussian primitives using the estimated depth. This similarly leaves errors in depth estimation unhandled, which directly leads to rendering artifacts. Even on the test split of the training dataset, artifacts caused by depth estimation errors in occluded or complex regions persist, such as the ghosting and blurring on details like ears and hands shown in Figure 4. Moreover, because DepthSplat incorporates self-attention in its depth estimation pipeline, a change in input image resolution disrupts the information represented by each patch, severely impairing the model's generalization capability. This explains DepthSplat's poor performance on the real-world dataset, which uses a resolution of 1024×1024, in contrast to the 512×512 resolution of the other datasets and the training data. This design choice also limits DepthSplat's ability to seamlessly switch between images of different resolutions, thereby restricting its practical application potential.

4.5 Qualitative analysis of specific artifact types

To further understand how different methods behave under challenging conditions, we categorize common rendering artifacts into three representative types—floaters, broken limbs, and facial distortions—and present classified visual comparisons in Figure 7. Each example includes a zoomed-in region to better highlight local artifacts.

Floaters. Floater artifacts mainly appear around foreground boundaries, as shown in the first two rows of Figure 7. These artifacts are closely related to inaccurate geometry near depth discontinuities and insufficiently Gaussian primitives around object silhouettes. GPS-Gaussian and GPS-Gaussian+ still exhibit visible boundary noise and discontinuous foreground structures. In contrast, our method produces cleaner object boundaries and more continuous foreground regions, reducing the floating artifacts observed in the baselines.

Broken limbs. The third and fourth rows in Figure 7 show examples of the broken limb artifact. In this category, the artifact does not always appear as a complete disconnection of limbs; instead, it also includes severe limb thinning, local shrinkage, and structurally unstable body parts that visually approach disconnection. Such artifacts are particularly noticeable under challenging poses and self-occlusions. The baseline methods tend to produce overly thin limb regions and unstable boundaries, mainly due to unreliable depth estimates around occluded or ambiguous body regions, which affects the placement and scaling of Gaussian primitives. In contrast, our method better preserves the thickness and continuity of limbs by combining attention-driven fusion with PSV-guided blending. Even when local depth estimation is unreliable, the PSV provides a geometrically stable reference, helping maintain more complete and natural limb structures.

Facial distortions. The last two rows in Figure 7 demonstrate comparisons of facial reconstruction quality. On the real-world dataset, where none of the methods were trained, all competing approaches exhibit severe facial distortion. Even on side-profile views, baseline methods suffer from noticeable distortion of facial features. In contrast, our method reconstructs high-fidelity faces with accurate geometric

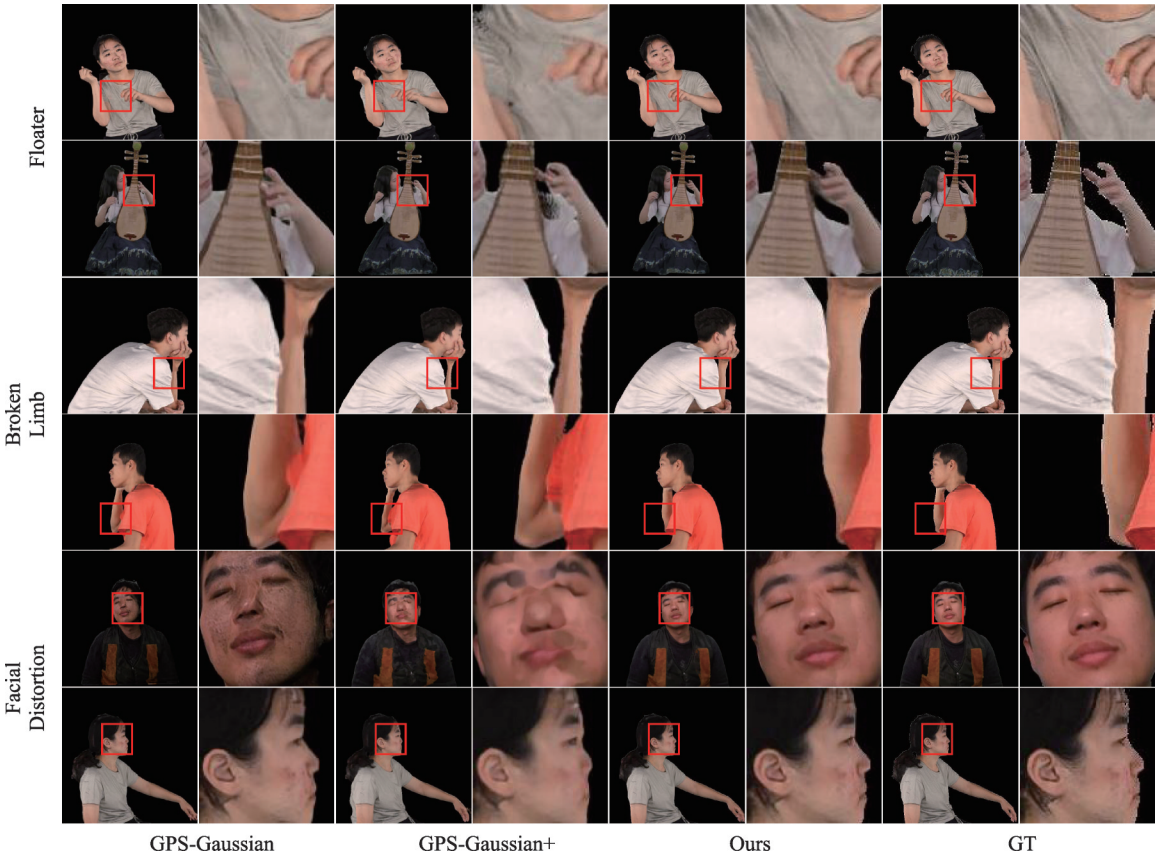


Figure 7 Visual comparisons categorized by three representative artifact types: floaters, broken limbs, and facial distortions. For each case, we show the rendered result together with a zoomed-in region. Compared with GPS-Gaussian and GPS-Gaussian+, our method more effectively suppresses these artifacts and produces results closer to the ground truth.

structure and sharp texture details. This robustness stems from our explicit error-correction mechanism: the attention module filters out erroneous features caused by depth inaccuracies, while the blending module preserves fine-grained details from the depth-based fusion, ensuring that facial regions, which are highly sensitive to perceptual quality, remain artifact-free.

In summary, this categorized analysis demonstrates that our method consistently suppresses all three prevalent artifact types across diverse datasets and challenging scenarios, further validating the effectiveness of our error-correction framework.

4.6 Computational efficiency

Although our method significantly outperforms existing baselines in overall visual quality, it still presents certain limitations regarding real-time inference speed. To objectively evaluate computational efficiency, we benchmarked our approach against other methods at a resolution of 1024×1024. All experiments were conducted on a single NVIDIA A100 80 GB PCIe GPU using identical datasets. We utilized the official GitHub implementations for all evaluated methods without any acceleration.

As shown in Table 2, while our approach is considerably faster than MuRF, MVSplat and DepthSplat, delivering superior rendering results without severely compromising efficiency, it falls short of the speed achieved by GPS-Gaussian and GPS-Gaussian+, thus not yet achieving strict real-time performance. To address this, we introduced a series of engineering optimizations, including post-training quantization and parallelizing loops to improve GPU utilization. After basic acceleration, our method achieves about 15 FPS at 1024×1024 resolution, basically meeting the deployment requirement. Meanwhile, the optimization

incurs negligible quality loss: on the same dataset, PSNR only drops from 30.16 to 29.99, and SSIM from 0.940 to 0.934. Considering further model simplification and distillation, additional optimization headroom is expected. Therefore, our future work will focus on leveraging techniques such as knowledge distillation to further accelerate inference speed, thereby enhancing the practical application value of our approach in real-world scenarios.

Table 2 Comparisons of efficiency and memory consumption. "w/o acceleration" indicates the method without any optimization, while "w/ acceleration" denotes the method after engineering optimization

Method	Efficiency (ms/frame)	GPU memory (MB)
GPS-Gaussian	105.78	4846
GPS-Gaussian+	100.49	4318
MuRF	6759.01	79528
MVSplat	1039.54	31160
DepthSplat	1448.68	41366
Ours (w/o acceleration)	270.34	11742
Ours (w/ acceleration)	64.85	10166

4.7 Ablation studies

We train our model using various combinations of modules and evaluate the rendered results on 500 samples of the THumanSit dataset. Table 3 provides a quantitative analysis. In the following sections, we will discuss the impact of different modules in detail.

Table 3 Ablation study of different modules. We report the average PSNR, SSIM, and LPIPS on our collected dataset. The best performance is marked in bold

Model	PSNR (dB) ↑	SSIM ↑	LPIPS ↓
Full model	**28.31**	**0.958**	**0.0590**
w/o depth refinement	27.92	0.953	0.0665
w/o self attention	27.89	0.954	0.0664
w/o soft attention	27.69	0.952	0.0691
w/o dual-branch compositing	27.25	0.952	0.0592

Monocular depth feature. We first remove the integration of features from the monocular depth estimator as introduced in Section 3.1. By incorporating monocular geometric features, this module significantly enhances the accuracy of depth estimation in occluded regions. As shown in Figure 8, removing this module leads to severe distortions in occluded facial areas and noticeable artifacts along clothing boundaries.

Attention-driven information fusion. This module integrates a self-attention encoder with a soft-attention decoder. The former is designed to enhance the network's feature extraction capabilities, while the latter improves its capacity to aggregate information across different scales. We conducted ablation experiments by removing each component individually. As illustrated in Figure 8, the absence of either module weakens the network's ability to handle projection errors caused by depth inaccuracies during the fusion process, resulting in noticeable ghosting artifacts.

Dual-branch compositing module. The dual-branch compositing module is crucial for correcting severe image degradation that arises from failures in depth estimation. As shown in the third row of Figure 8, due to severe self-occlusions induced by complex human poses, the depth estimation module suffers from large-scale failures. Consequently, removing the dual-branch compositing module leads to pronounced limb disconnections, severely degrading the overall rendering quality. Conversely, by leveraging the stable



Figure 8 Ablation study of our key modules. Our attention-based fusion network improves the discrimination of effective information in occluded regions. The depth refinement module enhances estimation accuracy and stability, reducing errors in ill-posed areas. Meanwhile, the dual-branch compositing module prevents artifacts and holes caused by depth estimation failures, ensuring a stable and robust final rendering. Key areas of comparison are highlighted with red rectangles.

geometric cues from the PSV, our complete pipeline effectively rectifies these structural anomalies, ensuring a robust and structurally intact novel view synthesis.

4.8 Analysis of the dual-branch compositing

We conduct a dedicated analysis to validate the design choices of our dual-branch compositing module. An important hyper-parameter is the number of depth planes D used in the plane-sweep volume. Table 4 reports the rendering quality under varying D.

Table 4 Ablation on the number of PSV depth planes

Number of depth planes	PSNR (dB) ↑	SSIM ↑	LPIPS ↓	Efficiency (ms) ↓
8	27.36	0.924	0.1023	227.88
16	27.30	0.951	0.0744	241.85
32	28.20	0.963	0.0568	270.34
64	27.45	0.953	0.0714	282.24

The results reveal a non-monotonic trend. With only 8 or 16 planes, the discrete depth sampling is too sparse to capture accurate correspondences; most warped pixels cannot find a correct match, and the module barely contributes to artifact correction, resulting in almost identical, low performance. When 64 planes are used, performance drops again despite the denser sampling. This counter-intuitive behavior is caused by the architecture of our 2D U-Net: the first convolutional layer compresses the input from $6D+3$ channels to merely 96 channels. For $D=64$, the input has 387 channels, so a substantial portion of the information is discarded at the very first layer before being processed by the rest of the network, introducing noise rather than useful cues. Consequently, the 32-plane configuration achieves the optimal balance between sampling density and information retention, yielding the best result. Regarding computational efficiency, the trend follows the expected pattern—fewer layers result in faster processing, as less input data needs to be warped

and processed.

To further validate the design of our blending strategy, we compare three alternative approaches using the same lightweight 2D U-Net backbone: (1) Confidence-based blending, which predicts per-pixel confidence scores for each PSV plane and outputs a confidence-weighted average; (2) Lightweight refinement without PSV, which directly refines $\boldsymbol{I}_f$ using the identical U-Net but without any PSV input; and (3) our Alpha compositing, which treats $\boldsymbol{I}_f$ as the background and selectively composites PSV foreground patches via predicted blending weights and opacity.

As shown in Table 5, Alpha compositing significantly outperforms both alternatives. Lightweight refinement without PSV achieves the worst performance (PSNR 23.19 dB). Deprived of any geometric prior, this lightweight network cannot distinguish genuine scene structure from artifacts and merely smooths the input, leaving structural distortions such as limb deformations unaddressed. Confidence-based blending achieves better results (PSNR 26.49 dB) by leveraging the geometrically stable PSV information. However, because it operates by soft-averaging across depth planes, when no single plane receives a high confidence score, which is a common situation in occluded or error-prone regions, the output degenerates into a weighted average of multiple misaligned layers, introducing blur and, in some cases, amplifying the original artifacts. Moreover, because confidence scores are normalized across planes, the network is forced to distribute its weight across all candidates, making it difficult to fully exclude irrelevant information. In contrast, our Alpha compositing bypasses this limitation by treating $\boldsymbol{I}_f$ as the primary background and using the PSV only to generate a foreground patch composited via the predicted opacity $\boldsymbol{\sigma}$. In well-reconstructed regions, $\boldsymbol{\sigma}$ remains low, fully preserving the sharp texture of $\boldsymbol{I}_f$; only in artifact regions does the network increase $\boldsymbol{\sigma}$ to incorporate the structurally stable PSV foreground. This mechanism avoids the forced averaging in confidence-based blending and explains the significant quality margin observed in both metrics and visual results.

Table 5 Comparison of different blending strategies. All variants share the same 2D U-Net backbone. The best result is in bold

Method	PSNR (dB) ↑	SSIM ↑	LPIPS ↓
Confidence-based blending	26.49	0.949	0.0808
Lightweight refinement (w/o PSV)	23.19	0.937	0.1056
Alpha compositing (Ours)	**28.20**	**0.963**	**0.0568**

4.9 Analysis of wide-baseline robustness

Our method relies on stereo depth estimation to warp the source views to the target viewpoint, where the depth is directly converted from the predicted disparity. Any failure in disparity estimation, therefore, directly compromises subsequent warping and fusion. We evaluate the accuracy of the estimated disparity under different viewpoint separations and demonstrate the difference in rendering quality across baselines, as reported in Table 6.

As shown in Table 6, at 48° the network still produces reasonably usable disparity with a 3-pixel accuracy of 0.75. At 72°, the accuracy drops substantially, yet some valid correspondences can still be extracted. At

Table 6 Disparity estimation accuracy under wide baselines. EPE denotes the end-point error in pixels; 1 px and 3 px denote the fraction of pixels with error below 1 and 3 pixels, respectively

Baseline angle	EPE (px)	1 px accuracy	3 px accuracy
24°	1.29	0.70	0.92
48°	3.48	0.42	0.75
72°	7.75	0.21	0.50
96°	–	–	–

96°, the common visible area between the two input views becomes too small to establish reliable matches, and the stereo matching module fails entirely. This degradation is an inherent limitation of stereo matching under extreme viewpoint separation. Overall, we have validated the effectiveness of our method under moderate baselines (e.g., 24°–48°), and extending its capability to wider baselines represents an important direction for future work.

4.10 Analysis of failure cases

While our method demonstrates robust performance across diverse scenarios, it can be challenged by several extreme conditions, including severe calibration errors, thin structures, extreme viewpoint extrapolation, and highly complex object interactions.

When the camera calibration is perturbed (20 cm translation error in opposite directions for the two input views as in Figure 9(a)), holes and artifacts appear in the rendered results. This occurs because the mismatch between the actual and assumed baselines leads to incorrect depth estimates even when flow is accurate, degrading the fused image $\boldsymbol{I}_f$; meanwhile, the PSV module, relying on the same erroneous calibration, can no longer provide accurate geometric cues for artifact correction. In the case of extreme extrapolation, where the target viewpoint lies far outside the input views (36° to the right of the right view in Figure 9(b)), only one side of the scene is observed. Even with accurate depth on the input side, the unseen region cannot be reconstructed, and the PSV lacks the necessary multi-view information to fill the missing content, leading to severe artifacts. For thin structures such as narrow garment straps, depth estimation tends to be unreliable, and the discrete depth planes of the PSV often fail to sample the exact geometry; as a result, these structures may appear broken or vanish in the rendered output, as shown in Figure 9(c). Regarding highly complex object interactions, while our method successfully handles a range of human-object interactions in the current benchmarks, extremely complex scenarios with extensive single-view or mutual occlusions remain challenging. In such cases, large portions of the scene are visible from only one view or completely occluded, which severely destabilizes both depth estimation and the attention-based fusion, and is difficult to refine with the PSV. Addressing these challenges represents an important direction for our future work toward more robust novel view synthesis.



Figure 9 Visualization of failure cases under different conditions. In each case, the left image shows our rendered result and the right image shows the ground truth.

5 Conclusion

This paper proposes a method for robust and stable human novel view synthesis from a single image pair. Our approach integrates an attention-driven information fusion module with a dual-branch compositing module, which work together to substantially enhance rendering quality and geometric stability. Extensive experiments demonstrate that our method outperforms existing state-of-the-art human NVS methods. Its effectiveness is validated on the challenging THumanSit and Dual-GS datasets and our own captured real-world data, where it shows robust generalization to unseen scenes and capably models complex human-object interactions. For future work, we plan to further enhance computational efficiency by accelerating

inference speed and reducing information redundancy. These optimizations are critical for moving our method towards practical, real-world deployment.

Declaration of competing interest

The authors declare that they have no conflict of interest.

CRediT author statement

Kaijin ZHAO: Conceptualization, Methodology, Writing-Original draft. **Chuyue ZHAO:** Writing-Review & Editing. **Zhiheng JIANG:** Data curation, Methodology. **Guoqing ZHOU:** Validation. **Qing WANG:** Conceptualization, Supervision.

References

1 Chen S E, Williams L. View interpolation for image synthesis. In: Proceedings of the 20th Annual Conference on Computer Graphics and Interactive Techniques. Anaheim, CA, ACM, 1993. 279–288
DOI: 10.1145/166117.166153

2 Levoy M, Hanrahan P. Light field rendering. In: Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques. ACM, 1996. 31–42
DOI: 10.1145/237170.237199

3 Kerbl B, Kopanas G, Leimkuehler T, Drettakis G. 3D Gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics, 2023, 42(4): 1–14
DOI: 10.1145/3592433

4 Charatan D, Li S L, Tagliasacchi A, Sitzmann V. PixelSplat: 3D Gaussian splats from image pairs for scalable generalizable 3D reconstruction. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA, IEEE, 2024. 19457–19467
DOI: 10.1109/cvpr52733.2024.01840

5 Chen Y D, Xu H F, Zheng C X, Zhuang B H, Pollefeys M, Geiger A, Cham T J, Cai J F. MVSplat: efficient 3D Gaussian splatting from Sparse multi-view images. In: Computer Vision – ECCV 2024. Cham: Springer, 2025. 370–386
DOI: 10.1007/978-3-031-72664-4_21

6 Mildenhall B, Srinivasan P P, Tancik M, Barron J T, Ramamoorthi R, Ng R. NeRF: representing scenes as neural radiance fields for view synthesis. In: Computer Vision – ECCV 2020. Cham: Springer, 2020. 405–421
DOI: 10.1007/978-3-030-58452-8_24

7 Wang Q Q, Wang Z C, Genova K, Srinivasan P, Zhou H, Barron J T, Martin-Brualla R, Snavely N, Funkhouser T. IBRNet: learning multi-view image-based rendering. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN, USA, IEEE, 2021. 4688–4697
DOI: 10.1109/cvpr46437.2021.00466

8 Xu H F, Chen A P, Chen Y D, Sakaridis C, Zhang Y L, Pollefeys M, Geiger A, Yu F. MuRF: multi-baseline radiance fields. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA, IEEE, 2024. 20041–20050
DOI: 10.1109/cvpr52733.2024.01894

9 Hu X R, Yang K F, Luo R Q, Peng T, Liu J P. Learning monocular face reconstruction from in the wild images using rotation cycle consistency. Virtual Reality & Intelligent Hardware, 2025, 7(4): 379–392
DOI: 10.1016/j.vrih.2022.08.014

10 Su Z Q, Zhou T S, Li K, Brady D, Liu Y B. View Synthesis from multi-view RGB data using multilayered representation and volumetric estimation. Virtual Reality & Intelligent Hardware, 2020, 2(1): 43–55
DOI: 10.1016/j.vrih.2019.12.001

11 Wang Y R, Huang Z H, Zhu H, Li W, Cao X, Yang R G. Interactive free-viewpoint video generation. Virtual Reality & Intelligent Hardware, 2020, 2(3): 247–260
DOI: 10.1016/j.vrih.2020.04.004

12 Lin H T, Peng S D, Xu Z, Yan Y Z, Shuai Q, Bao H J, Zhou X W. Efficient neural radiance fields for interactive free-viewpoint video. In: SIGGRAPH Asia 2022. ACM, 2022. 1–9
DOI: 10.1145/3550469.3555376

13 Szymanowicz S, Rupprecht C, Vedaldi A. Splatter image: ultra-fast single-view 3D reconstruction. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA, IEEE, 2024. 10208–10217
DOI: 10.1109/cvpr52733.2024.00972

14 Wewer C, Raj K, Ilg E, Schiele B, Lenssen J E. LatentSplat: autoencoding variational Gaussians for fast generalizable 3D reconstruction. In: Computer Vision – ECCV 2024. Cham: Springer, 2025. 456–473

DOI: 10.1007/978-3-031-73021-4_27

15 Zhang J J, Zhang Y X, Zhang H W, Zhou X, Zhou B Y, Shao R Z, Hu Z H, Liu Y B. Ins-HOI: instance aware human-object interactions recovery. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025, 47(11): 9655–9672
DOI: 10.1109/tpami.2025.3588268

16 Jiang Y H, Shen Z H, Hong Y, Guo C C, Wu Y Z, Zhang Y L, Yu J Y, Xu L. Robust dual Gaussian splatting for immersive human-centric volumetric videos. ACM Transactions on Graphics, 2024, 43(6): 1–15
DOI: 10.1145/3687926

17 Zhou T H, Tucker R, Flynn J, Fyffe G, Snavely N. Stereo magnification: learning view synthesis using multiplane images. ACM Transactions on Graphics, 2018, 37(4): 1–12
DOI: 10.1145/3197517.3201323

18 Mildenhall B, Srinivasan P P, Ortiz-Cayon R, Kalantari N K, Ramamoorthi R, Ng R, Kar A. Local light field fusion: practical view synthesis with prescriptive sampling guidelines. ACM Transactions on Graphics, 2019, 38(4): 1–14
DOI: 10.1145/3306346.3322980

19 Li J X, Feng Z J, She Q, Ding H H, Wang C H, Lee G H. MINE: towards continuous depth MPI with NeRF for novel view synthesis. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, QC, Canada, IEEE, 2021. 12558–12568
DOI: 10.1109/iccv48922.2021.01235

20 Yoon G J, Jung G, Song J, Yoon S M. View synthesis with multiplane images from computationally generated RGB-D light fields. Engineering Applications of Artificial Intelligence, 2024, 132: 107930
DOI: 10.1016/j.engappai.2024.107930

21 Cao A, Rockwell C, Johnson J. FWD: real-time novel view synthesis with forward warping and depth. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA, IEEE, 2022. 15692–15703
DOI: 10.1109/cvpr52688.2022.01526

22 Li D Q, Huang S S, Shen T Y, Huang H. Dynamic view synthesis with spatio-temporal feature warping from sparse views. In: Proceedings of the 31st ACM International Conference on Multimedia. New York, NY, USA, ACM, 2023. 1565–1576
DOI: 10.1145/3581783.3612419

23 Yao Y, Luo Z X, Li S W, Fang T, Quan L. MVSNet: depth inference for unstructured multi-view stereo. In: Computer Vision – ECCV 2018. Cham: Springer, 2018. 785–801
DOI: 10.1007/978-3-030-01237-3_47

24 Xu H F, Peng S Y, Wang F, Blum H, Barath D, Geiger A, Pollefeys M. DepthSplat: connecting Gaussian splatting and depth. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN, USA, IEEE, 2025. 16453–16463
DOI: 10.1109/cvpr52734.2025.01534

25 Zhang C R, Zou Y S, Li Z L, Yi M M, Wang H Q. TranSplat: generalizable 3D Gaussian splatting from sparse multi-view images with transformers. Proceedings of the AAAI Conference on Artificial Intelligence, 2025, 39(9): 9869–9877
DOI: 10.1609/aaai.v39i9.33070

26 Yang L H, Kang B Y, Huang Z L, Xu X G, Feng J S, Zhao H S. Depth anything: unleashing the power of large-scale unlabeled data. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA, IEEE, 2024. 10371–10381
DOI: 10.1109/cvpr52733.2024.00987

27 Yang L H, Kang B Y, Huang Z L, Zhao Z, Xu X G, Feng J S, Zhao H S. Depth anything V2. In: Advances in Neural Information Processing Systems 37. Vancouver, BC, Canada, Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024. 21875–21911
DOI: 10.52202/079017-0688

28 Zhang S J, Fei X, Liu F F, Song H X, Duan Y Q. Gaussian graph network: learning efficient and generalizable Gaussian representations from multi-view images. In: Advances in Neural Information Processing Systems 37. Vancouver, BC, Canada, Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024. 50361–50380
DOI: 10.52202/079017-1595

29 Wang Y S, Huang T X, Chen H L, Lee G H. FreeSplat: generalizable 3D Gaussian splatting towards free view synthesis of indoor scenes. In: Advances in Neural Information Processing Systems 37. Vancouver, BC, Canada, Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024. 107326–107349
DOI: 10.52202/079017-3409

30 Wang Y S, Huang T X, Chen H L, Lee G H. FreeSplat++: generalizable 3D Gaussian splatting for efficient indoor scene reconstruction. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2026, 48(7): 7749–7765
DOI: 10.1109/tpami.2026.3665771

31 Zhang K, Bi S, Tan H, Xiangli Y, Zhao N X, Sunkavalli K, Xu Z X. GS-LRM: large reconstruction model for 3D Gaussian splatting. In: Computer Vision – ECCV 2024. Cham: Springer, 2025. 1–19
DOI: 10.1007/978-3-031-72670-5_1

32 Chen A P, Xu H F, Esposito S, Tang S Y, Geiger A. LaRa: efficient large-baseline radiance fields. In: Computer Vision – ECCV 2024. Cham: Springer, 2025. 338–355

DOI: 10.1007/978-3-031-73247-8_20

33 Jin H A, Jiang H W, Tan H, Zhang K, Bi S, Zhang T Y, Luan F J, Snavely N, Xu Z X. LVSM: a large view synthesis model with minimal 3D inductive bias. In: The Thirteenth International Conference on Learning Representations. Singapore, IMLS, 2025. 60001–60021
https://proceedings.iclr.cc/paper_files/paper/2025/file/9676c5283df26cabca412ca66b164a7d-Paper-Conference.pdf

34 Barron J T, Mildenhall B, Tancik M, Hedman P, Martin-Brualla R, Srinivasan P P. Mip-NeRF: a multiscale representation for anti-aliasing neural radiance fields. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, QC, Canada, IEEE, 2021. 5835–5844
DOI: 10.1109/iccv48922.2021.00580

35 Barron J T, Mildenhall B, Verbin D, Srinivasan P P, Hedman P. Mip-NeRF 360: unbounded anti-aliased neural radiance fields. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA, IEEE, 2022. 5460–5469
DOI: 10.1109/cvpr52688.2022.00539

36 Barron J T, Mildenhall B, Verbin D, Srinivasan P P, Hedman P. Zip-NeRF: anti-aliased grid-based neural radiance fields. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France, IEEE, 2023. 19640–19648
DOI: 10.1109/iccv51070.2023.01804

37 Han X F, Liu Z, Nan H, Zhao K, Zhao D J, Jin X D. PW-NeRF: progressive wavelet-mask guided neural radiance fields view synthesis. Image and Vision Computing, 2024, 147: 105073
DOI: 10.1016/j.imavis.2024.105073

38 Zhao C Y, Huang X, Wang X, Zhou G Q, Wang Q. Phase shift guided dynamic view synthesis from monocular video. Image and Vision Computing, 2025, 162: 105702
DOI: 10.1016/j.imavis.2025.105702

39 Müller T, Evans A, Schied C, Keller A. Instant neural graphics primitives with a multiresolution hash encoding. ACM Transactions on Graphics, 2022, 41(4): 1–15
DOI: 10.1145/3528223.3530127

40 Jain A, Tancik M, Abbeel P. Putting NeRF on a diet: semantically consistent few-shot view synthesis. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, QC, Canada, IEEE, 2021. 5865–5874
DOI: 10.1109/iccv48922.2021.00583

41 Chen A P, Xu Z X, Wei X Y, Tang S Y, Su H, Geiger A. Dictionary fields: learning a neural basis decomposition. ACM Transactions on Graphics, 2023, 42(4): 1–12
DOI: 10.1145/3592135

42 Somraj N, Karanayil A, Soundararajan R. SimpleNeRF: regularizing sparse input neural radiance fields with simpler solutions. In: SIGGRAPH Asia 2023. New York, NY, USA, ACM, 2023. 1–11
DOI: 10.1145/3610548.3618188

43 Wang G C, Chen Z X, Loy C C, Liu Z W. SparseNeRF: distilling depth ranking for few-shot novel view synthesis. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France, IEEE, 2023. 9031–9042
DOI: 10.1109/iccv51070.2023.00832

44 Truong P, Rakotosaona M J, Manhardt F, Tombari F. SPARF: neural radiance fields from sparse and noisy poses. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, BC, Canada, IEEE, 2023. 4190–4200
DOI: 10.1109/cvpr52729.2023.00408

45 Xu Q G, Xu Z X, Philip J, Bi S, Shu Z X, Sunkavalli K, Neumann U. Point-NeRF: point-based neural radiance fields. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA, IEEE, 2022. 5428–5438
DOI: 10.1109/cvpr52688.2022.00536

46 Zhao K J, Huang X, Zhou G Q, Wang Q. Generalizable occlusion-aware human novel view synthesis from image pairs. IEEE Transactions on Circuits and Systems for Video Technology, 2026, 36(6): 8021–8034
DOI: 10.1109/tcsvt.2026.3656967

47 Johari M M, Lepoittevin Y, Fleuret F. GeoNeRF: generalizing NeRF with geometry priors. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA, IEEE, 2022. 18344–18347
DOI: 10.1109/cvpr52688.2022.01782

48 Chen A P, Xu Z X, Zhao F Q, Zhang X S, Xiang F B, Yu J Y, Su H. MVSNeRF: fast generalizable radiance field reconstruction from multi-view stereo. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, QC, Canada, IEEE, 2021. 14104–14113
DOI: 10.1109/iccv48922.2021.01386

49 Liu Y, Peng S D, Liu L J, Wang Q Q, Wang P, Theobalt C, Zhou X W, Wang W P. Neural rays for occlusion-aware image-based rendering. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA, IEEE, 2022. 7814–7823
DOI: 10.1109/cvpr52688.2022.00767

50 Loper M, Mahmood N, Romero J, Pons-Moll G, Black M J. SMPL: a skinned multi-person linear model. ACM Transactions on Graphics, 2015, 34(6): 1–16

DOI: 10.1145/2816795.2818013

51 Pavlakos G, Choutas V, Ghorbani N, Bolkart T, Osman A A, Tzionas D, Black M J. Expressive body capture: 3D hands, face, and body from a single image. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA, IEEE, 2019. 10967–10977
DOI: 10.1109/cvpr.2019.01123

52 Kwon Y, Fang B L, Lu Y X, Dong H Y, Zhang C, Carrasco F V, Mosella-Montoro A, Xu J J, Takagi S, Kim D, Prakash A, De la Torre F. Generalizable human Gaussians for sparse view synthesis. In: Computer Vision – ECCV 2024. Cham: Springer, 2025. 451–468
DOI: 10.1007/978-3-031-73229-4_26

53 Li Z, Zheng Z R, Wang L Z, Liu Y B. Animatable Gaussians: learning pose-dependent Gaussian maps for high-fidelity human avatar modeling. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA, IEEE, 2024. 19711–19722
DOI: 10.1109/cvpr52733.2024.01864

54 Fan J L, Lv X D, Zeng X P, Bao Z Y, He Z W, Gao M Y. Unified volumetric avatar: enabling flexible editing and rendering of neural human representations. Image and Vision Computing, 2025, 153: 105345
DOI: 10.1016/j.imavis.2024.105345

55 Chen J C, Yi W T, Ma L Q, Jia X, Lu H C. GM-NeRF: learning generalizable model-based neural radiance fields from multi-view images. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, BC, Canada, IEEE, 2023. 20648–20658
DOI: 10.1109/cvpr52729.2023.01978

56 Pan X, Yang Z X, Ma J X, Zhou C, Yang Y. TransHuman: a transformer-based human representation for generalizable neural human rendering. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France, IEEE, 2023. 3521–3532
DOI: 10.1109/iccv51070.2023.00328

57 Shao R Z, Zhang H W, Zhang H, Chen M J, Cao Y P, Yu T, Liu Y B. DoubleField: bridging the neural surface and radiance fields for high-fidelity human reconstruction and rendering. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA, IEEE, 2022. 15851–15861
DOI: 10.1109/cvpr52688.2022.01541

58 Zheng R C, Li P, Wang H Q, Yu T. Learning visibility field for detailed 3D human reconstruction and relighting. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, BC, Canada, IEEE, 2023. 216–226
DOI: 10.1109/cvpr52729.2023.00029

59 Weng C, Curless B, Srinivasan P P, Barron J T, Kemelmacher-Shlizerman I. HumanNeRF: free-viewpoint rendering of moving people from monocular video. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA, IEEE, 2022. 16189–16199
DOI: 10.1109/cvpr52688.2022.01573

60 Kocabas M, Chang J R, Gabriel J, Tuzel O, Ranjan A. HUGS: human Gaussian splats. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA, IEEE, 2024. 505–515
DOI: 10.1109/cvpr52733.2024.00055

61 Xu Z, Peng S D, Lin H T, He G Z, Sun J M, Shen Y J, Bao H J, Zhou X W. 4K4D: real-time 4D view synthesis at 4K resolution. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA, IEEE, 2024. 20029–20040
DOI: 10.1109/cvpr52733.2024.01893

62 Qian Z Y, Wang S F, Mihajlovic M, Geiger A, Tang S Y. 3DGS-avatar: animatable avatars via deformable 3D Gaussian splatting. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA, IEEE, 2024. 5020–5030
DOI: 10.1109/cvpr52733.2024.00480

63 Somraj N, Choudhary K, Mupparaju S H, Soundararajan R. Factorized motion fields for fast sparse input dynamic view synthesis. In: Special Interest Group on Computer Graphics and Interactive Techniques Conference. Denver, CO, USA, ACM, 2024. 1–12
DOI: 10.1145/3641519.3657498

64 Zheng S Y, Zhou B Y, Shao R Z, Liu B N, Zhang S P, Nie L Q, Liu Y B. GPS-Gaussian: generalizable pixel-wise 3D Gaussian splatting for real-time human novel view synthesis. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA, IEEE, 2024. 19680–19690
DOI: 10.1109/cvpr52733.2024.01861

65 Sun Y, Zhang S P, Wang C Y, Zheng S Y, Li Z L, Ji X Y. Generalizable and relightable Gaussian splatting for human novel view synthesis. 2025, arXiv: 2505.21502
https://arxiv.org/abs/2505.21502

66 Zhou B Y, Zheng S Y, Tu H Z, Shao R Z, Liu B N, Zhang S P, Nie L Q, Liu Y B. GPS-Gaussian: generalizable pixel-wise 3D Gaussian splatting for real-time human-scene rendering from sparse views. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025.
DOI: 10.1109/tpami.2025.3561248

67 Lipson L, Teed Z, Deng J. RAFT-stereo: multilevel recurrent field transforms for stereo matching. In: 2021 International Conference on

3D Vision (3DV). London, United Kingdom, IEEE, 2021. 218–227
DOI: 10.1109/3dv53792.2021.00032

68 Ulyanov D, Vedaldi A, Lempitsky V. Instance normalization: the missing ingredient for fast stylization. 2017, arXiv: 1607.08022v3
https://arxiv.org/abs/1607.08022v3

69 Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. 2015, arXiv: 1502.03167
https://arxiv.org/abs/1502.03167

70 Liu Z, Lin Y T, Cao Y, Hu H, Wei Y X, Zhang Z, Lin S, Guo B N. Swin transformer: hierarchical vision transformer using shifted windows. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, QC, Canada, IEEE, 2021. 9992–10002
DOI: 10.1109/iccv48922.2021.00986

71 Guo C L, Yan Q X, Anwar S, Cong R M, Ren W Q, Li C Y. Image dehazing transformer with transmission-aware 3D position embedding. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA, IEEE, 2022. 5802–5810
DOI: 10.1109/cvpr52688.2022.00572

72 Zamir S W, Arora A, Khan S, Hayat M, Khan F S, Yang M H, Shao L. Learning enriched features for real image restoration and enhancement. In: Computer Vision – ECCV 2020. Cham: Springer, 2020. 492–511
DOI: 10.1007/978-3-030-58595-2_30

73 Wang Z, Bovik A C, Sheikh H R, Simoncelli E P. Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing, 2004, 13(4): 600–612
DOI: 10.1109/tip.2003.819861

74 Loshchilov I, Hutter F. Decoupled weight decay regularization. 2019, arXiv: 1711.05101
https://arxiv.org/abs/1711.05101

75 Lin S C, Yang L J, Saleemi I, Sengupta S. Robust high-resolution video matting with temporal guidance. In: 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Waikoloa, HI, USA, IEEE, 2022. 3132–3141
DOI: 10.1109/wacv51458.2022.00319

76 Zhang R, Isola P, Efros A A, Shechtman E, Wang O. The unreasonable effectiveness of deep features as a perceptual metric. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, IEEE, 2018. 586–595
DOI: 10.1109/cvpr.2018.00068