Pattern Recognition Letters 210 (2026) 83–88

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Fast generalizable human novel view synthesis from image pairs

Kaijin Zhao, Bingjie Song, Chen Gong, Guoqing Zhou, Qing Wang *

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

ARTICLE INFO

Editor: Yonghuai Liu

Keywords:
Novel view synthesis
3D Gaussian splatting
Generalizable rendering

ABSTRACT

We propose an approach for efficient and generalizable human novel view synthesis that takes an image pair as input. Existing generalizable approaches often employ complex network architectures, resulting in significant memory overhead and inference latency that hinder their practical deployment. Furthermore, current human reconstruction methods typically rely on parametric human templates or direct back-projection based on estimated depth. Such excessive dependence on priors often compromises rendering fidelity. To address these challenges, we propose a novel framework based on 3D Gaussian Splatting (3DGS). By leveraging the geometric consistency of depth information between the input views, our method obtains a more reliable understanding of the 3D scene. Additionally, we integrate monocular features with stereo cues to improve the network's ability to handle occluded regions. We further decouple the processing of high-frequency texture details from low-frequency semantic information and implement a Gaussian pruning strategy. These designs achieve a rendering speed of approximately 56 ms for 2K-resolution images without any engineering optimization (*e.g.*, TensorRT). Extensive experiments on THuman, THuman-Sit, Dual-GS and DNA-Rendering datasets demonstrate that our method achieves state-of-the-art performance in both qualitative and quantitative evaluations.

1. Introduction

Novel View Synthesis (NVS) aims to synthesize dense views from a limited set of input views. By eliminating the need for complex dense-capture systems, it holds tremendous potential for applications such as 3D video conferencing, immersive sports broadcasting, and free viewpoint video (FVV), thereby attracting widespread attention. Early approaches typically relied on view interpolation or depth-based warping to synthesize target views. However, these methods still require relatively dense input and struggle to produce high-quality syntheses under sparse input settings.

In recent years, neural rendering approaches, such as NeRF [1], have made significant strides in novel view synthesis. Although early NeRFs required per-scene optimization, feed-forward methods [2–4] were later introduced to enhance rendering speed and generalization. Nevertheless, the need for per-view ray marching and dense spatial sampling still constrains the inference efficiency of NeRF-based methods, limiting their practical deployment.

Subsequently, 3D Gaussian Splatting (3DGS) [5] has emerged with improved efficiency and rendering quality. Although the original 3DGS lacks generalization due to its gradient-based optimization of Gaussian primitives, later works [6–8] enabled feed-forward 3DGS via depth estimation. However, these approaches introduce complex depth-estimation and regression networks, diminishing the efficiency benefits of 3DGS. Furthermore, without error-correction mechanisms, inaccuracies in the initial estimates are propagated to the final representation, resulting in a noticeable drop in visual quality.

Simultaneously, human novel view synthesis has attracted increasing attention. Yet many existing methods rely on parametric human templates (*e.g.*, SMPL [9]), which struggle to capture complex, loose-fitting clothing or intricate hairstyles and complicate human–object interactions, thereby limiting their practical applicability. Alternatively, some approaches perform human NVS from monocular videos using techniques such as deformation fields [10], but their limited generalization limits their applicability in real-world scenarios where pre-recorded complete videos are unavailable, such as immersive sports broadcasting or 3D video conferencing.

To address these challenges, we propose a fast and generalizable human novel view synthesis method that operates on an image pair. We first estimate the depth maps of the input views via stereo matching. Rather than relying solely on learned features to infer geometric correction, we explicitly evaluate the reliability of the estimated geometry through bidirectional depth reprojection. The resulting cross-view depth discrepancy indicates unreliable regions and retains the magnitude of

* Corresponding author.
E-mail addresses: kjzhao@mail.nwpu.edu.cn (K. Zhao), sbjchn@mail.nwpu.edu.cn (B. Song), gchensfl@mail.nwpu.edu.cn (C. Gong), zhouguoqing@nwpu.edu.cn (G. Zhou), qwang@nwpu.edu.cn (Q. Wang).

https://doi.org/10.1016/j.patrec.2026.09.036
Received 28 March 2026; Received in revised form 19 September 2026; Accepted 23 September 2026
Available online 26 September 2026
0167-8655/© 2026 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

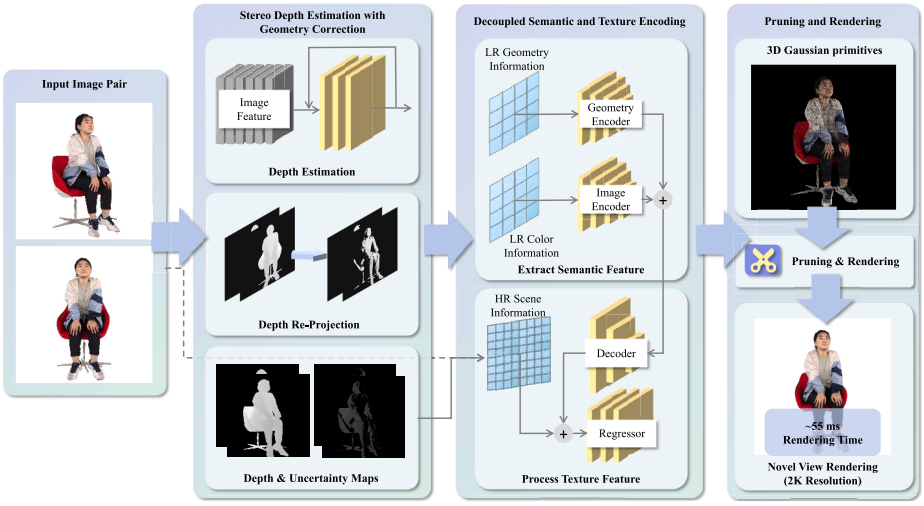


Fig. 1. Overview of Our Pipeline. We first estimate depth maps and assess their uncertainty and occlusion probability through a mutual re-projection consistency check. To balance efficiency and quality, we extract semantic features from low-resolution and subsequently integrate them with high-resolution scene information to preserve high-frequency texture details. Finally, 3D Gaussian attributes are regressed from the enriched features, followed by a pruning and rendering process to generate the final output.

the observed geometric inconsistency, and is directly provided to the subsequent correction network to guide the refinement of Gaussian geometry.

Furthermore, to enable efficient high-resolution reconstruction, we decouple low-resolution semantic reasoning from high-resolution texture processing. Computationally expensive semantic and geometric feature extraction is performed at a fixed low resolution, while the resulting features are upsampled and combined with the original-resolution image and geometry information through a lightweight high-resolution branch. This design preserves fine texture and boundary information while avoiding expensive full-resolution feature extraction before Gaussian regression and rendering.

In summary, our contributions are as follows:

- We propose a novel 3DGS-based framework for human novel-view synthesis, capable of achieving fast, high-quality rendering at 2K resolution.
- We integrate uncertainty-guided geometry correction with decoupled semantic and texture encoding, improving geometric and textural fidelity while enabling efficient high-resolution rendering.
- Our method outperforms the SOTA methods in multiple human datasets in both qualitative and quantitative evaluations.

2. Related work

2.1. Generalizable novel view synthesis

Feed-forward 3DGS methods commonly determine Gaussian positions from estimated depth. MVSplat [6] constructs a cost volume for depth regression. DepthSplat [7] and TranSplat [8] further incorporate monocular features to improve depth estimation. Nevertheless, residual errors in the estimated geometry can propagate to subsequent Gaussian reconstruction and degrade rendering quality.

Uncertainty has also been explored to improve geometric reconstruction. SparseRecon [11] estimates geometric confidence from multi-view depth consistency and uses it to selectively introduce monocular depth constraints during scene-specific optimization. UNG-GS [12] jointly optimizes a spatial uncertainty field with Gaussian primitives and uses the estimated uncertainty to regulate depth rendering and geometric refinement. More recently, CoSplat [13] incorporates depth uncertainty into a generalizable framework to perform confidence-aware multi-view feature aggregation and establish a consistent canonical representation. These studies demonstrate the benefit of uncertainty for geometry reasoning. Different from them, we directly use reprojection-derived uncertainty as an input to a feed-forward correction network, explicitly connecting geometric reliability with Gaussian position refinement.

Recent works have also explored more efficient representations and multi-resolution processing. HiSplat [14] adopts a hierarchical coarse-to-fine Gaussian representation, while latentSplat [15] performs scene representation and rendering in a compact latent space, followed by image decoding. In contrast, our method retains the standard Gaussian representation but separates low-resolution semantic reasoning from lightweight high-resolution texture processing to reduce the cost of high-resolution feature extraction.

2.2. Human novel view synthesis

Current human novel view synthesis methods typically leverage parametric human body models or estimated human depth maps to unproject Gaussian primitives. For instance, HumanSplat [16] uses SMPL to assist Gaussian-center prediction, while GHG [17] expands the human template into multiple scaffolds to better model loose clothing and hairstyles. GPS-Gaussian [18] establishes a feed-forward stereo-pair paradigm that estimates human depth and back-projects pixel-aligned Gaussian primitives into 3D space. GPS-Gaussian+ [19] further introduces epipolar attention and geometry regularization to improve cross-view consistency and predicts a depth residual to refine the initial geometry. EVA-Gaussian [20] similarly enhances cross-view geometry estimation and employs a recurrent feature refiner to alleviate rendering artifacts caused by inaccurate Gaussian positions. These approaches improve either the initial geometry estimation or the subsequent refinement, but the reliability of the estimated geometry is not explicitly exposed to the correction process. In contrast, our method explicitly measures the residual inconsistency through bidirectional depth reprojection and uses it as a reliability cue to guide subsequent geometry correction.

Per-scene methods such as HumanNeRF [21], 4K4D [22] and TAGA [23] achieve high-quality human reconstruction through scene-specific modeling or optimization, but require per-scene training and therefore differ from our feed-forward setting.

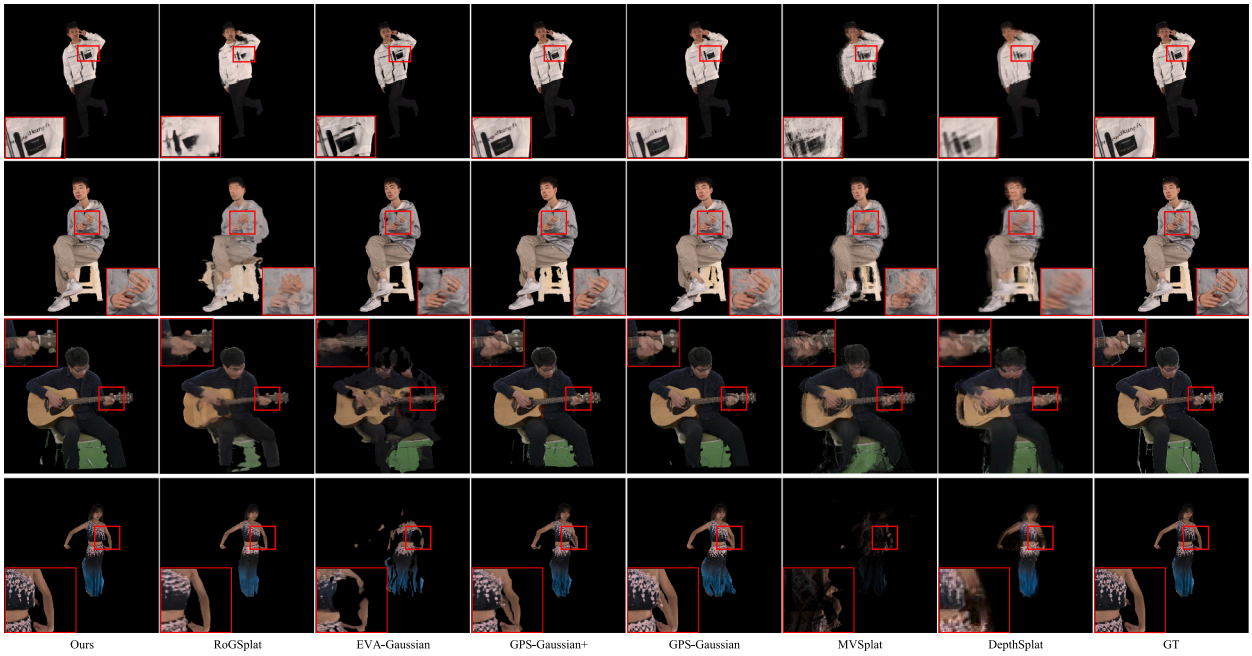


Fig. 2. Qualitative Results across Different Datasets. The first, second, third, and fourth rows correspond to results on THuman, THuman-Sit, Dual-GS, and DNA-Rendering, respectively.

3. Method

3.1. Stereo depth estimation

The overall pipeline is illustrated in Fig. 1. Our depth estimation framework is primarily built upon RAFT-Stereo [24]. We retain its main stereo-matching components, including 1D correlation construction along the epipolar direction, recurrent disparity refinement, convex upsampling, and sequence supervision, without introducing a new correlation formulation or recurrent update rule. In our implementation, the recurrent update is simplified to a single 1/8-resolution ConvGRU, and we use $K = 3$ recurrent iterations during both training and inference.

Given two input views V_1, V_2 with images $\mathbf{I}_1, \mathbf{I}_2$ and corresponding camera parameters, we first perform stereo rectification and use the rectified inputs thereafter. The shared encoder extracts context and correlation features, from which the 1D correlation volume is constructed as

$$\mathbf{C}_{ijk} = \sum_h \mathbf{F}_{ijh} \cdot \mathbf{G}_{ikh}, \qquad \mathbf{C} \in \mathbb{R}^{H\times W\times W}. \tag{1}$$

Starting from zero disparity, the recurrent unit iteratively samples the correlation volume around the current estimate and refines the disparity for K iterations.

The final disparity **disp** is converted to depth using the rectified stereo geometry:

$$\mathbf{D} = \frac{bf}{\mathbf{disp}}, \tag{2}$$

where b and f denote the rectified baseline and focal length, respectively, and **disp** accounts for the horizontal principal-point offset after rectification. We apply the shared stereo estimator in both directions to obtain the depth maps $\{\mathbf{D}_1, \mathbf{D}_2\}$ for the subsequent bidirectional consistency analysis.

3.2. Geometry correction

After obtaining the depth maps for the input views, Gaussian primitives can be directly unprojected into 3D space. However, relying solely on estimated depth maps often propagates estimation errors directly into the rendered images, resulting in artifacts such as disconnected limbs and floating edges. To further enhance the geometric accuracy of the reconstructed 3DGS, we propose a geometry correction mechanism. Specifically, given the input depths $\mathbf{D}_1, \mathbf{D}_2$, we first unproject the depth maps into 3D space using the rectified camera intrinsics and extrinsics:

$$\mathbf{P}_i = \mathbf{D}_i \cdot \mathbf{T}_i^{-1}\mathbf{K}_i^{-1}\mathbf{Coord}, \quad i \in \{1, 2\}, \tag{3}$$

where $\mathbf{P}_i$ denotes the 3D point clouds unprojected from $\mathbf{D}_i$, and **Coord** represents the homogeneous coordinate grid of the image plane. Subsequently, we project $\mathbf{P}_1$ and $\mathbf{P}_2$ onto view V_2 and V_1, respectively, obtaining the corresponding reprojected depths $\mathbf{D}'_1$ and $\mathbf{D}'_2$.

Ideally, the estimated depth should be consistent with reprojections from other views. We exploit the discrepancy to evaluate depth reliability and compute the uncertainty $\mathbf{R}_i$ as follows:

$$\mathbf{R}_i = |\mathbf{D}_i - \mathbf{D}'_i|. \tag{4}$$

This metric simultaneously characterizes the probability of occlusion. Leveraging this information, we enable the network to perceive both depth estimation errors and occlusion cues, thereby facilitating more effective correction of the reconstructed scene geometry. Specifically, the estimated depth $\mathbf{D}_i$, reprojected depth $\mathbf{D}'_i$, and uncertainty $\mathbf{R}_i$ are concatenated to form the geometry input $\mathbf{G}_i$. As detailed in Sec. 3.3, $\mathbf{G}_i$ is fused with the image features to predict the depth deformation M_δ, which refines the initial depth before determining the Gaussian centers.

3.3. Decoupled semantic and texture encoding

To enable efficient high-resolution reconstruction, we perform semantic and geometric feature extraction at a fixed 512×512 resolution and reserve only lightweight texture processing for the original resolution. Specifically, the image and geometry inputs are resized to 512×512, where DINOv2 [25] and the geometry encoder–decoder extract and fuse semantic and depth features. The resulting low-resolution feature is then upsampled and concatenated with the original-resolution image and geometry information, followed by a lightweight convolutional refinement module before Gaussian regression. This design preserves high-frequency texture and boundary information while reducing

Table 1
Quantitative Comparison. Quantitative results on THuman, THuman-Sit, Dual-GS, and DNA-Rendering at different output resolutions. MAE is reported on THuman and THuman-Sit, while FVD is reported on Dual-GS and DNA-Rendering.

Method	THuman				THuman-Sit				Dual-GS				DNA-Rendering			
	PSNR↑	SSIM↑	LPIPS↓	MAE↓	PSNR↑	SSIM↑	LPIPS↓	MAE↓	PSNR↑	SSIM↑	LPIPS↓	FVD↓	PSNR↑	SSIM↑	LPIPS↓	FVD↓
MVSplat (1024^2)	18.90	0.454	0.3478	0.5517	18.57	0.528	0.2800	0.4428	22.04	0.757	0.2056	886.41	15.23	0.471	0.3499	1195.01
RoGSplat (1024^2)	20.75	0.812	0.2556	0.0272	17.07	0.762	0.3168	0.0764	22.87	0.876	0.1758	912.67	23.22	0.847	0.1770	324.67
EVA-Gaussian (1024^2)	24.48	0.874	0.1798	0.0132	23.67	0.882	0.1774	0.0168	15.58	0.584	0.3364	2153.43	14.53	0.518	0.3785	851.38
GPS-Gaussian (1024^2)	22.28	0.845	0.1797	0.0082	20.40	0.840	0.1889	0.0180	24.33	0.893	0.1353	452.48	20.29	0.774	0.1947	389.74
GPS-Gaussian+ (1024^2)	24.90	0.900	0.1616	0.0090	23.34	0.904	0.1545	0.0134	23.37	0.885	0.1391	565.32	21.17	0.806	0.1921	405.50
Ours (1024^2)	**25.87**	**0.917**	**0.1365**	**0.0054**	**25.33**	**0.926**	**0.1235**	**0.0074**	**25.45**	**0.915**	**0.1205**	**411.68**	**23.99**	**0.871**	**0.1247**	**188.84**
EVA-Gaussian (2048^2)	21.24	0.828	0.2361	0.0144	21.41	0.857	0.2236	0.0189	15.78	0.640	0.2837	2216.75	14.85	0.644	0.3045	1258.87
GPS-Gaussian (2048^2)	22.48	0.837	0.2265	0.0080	20.52	0.844	0.2371	0.0177	24.38	0.917	0.1304	466.96	20.10	0.803	0.2024	395.96
GPS-Gaussian+ (2048^2)	24.78	0.871	0.2131	0.0092	23.37	0.887	0.2103	0.0135	23.56	0.907	0.1310	584.75	20.83	0.833	0.2022	400.71
Ours (2048^2)	**25.76**	**0.886**	**0.1812**	**0.0054**	**25.35**	**0.907**	**0.1748**	**0.0073**	**25.50**	**0.931**	**0.1161**	**438.95**	**23.48**	**0.880**	**0.1464**	**194.73**
DepthSplat (256^2)	20.08	0.653	0.1948	0.1161	18.23	0.626	0.2264	0.1186	20.21	0.695	0.2443	1193.57	20.66	0.594	0.2393	760.21
Ours (256^2)	**23.64**	**0.867**	**0.1720**	**0.0143**	**22.53**	**0.871**	**0.1521**	**0.0184**	**25.07**	**0.875**	**0.1488**	**631.75**	**24.48**	**0.860**	**0.1248**	**198.56**

Table 2
Efficiency comparison. Inference time (milliseconds) and peak GPU memory (MB) usage at different output resolutions evaluated on an NVIDIA RTX 5090. OOM indicates an out-of-memory failure.

Method	256 × 256	512 × 512	1024 × 1024	1536 × 1536	2048 × 2048
MVSplat	47.49 / 1,219	120.21 / 4,346	1145.39 / 21,717	OOM	OOM
DepthSplat	96.56 / 2,605	208.94 / 6,320	OOM	OOM	OOM
RoGSplat	144.85 / 2,684	159.85 / 2,476	617.78 / 9,169	1854.64 / 20,517	OOM
EVA-Gaussian	28.91 / 247	32.40 / 487	93.59 / 1,503	188.50 / 3,197	372.84 / 5564
GPS-Gaussian	13.72 / 64	15.81 / 173	18.35 / 505	29.92 / 1,091	50.23 / 1910
GPS-Gaussian+	18.32 / 67	14.79 / 179	21.89 / 507	35.67 / 1,099	60.55 / 1922
Ours	25.70 / 153	26.40 / 258	29.27 / 683	41.72 / 1,386	55.95 / 2360

full-resolution neural computation; the conventional Gaussian representation and splatting process remain unchanged.

The decoupled encoding process is formulated as follows:

$$\mathbf{F}_s = \mathcal{D}_{sem}\big(\mathcal{E}_{img}(\mathbf{I}_{\downarrow}) \oplus \mathcal{E}_{geo}(\mathbf{G}_{\downarrow})\big), \tag{5}$$

$$\mathbf{F}_g = \mathcal{D}_{tex}\big(\mathbf{F}_{s\uparrow} \oplus \mathbf{I} \oplus \mathbf{G}\big), \tag{6}$$

where $\mathbf{I}$ denotes the rectified input image, and $\mathbf{G} = [\mathbf{D}, \mathbf{D}', \mathbf{R}]$ denotes the geometry input formed by concatenating the estimated depth, reprojected depth, and uncertainty map. $\mathbf{I}_{\downarrow}$ and $\mathbf{G}_{\downarrow}$ denote their downsampled versions. $\mathcal{E}_{img}$ and $\mathcal{E}_{geo}$ refer to the image and geometry encoders, respectively. $\mathbf{F}_s$ is the extracted low-frequency semantic feature. $\mathcal{D}_{sem}$ denotes the low-resolution semantic decoder, while $\mathcal{D}_{tex}$ denotes the lightweight high-resolution refinement module. The computationally intensive encoding and decoding operations are therefore confined to the low-resolution branch, whereas $\mathcal{D}_{tex}$ performs only lightweight convolutional processing at the original resolution. $\mathbf{F}_g$ is the high resolution Gaussian feature. $(\cdot)\uparrow$ indicates the upsampling operation, and $\oplus$ denotes channel-wise concatenation.

Finally, $\mathbf{F}_g$ is used to regress the Gaussian attributes $\{M_s, M_\alpha, M_r\}$ and depth deformation M_δ. The refined depth determines the Gaussian centers M_x through back-projection, while the input color is directly assigned to M_c. For co-visible regions, primitives with $\mathbf{R} < \varepsilon$ are treated as coincident and pruned from V_2 while retaining the full set from V_1.

3.4. Training loss

Our training consists of stereo pretraining followed by joint optimization of geometry and appearance. For image quality supervision, we employ a combination of L1 loss (L_1) and SSIM loss (L_2).

Regarding geometry supervision, we constrain both the initial estimated depth and the final refined depth after deformation. The final refined depth is directly supervised using L1 loss (L_3). For the initial estimation, we supervise it in the disparity domain to handle varying depth ranges effectively:

$$L_4 = \sum_{k=1}^{K} \gamma^{K-k} |\mathbf{d}_{gt} - \mathbf{d}_k|_1, \tag{7}$$

Table 3
Ablation study on the THuman-Sit dataset. We evaluate the contributions of geometry correction, reprojection-derived reliability guidance, and decoupled encoding. Inference time is measured in milliseconds.

Method	PSNR↑	SSIM↑	LPIPS↓	MAE↓	Time↓
w/o Geometry Correction	23.92	0.907	0.1354	0.0079	48.16
w/o Reprojection-Derived Reliability	24.76	0.918	0.1310	0.0080	56.85
w/o Decoupled Encoding	25.54	0.930	0.1198	0.0071	165.80
Full Model	25.33	0.926	0.1235	0.0074	52.29

where L_4 denotes the disparity loss, K is the total number of iterations, and $\mathbf{d}_k$ represents the prediction at the k-th iteration. We set the decay factor $\gamma = 0.9$. Consequently, the total loss is defined as:

$$L_{total} = \lambda_1 L_1 + \lambda_2 L_2 + \lambda_3 L_3 + \lambda_4 L_4. \tag{8}$$

Our optimization follows a two-stage training strategy. In the first stage, only the stereo depth estimation module is trained using L_4. In the second stage, the complete framework is jointly optimized with all four objectives. We set $(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = (0.8, 0.1, 0.1, 1.0)$, maintaining a $1:1$ coefficient ratio between the stereo supervision retained from the first stage and the newly introduced objectives in the second stage. Regarding the hyperparameters, we set the number of depth estimation iterations K to 3, and the empirical threshold ε to 1×10^{-3}. The stereo estimator is first trained for 50K iterations in the first stage, followed by 150K iterations of joint optimization in the second stage.

4. Experiment

4.1. Datasets and metrics

We assess the quantitative performance using PSNR, SSIM [26], and LPIPS [27]. For the experimental datasets, we use THuman [28], THuman-Sit [29], Dual-GS [30], and DNA-Rendering [31]. THuman contains human subjects with diverse poses and appearances, while THuman-Sit further includes challenging human–object interactions.

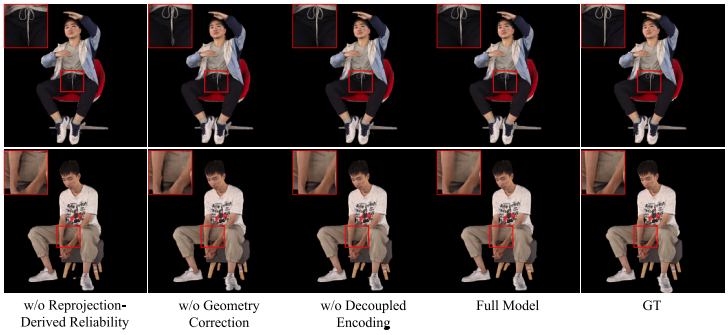


Fig. 3. Ablation Study. Qualitative ablation results on the THuman-Sit dataset.

Dual-GS and DNA-Rendering contain dynamic human sequences with diverse motions and scene configurations.

Beyond PSNR, SSIM, and LPIPS, we adopt additional complementary metrics tailored to the characteristics of each dataset. Since THuman-Sit and THuman consist of static human instances, we report depth MAE to directly evaluate reconstructed geometry. Dual-GS and DNA-Rendering contain dynamic sequences but lack sufficiently dense and accurate depth ground truth for reliable geometric evaluation; we therefore report FVD as a complementary video-level metric. The main quantitative tables retain the conventional frame-averaged evaluation protocol. For statistical analysis, we additionally treat each scene as an independent observation and report scene-averaged results with 95% bootstrap confidence intervals, together with favorable-scene counts and median paired differences. Complete statistical and per-scene results are provided in the supplementary material.

Our method is trained only on the THuman and THuman-Sit datasets, with all evaluation samples excluded from the training set. Dual-GS and DNA-Rendering are used solely for evaluation and are completely excluded from training and fine-tuning. We therefore use these two datasets as held-out benchmarks for evaluating cross-dataset generalization.

4.2. Main results

Baselines. We compare our method with human NVS methods GPS-Gaussian [18], GPS-Gaussian+ [19], EVA-Gaussian [20], and RoGSplat [32], and generalizable novel view synthesis methods MVSplat [6] and DepthSplat [7]. EVA-Gaussian and RoGSplat are retrained from their official public implementations, as pretrained checkpoints were not available. Both are trained only on THuman and THuman-Sit following their official training configurations, while Dual-GS and DNA-Rendering are used only for evaluation. EVA-Gaussian retains its native two-view setting and uses the same 36° source-view separation as our method, whereas RoGSplat retains its native four-view input rather than being modified to two views.

Comparisons. Qualitative and quantitative results are presented in Fig. 2 and Table 1. At 1024×1024, our method achieves the best PSNR, SSIM, and LPIPS on all four datasets among the methods evaluated at this resolution. The improvement is also reflected in the complementary metrics, with the lowest depth MAE on THuman and THuman-Sit and the lowest FVD on Dual-GS and DNA-Rendering. In particular, on DNA-Rendering, RoGSplat achieves 23.22 dB PSNR and an FVD of 324.67, whereas our method achieves 23.99 dB and 188.84, respectively. Similar advantages are observed at the other evaluated resolutions.

The evaluated baselines differ across resolutions because of their computational feasibility and native inference settings. MVSplat and RoGSplat cannot be evaluated at 2048×2048 due to excessive GPU memory consumption. DepthSplat is sensitive to the mismatch between training and inference resolutions and is therefore reported at its native 256×256 resolution. GPS-Gaussian, GPS-Gaussian+, EVA-Gaussian, and our method can be evaluated at both 1024×1024 and 2048×2048.

Table 2 reports the inference time and peak GPU memory usage from 256×256 to 2048×2048. Overall, lower output resolutions generally require less inference time and GPU memory, while the scalability advantage of our method becomes more evident as the output resolution increases. In particular, our method maintains moderate runtime and memory growth at high resolutions, whereas several baselines exhibit substantially increased computational cost or encounter out-of-memory failures.

4.3. Ablation studies

We validate the effectiveness of each component of our method through ablation studies, as summarized in Table 3 and demonstrated in Fig. 3. Removing the geometry correction module leads to a noticeable decline in performance. This occurs because the network loses the capability to mitigate depth estimation errors, resulting in degraded quality in both geometry and textural details.

Furthermore, we evaluate the impact of the decoupled semantic and texture encoding module by replacing it with full-resolution computation. As shown in Table 3, full-resolution processing slightly improves the rendering-quality metrics, but increases the inference time from 52.29 ms to 165.80 ms. The proposed decoupled design therefore represents a deliberate efficiency-quality trade-off, substantially reducing the computational cost while maintaining only a small difference in rendering quality.

To further isolate the contribution of explicit reliability guidance, we retain the complete geometry-correction network while setting only the reprojection-derived reliability input to zero. This variant reduces PSNR from 25.33 dB to 24.76 dB and increases the target-view depth MAE from 0.0074 to 0.0080, indicating that explicitly exposing reprojection-derived reliability helps the correction network identify regions requiring geometric refinement.

5. Limitations

Our method currently assumes accurately calibrated and rectified stereo inputs, as both disparity-based depth estimation and bidirectional depth reprojection depend on reliable camera parameters. The framework is designed for two input views and does not explicitly exploit additional observations when more views are available. Moreover, very large viewpoint differences and severe occlusions can make stereo correspondence and cross-view consistency estimation less reliable, thereby reducing the effectiveness of subsequent geometry correction. Finally, our high-resolution runtime is evaluated on an NVIDIA RTX 5090 GPU; therefore, the reported efficiency may not directly transfer to lower-end hardware. Extending the framework to more flexible multi-view inputs and improving robustness under challenging camera configurations are promising directions for future work.

6. Conclusion

We propose a fast and generalizable framework for human novel-view synthesis from image pairs. Uncertainty-guided geometry correction improves geometric and rendering fidelity, while decoupled semantic and texture encoding enables efficient high-resolution inference. Extensive evaluations demonstrate strong reconstruction quality and generalization across multiple human datasets, highlighting the practical potential of the proposed framework.

CRediT authorship contribution statement

Kaijin Zhao: Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Conceptualization; **Bingjie Song:** Writing – review & editing, Validation, Software; **Chen Gong:** Investigation, Data curation; **Guoqing Zhou:** Writing – review & editing, Supervision, Resources; **Qing Wang:** Writing – review & editing, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by the NSFC under grant 62571449.

Data availability

The authors do not have permission to share data.

Supplementary material

Supplementary material associated with this article can be found in the online version at 10.1016/j.patrec.2026.09.036.

References

[1] B. Mildenhall, P.P. Srinivasan, M. Tancik, J.T. Barron, R. Ramamoorthi, R. Ng, NeRF: representing scenes as neural radiance fields for view synthesis, in: European Conference on Computer Vision, 2020.
[2] H. Xu, A. Chen, Y. Chen, C. Sakaridis, Y. Zhang, M. Pollefeys, A. Geiger, F. Yu, MuRF: multi-baseline radiance fields, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024.
[3] Q. Wang, Z. Wang, K. Genova, P. Srinivasan, H. Zhou, J.T. Barron, R. Martin-Brualla, N. Snavely, T. Funkhouser, IBRNet: learning multi-view image-based rendering, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.
[4] D. Zimny, J. Waczyńska, T. Trzciński, P. Spurek, Points2NeRF: generating neural radiance fields from 3D point cloud, Pattern Recognit. Lett. 185 (2024) 8–14.
[5] B. Kerbl, G. Kopanas, T. Leimkühler, G. Drettakis, 3D Gaussian splatting for real-time radiance field rendering, ACM Trans. Graph. 42 (4) (2023).
[6] Y. Chen, H. Xu, C. Zheng, B. Zhuang, M. Pollefeys, A. Geiger, T.-J. Cham, J. Cai, MVSplat: efficient 3D Gaussian splatting from sparse multi-view images, in: European Conference on Computer Vision, Springer, 2024, pp. 370–386.
[7] H. Xu, S. Peng, F. Wang, H. Blum, D. Barath, A. Geiger, M. Pollefeys, Depthsplat: connecting Gaussian splatting and depth, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2025, pp. 16453–16463.
[8] C. Zhang, Y. Zou, Z. Li, M. Yi, H. Wang, Transplat: generalizable 3D gaussian splatting from sparse multi-view images with transformers, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, 2025, pp. 9869–9877.
[9] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, M.J. Black, SMPL: a skinned multi-person linear model, ACM Trans. Graph. (Proc. SIGGRAPH Asia) 34 (6) (2015) 248:1–248:16.
[10] K. Park, U. Sinha, J.T. Barron, S. Bouaziz, D.B. Goldman, S.M. Seitz, R. Martin-Brualla, Nerfies: deformable neural radiance fields, IEEE/CVF Int. Conf. Comput. Vis. (2020) 5845–5854.
[11] L. Han, X. Zhang, H. Song, K. Shi, Y.-S. Liu, Z. Han, SparseRecon: neural implicit surface reconstruction from sparse views with feature and depth consistencies, in: 2025 IEEE/CVF International Conference on Computer Vision (ICCV), 2025, pp. 28514–28524.
[12] Z. Tan, X. Chen, J. Zhang, L. Feng, D. Hu, Uncertainty-aware normal-guided Gaussian splatting for surface reconstruction from sparse image sequences, (2025). arXiv preprint arXiv:2503.11172
[13] F. Hu, B. Liu, L. Fang, L. Ye, Q. Zhang, CoSplat: resolving multi-view inconsistency in feed-forward 3D Gaussian splatting, Displays 95 (2026) 103549.
[14] S. Tang, W. Ye, P. Ye, W. Lin, Y. Zhou, T. Chen, W. Ouyang, Hisplat: hierarchical 3d gaussian splatting for generalizable sparse-view reconstruction, in: International Conference on Learning Representations, 2025, 2025, pp. 48436–48449.
[15] C. Wewer, K. Raj, E. Ilg, B. Schiele, J.E. Lenssen, LatentSplat: autoencoding variational Gaussians for fast generalizable 3D reconstruction, in: A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, G. Varol (Eds.), Computer Vision – ECCV 2024, Springer Nature Switzerland, Cham, 2025, pp. 456–473.
[16] P. Pan, Z. Su, C. Lin, Z. Fan, Y. Zhang, Z. Li, T. Shen, Y. Mu, Y. Liu, Humansplat: generalizable single-image human Gaussian splatting with structure priors, Adv. Neural Inf. Process. Syst. 37 (2024) 74383–74410.
[17] Y. Kwon, B. Fang, Y. Lu, H. Dong, C. Zhang, F.V. Carrasco, A. Mosella-Montoro, J. Xu, S. Takagi, D. Kim, et al., Generalizable human Gaussians for sparse view synthesis, in: European Conference on Computer Vision, Springer, 2024, pp. 451–468.
[18] S. Zheng, B. Zhou, R. Shao, B. Liu, S. Zhang, L. Nie, Y. Liu, Gps-gaussian: generalizable pixel-wise 3D Gaussian splatting for real-time human novel view synthesis, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 19680–19690.
[19] B. Zhou, S. Zheng, H. Tu, R. Shao, B. Liu, S. Zhang, L. Nie, Y. Liu, Gps-Gaussian+: generalizable pixel-wise 3D Gaussian splatting for real-time human-scene rendering from sparse views, IEEE Trans. Pattern Anal. Mach. Intell. (2025).
[20] Y. Hu, Z. Liu, J. Shao, Z. Lin, J. Zhang, EVA-Gaussian: 3D Gaussian-based real-time human novel view synthesis under diverse multi-view camera settings, in: 2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), 2025, pp. 2634–2643.
[21] C.-Y. Weng, B. Curless, P.P. Srinivasan, J.T. Barron, I. Kemelmacher-Shlizerman, Humannerf: free-viewpoint rendering of moving people from monocular video, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16210–16220.
[22] Z. Xu, S. Peng, H. Lin, G. He, J. Sun, Y. Shen, H. Bao, X. Zhou, 4K4d: real-time 4d view synthesis at 4k resolution, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 20029–20040.
[23] Z. Zhai, G. Chen, W. Wang, D. Zheng, J. Xiao, TAGA: self-supervised learning for template-free animatable Gaussian articulated model, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 21159–21169.
[24] L. Lipson, Z. Teed, J. Deng, RAFT-stereo: multilevel recurrent field transforms for stereo matching, in: 2021 International Conference on 3D Vision (3DV), 2021, pp. 218–227.
[25] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al., DINOv2: learning robust visual features without supervision, Trans. Mach. Learn. Res. J. (2024) 1–31.
[26] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.
[27] R. Zhang, P. Isola, A.A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (2018) 586–595.
[28] T. Yu, Z. Zheng, K. Guo, P. Liu, Q. Dai, Y. Liu, Function4d: real-time human volumetric capture from very sparse consumer rgbd sensors, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 5746–5756.
[29] J. Zhang, Y. Zhang, H. Zhang, X. Zhou, B. Zhou, R. Shao, Z. Hu, Y. Liu, Ins-HOI: instance aware human-object interactions recovery, IEEE Trans. Pattern Anal. Mach. Intell. (2025).
[30] Y. Jiang, Z. Shen, Y. Hong, C. Guo, Y. Wu, Y. Zhang, J. Yu, L. Xu, Robust dual Gaussian splatting for immersive human-centric volumetric videos, ACM Trans. Graph. 43 (6) (2024) 1–15.
[31] W. Cheng, R. Chen, S. Fan, W. Yin, K. Chen, Z. Cai, J. Wang, Y. Gao, Z. Yu, Z. Lin, D. Ren, L. Yang, Z. Liu, C.C. Loy, C. Qian, W. Wu, D. Lin, B. Dai, K.-Y. Lin, DNA-rendering: a diverse neural actor repository for high-fidelity human-centric rendering, in: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE Computer Society, Los Alamitos, CA, USA, 2023, pp. 19925–19936.
[32] J. Xiao, Q. Zhang, Y. Nie, L. Zhu, W.-S. Zheng, RoGSplat: learning robust generalizable human Gaussian splatting from sparse multi-view images, in: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025, pp. 5980–5990.