STEREOGAUSSIANS: FEED-FORWARD 3D GAUSSIAN SPLATTING FROM STEREO IMAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Feed-forward 3D Gaussian Splatting (3DGS) enables reconstruction without per-scene optimisation, but practical stereo-camera applications require nearby-view extrapolation beyond the input views. Stereo depth anchors visible surfaces, yet rendering newly exposed regions also requires learned appearance and additional scene capacity. We introduce StereoGaussians, which predicts a metric 3DGS representation from a single calibrated stereo pair. It reuses intermediate representations from frozen pretrained stereo networks to predict Gaussian attributes, while calibrated disparity anchors the geometry. A second Gaussian layer and an expanded image canvas provide capacity for disoccluded and outside-field-of-view content. For training, we construct SceneSplat-Stereo from quality-filtered 3DGS teachers, pairing stereo inputs with nearby target views across 803 training scenes. Experiments on unseen real and photorealistic stereo benchmarks demonstrate improvements over strong view-synthesis baselines, while ablation studies support our main design choices.

1 INTRODUCTION

Feed-forward 3D Gaussian Splatting (3DGS) predicts renderable scenes without per-scene optimisation (Charatan et al., 2024; Chen et al., 2024; Xu et al., 2025b). For XR headsets, stereo rigs, and spatial video, a single calibrated stereo pair must support nearby-view extrapolation, revealing surfaces hidden from the inputs or extending beyond their field of view. Unlike interpolation between posed context views, this requires predicting content outside the observed coverage. Stereo calibration provides metric geometry for visible surfaces, but accurate matching alone cannot determine how newly exposed regions render. Monocular predictors infer geometry and scale from image priors; calibrated stereo instead supplies a metric reference for learning a renderable representation.

Our central idea is to transfer pretrained stereo representations beyond their final disparity output. Stereo networks encode both correspondence and image context (Wen et al., 2025; 2026), providing features from which to learn Gaussian appearance, opacity, and shape. StereoGaussians couples these frozen features with a trainable predictor: calibrated depth anchors the geometry, while rendering supervision learns the attributes needed for novel views. A second Gaussian layer adds capacity behind visible surfaces, and an expanded prediction canvas adds support outside the input frustum. Together, these mechanisms connect stereo geometry with learned scene content for nearby-view extrapolation. The encoder remains frozen, allowing rendering supervision to train the Gaussian predictor while preserving the pretrained stereo mapping.

Learning this mapping requires stereo inputs paired with posed target views of the same static scene. Existing correspondence datasets often lack this supervision, while synthetic scenes and dynamic captures can introduce appearance gaps or scene inconsistencies. We address this training-data gap with SceneSplat-Stereo, rendering calibrated stereo pairs and nearby targets from quality-filtered real-scene 3DGS teachers. This combines controlled camera sampling, static-scene consistency, and reconstructed real-scene appearance. Target displacements beyond the stereo baseline expose regions that correspondence supervision alone does not cover.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Our contributions are:

1. **Stereo-to-target training data at scale.** SceneSplat-Stereo supplies approximately 324 K calibrated stereo pairs and 3.24 M target images across 803 training scenes. Its teacher-based construction enables consistent rendering supervision, supporting stronger cross-dataset generalisation and perceptual quality than the evaluated existing stereo training datasets on the two novel-view synthesis benchmarks.
2. **Stereo representation transfer for Gaussian prediction.** StereoGaussians maps a single stereo pair to a reusable metric 3DGS representation, combining frozen stereo features and depth-anchored attributes with two-layer, expanded-canvas prediction for disoccluded regions and scene content beyond the reference field of view.
3. **State-of-the-art zero-shot view synthesis.** StereoGaussians achieves state-of-the-art performance among the evaluated methods on real and photorealistic stereo benchmarks. Ablations establish the contributions of pretrained stereo features, two-layer geometry, and expanded prediction coverage to rendering quality.

2 Related Work

Feed-forward Gaussian prediction. pixelSplat (Charatan et al., 2024), MVSplat (Chen et al., 2024), MVSGaussian (Liu et al., 2024), and DepthSplat (Xu et al., 2025b) predict Gaussians from posed views. DepthSplat combines multi-view matching with pretrained monocular depth features to connect depth estimation and Gaussian prediction. GS-LRM (Zhang et al., 2024) predicts Gaussians with transformers, while Long-LRM (Chen et al., 2025) scales to longer image sequences and wider scene coverage. NoPoSplat (Ye et al., 2025b), Splatt3R (Smart et al., 2024), and AnySplat (Jiang et al., 2025) relax camera requirements; YoNoSplat (Ye et al., 2025a) supports both posed and unposed inputs. ReSplat (Xu et al., 2025a) refines Gaussians through rendering feedback. Single-image methods such as Splatter Image (Szymanowicz et al., 2024), Flash3D (Szymanowicz et al., 2025), and SHARP (Mescheder et al., 2025) infer geometry and appearance from monocular priors, including additional scene capacity for surfaces not directly visible in the input image.

Transformer-based view synthesis. LVSM (Jin et al., 2025) synthesises target images using learned latent tokens or direct input-to-target attention. Efficient-LVSM (Jia et al., 2026) separates input and target processing through decoupled co-refinement attention. These approaches synthesise views without predicting an explicit Gaussian scene.

Stereo representations and view synthesis. RAFT-Stereo (Lipson et al., 2021), IGEV (Xu et al., 2023), and FoundationStereo (Wen et al., 2025) learn correspondence and context features for disparity estimation; Fast-FoundationStereo (Wen et al., 2026) studies efficient zero-shot stereo. StereoNeRF (Lee et al., 2024) combines pretrained stereo features with depth-guided plane sweeping and a stereo depth loss for NeRF rendering from multiple pairs. Stereo-GS (Huang et al., 2025) uses global attention to fuse pairwise stereo features and separate geometry and appearance heads for pose-free multi-view Gaussian reconstruction. WorldMirror (Liu et al., 2026) incorporates camera and depth priors into joint geometry and Gaussian prediction. In our approach, calibrated disparity anchors metric geometry, while frozen stereo features condition Gaussian attributes. Two-layer prediction and an expanded canvas support nearby-view extrapolation from a single stereo pair.

Stereo and view-synthesis data. RealEstate10K (Zhou et al., 2018) and DL3DV (Ling et al., 2024) provide posed video for view synthesis. Middlebury (Scharstein et al., 2014) and DrivingStereo (Yang et al., 2019) benchmark stereo matching, while Flickr1024 (Wang et al., 2019b) supports stereo super-resolution. SceneFlow (Mayer et al., 2016) and IRS (Wang et al., 2019a) provide synthetic geometric supervision, but correspondence data need not include posed target views for rendering supervision. StereoNVS (Lee et al., 2024) provides stereo-to-target views, and SceneSplat-7K (Li et al., 2025) provides real-scene Gaussians. SceneSplat-Stereo builds on these teachers to jointly control stereo calibration and target-view sampling, producing appearance-consistent supervision for nearby-view extrapolation.

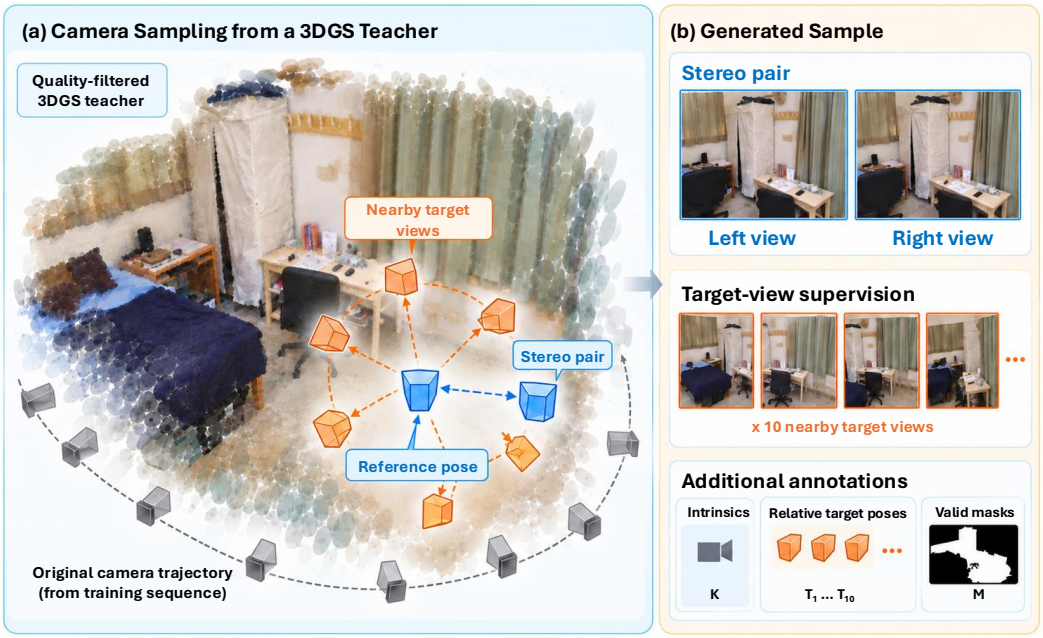

Figure 1: **SceneSplat-Stereo synthetic data generation. (a) Camera sampling.** A quality-filtered real-scene 3DGS teacher renders a calibrated stereo pair and ten nearby targets around a reference pose from the source trajectory. **(b) Generated sample.** Stereo inputs and target images are accompanied by intrinsics, relative target poses, and opacity-derived valid masks. Camera placement is schematic and illustrates the relative arrangement of input and target views.

3 SceneSplat-Stereo Dataset

SceneSplat-Stereo pairs calibrated stereo inputs with nearby target views rendered from the same static 3DGS teacher. This provides consistent appearance and known camera poses for learning stereo-to-3DGS prediction. Figure 1 illustrates the camera sampling and rendering pipeline, which preserves the reconstructed teacher's real-scene appearance. The input pair provides a metric geometric reference, while surrounding targets supervise rendering as surfaces become visible and image coverage changes. Table 1 compares its combination of stereo input, target-view supervision, static-scene consistency, and real-scene appearance with existing datasets used for stereo matching and novel-view synthesis under different supervision settings.

Sources and splits. We use the ScanNetGS and ScanNetPPGS-V2 subsets of SceneSplat-7K (Li et al., 2025), selecting them for their scene scale and capture coverage for nearby rendering. The quality-filtered pool is split into 803 training and 50 validation scenes, with no scene overlap. Training contains approximately 324 K stereo pairs, each with ten separately rendered targets (3.24 M target images). For the default model, StereoNVS and ReplicaGS serve only as test datasets. Source counts and validation sampling are detailed in Appendix B and Table 5.

Rendering and quality control. For each reference pose, we translate the right camera by baseline B along the local x-axis, retaining its orientation and intrinsics to form a rectified pair. Ten target cameras translate by radius ρ along fixed Fibonacci-sphere directions, with no rotation or overlap-based rejection. Samples include calibration, relative target poses, and opacity-derived valid masks, used for ReplicaGS evaluation but not training.

We retain teachers with released reconstruction PSNR above 30 dB at 30K optimisation steps, selecting scenes by reconstruction fidelity. Three quality tiers set B and ρ, with target displacement two to three times the stereo baseline. Training samples four of the ten targets per pair, exposing the predictor to varied nearby viewpoints across iterations. The exact tiers, filtering counts, and rendering protocol are given in Appendix B.

Table 1: Dataset comparison for feed-forward stereo-to-3DGS prediction. Scale reports scenes (or sequences) / input images, counting both stereo views or individual monocular frames; separately rendered targets are excluded. See Appendix B for details on dataset-specific counting conventions and the supervision available from each source.

Data source	Scenes / images	Stereo input	Target-view sup.	Static scenes	Real appearance
RealEstate10K (Zhou et al., 2018)	80K / 10M		✓	✓	✓
DL3DV-10K (Ling et al., 2024)	10.5K / 51.2M		✓	✓	✓
Middlebury v3 (Scharstein et al., 2014)	30 / 60	✓		✓	✓
Flickr1024 (Wang et al., 2019b)	1K / 2K	✓		✓	✓
InStereo2K (Bao et al., 2020)	– / 4.1K	✓		✓	✓
KITTI (Geiger et al., 2012)	789 / 1.6K	✓			✓
DrivingStereo (Yang et al., 2019)	42 / 364K	✓			✓
SceneFlow (Mayer et al., 2016)	2.2K / 79.6K	✓	✓		
IRS (Wang et al., 2019a)	70 / 207K	✓		✓	
StereoNVS (Lee et al., 2024)	103 / 17.6K	✓	✓	✓	✓
SceneSplat-Stereo	**803 / 648K**	✓	✓	✓	✓

4 StereoGaussians

4.1 Overview

Given a calibrated rectified stereo pair (I_L, I_R), intrinsics K, and baseline B, StereoGaussians predicts Gaussians $\mathcal{G} = \{(\mathbf{x}_i, \mathbf{s}_i, \mathbf{q}_i, \alpha_i, \mathbf{c}_i)\}_{i=1}^{N}$ in the left-camera frame, using the standard 3DGS position, scale, rotation, opacity, and colour attributes (Kerbl et al., 2023). A frozen stereo encoder supplies intermediate features for attribute prediction and disparity for metric geometry. A trainable predictor constructs two Gaussian layers on an expanded image canvas, providing capacity for disoccluded and outside-FoV content. The decoder and predictor learn through differentiable Gaussian rendering with target-image supervision from SceneSplat-Stereo; the stereo encoder remains frozen. Figure 2 shows the pipeline.

4.2 Transferring stereo representations to 3DGS

The encoder exposes a left-reference disparity map d and aligned multiscale features $\{F^s\}$ through a backbone-specific interface. A DPT-style decoder $\mathcal{D}$ concatenates each feature scale with resized reference RGB and stereo depth z (Eq. 3), then projects and fuses the features coarse-to-fine into a dense full-resolution descriptor for downstream Gaussian attribute prediction:

$$\Phi = \mathcal{D}\big(\{[F^s, \downarrow_s I_L, \downarrow_s z]\}_s\big) \in \mathbb{R}^{C\times H\times W}. \tag{1}$$

Here $[\cdot,\cdot]$ denotes channel concatenation and $\downarrow_s$ resizes inputs to feature scale s. After canvas extension (§4.4), a convolutional predictor $\mathcal{H}$ maps the extended inputs, denoted by tildes, to attribute residuals for both layers and a separate back-layer depth offset:

$$(\theta_{\text{front}}, \theta_{\text{back}}, \Delta z) = \mathcal{H}([\tilde{I}_L, \tilde{z}, \tilde{\Phi}]). \tag{2}$$

Its shared regressor feeds three output branches. This feature pathway conditions Gaussian prediction on pretrained stereo representations beyond the final disparity map. Encoder interfaces and decoder/head implementation details are given in Appendix A.

4.3 Geometry-anchored Gaussian prediction

For reference pixel $p = (u, v)$, calibrated disparity defines metric depth

$$z(p) = \text{clip}\left(\frac{f_x B}{d(p)}, z_{\min}, z_{\max}\right), \tag{3}$$

where f_x is the horizontal focal length in pixels. The front layer retains this processed depth. Each layer ℓ predicts a two-channel output o_ℓ giving image-plane offsets $(\delta_{u,\ell}, \delta_{v,\ell}) = \sigma(o_\ell) - \frac{1}{2}$, bounded to half a canvas pixel per axis. Its centre is

$$\mathbf{x}_\ell(p) = z_\ell(p)\tilde{K}^{-1}[u + \delta_{u,\ell}, v + \delta_{v,\ell}, 1]^\top, \tag{4}$$

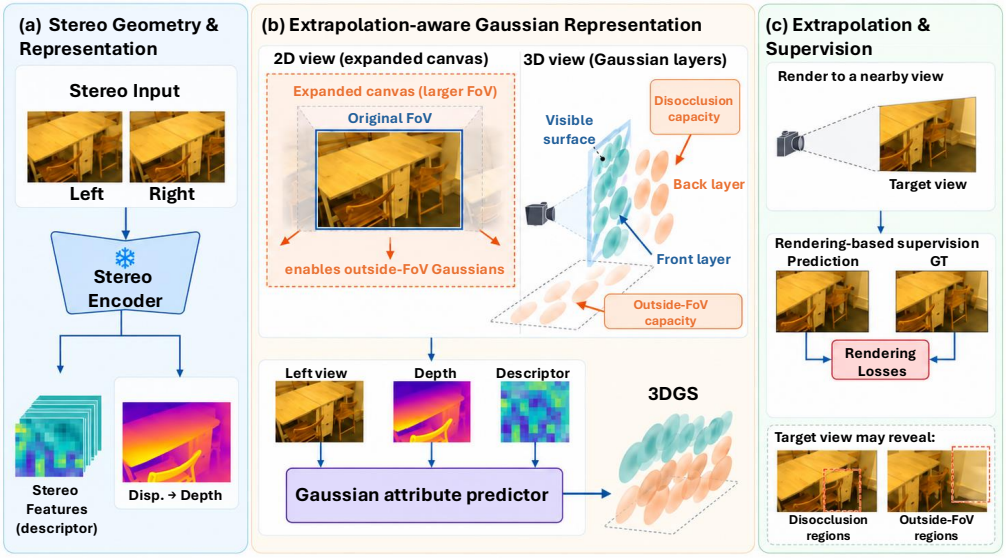

Figure 2: **StereoGaussians overview. (a) Stereo geometry and representation.** Frozen stereo features and calibrated depth condition Gaussian prediction. **(b) Extrapolation-aware representation.** Two layers and an expanded canvas support disoccluded and outside-FoV content. **(c) Extrapolation and supervision.** Target-view rendering trains the decoder and attribute predictor. Scene and Gaussian illustrations are schematic.

where z_ℓ and canvas intrinsics $\tilde{K}$ are specified below.

We adapt SHARP's residual attribute composition (Mescheder et al., 2025): geometry and reference colour define base attributes, while the predictor learns bounded multiplicative scale updates, additive quaternion updates, and opacity updates in logit space. The isotropic base scale $s_{0,\ell} = \kappa z_\ell / \sqrt{f_x f_y}$ follows the metric pixel footprint. Our adaptation fixes front-layer stereo depth and bounds image-plane offsets, whereas SHARP also refines inverse depth through its composer. A separate branch determines back-layer depth, with no subsequent depth residual. We also replace RGB composition with degree-2 spherical-harmonic (SH) residuals around a reference-image DC base. These constraints provide a structural prior without an additional supervision loss. Full attribute equations, activations, and constants are given in Appendix A.2.

4.4 Extrapolation-aware prediction

Back layer. Both layers have separate attribute outputs at every canvas pixel. The front layer uses $z_{\text{front}} = \tilde{z}$; the back layer uses

$$z_{\text{back}}(p) = \max_{q \in \mathcal{N}_9(p)} \tilde{z}(q) + \text{softplus}(\Delta z(p)), \tag{5}$$

where $\mathcal{N}_9(p)$ is a 9×9 neighbourhood. The local maximum provides a deeper anchor near depth boundaries, and the learned positive offset adjusts its depth under rendering supervision. Both the centre and base footprint use this final depth. We concatenate the two layers; learned attributes and visibility determine their contributions to a target view, without an edge mask restricting the back layer to a predefined subset of image pixels.

Expanded field of view. After decoding at the original resolution, we extend I_L, z, and Φ by edge replication to $(H + 2P) \times (W + 2P)$, where P is the padding per side. Canvas intrinsics retain the focal lengths and shift the principal point to $(c_x + P, c_y + P)$, so padded pixels define rays outside the original image frustum. The predictor operates on the full canvas, using the extended depth for both layer constructions above. Target cameras retain their original resolution and intrinsics; padded Gaussians receive the same rendering supervision through the target pixels they cover. Canvas implementation details are given in Appendix A.1.

Table 2: Zero-shot comparison using the default 16-pixel padding. Our model is trained on SceneSplat-Stereo. Baselines retain their respective pretraining. Bold indicates the best score in each column for the corresponding dataset and evaluation metric.

	StereoNVS			ReplicaGS		
Method	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Single-view feed-forward 3DGS						
SHARP-mono	18.845	0.6140	0.2604	23.388	0.7877	0.2103
SHARP-stereo	18.767	0.6030	0.2788	23.363	0.7808	0.2164
Sparse-view feed-forward 3DGS						
MVSplat	17.458	0.7248	0.2087	14.705	0.5342	0.4152
DepthSplat	17.977	0.7498	0.2024	16.388	0.6161	0.3270
Transformer-based NVS						
Efficient-LVSM	15.319	0.4646	0.4752	22.780	0.7539	0.2210
LVSM	19.796	0.6450	0.2460	25.772	0.8409	0.1701
Stereo geometry without learned prediction						
FS-forward-splat	15.734	0.6583	0.2891	17.017	0.7846	0.3009
FS-naive-GS	16.008	0.6636	0.2802	17.918	0.8090	0.2303
StereoGaussians	**21.261**	**0.7549**	**0.1653**	**30.266**	**0.9447**	**0.0818**

5 Experiments

5.1 Experimental setup

Training and evaluation. Our default uses the FoundationStereo encoder and predictor of §4, with two Gaussian layers, no back-layer edge mask, and 16-pixel FoV padding. Our architectural ablations train at 256×256 for 30K steps with batch size 32, using full-image MSE plus 0.05 LPIPS and selecting the checkpoint with highest validation PSNR. SceneSplat-Stereo pairs are sampled uniformly with four random targets per pair. Full training and validation details are in Appendices A and B, including optimisation, model selection, and target sampling.

We evaluate zero-shot on StereoNVS (8 real-world scenes, 224 pairs) and ReplicaGS (8 held-out photorealistic scenes, 800 pairs, 8 000 targets). Neither test set is used to train or validate the default SceneSplat-Stereo model. We report PSNR, SSIM, and LPIPS-VGG; ReplicaGS metrics use opacity-derived valid masks for teacher-supported regions (Appendix C).

Baselines. We compare MVSplat (Chen et al., 2024), DepthSplat (Xu et al., 2025b), SHARP (Mescheder et al., 2025), LVSM (Jin et al., 2025), and Efficient-LVSM (Jia et al., 2026). SHARP-mono uses the reference image; SHARP-stereo concatenates independent left/right predictions in the reference frame. FS-forward-splat and FS-naive-GS use FoundationStereo depth with forward splatting and hand-specified Gaussian attributes, respectively. We evaluate released models with their original pretraining under the zero-shot transfer protocol. Baseline settings are detailed in Appendix D, including pretrained model selection and LVSM camera preprocessing.

5.2 Zero-shot novel-view synthesis

StereoGaussians outperforms all compared baselines on both test sets across PSNR, SSIM, and LPIPS (Table 2). Relative to LVSM, the strongest learned baseline in PSNR, it gains 1.465 dB on StereoNVS and 4.494 dB on ReplicaGS. Both FS baselines lag behind despite using the same stereo backbone, indicating that strong geometry alone is insufficient without learned Gaussian prediction. Qualitative comparisons appear in Figure 3 and Appendix G.1.

Higher-resolution models. The same architecture supports 512^2 and 1024^2 stereo inputs and target-view rendering (Appendix E).

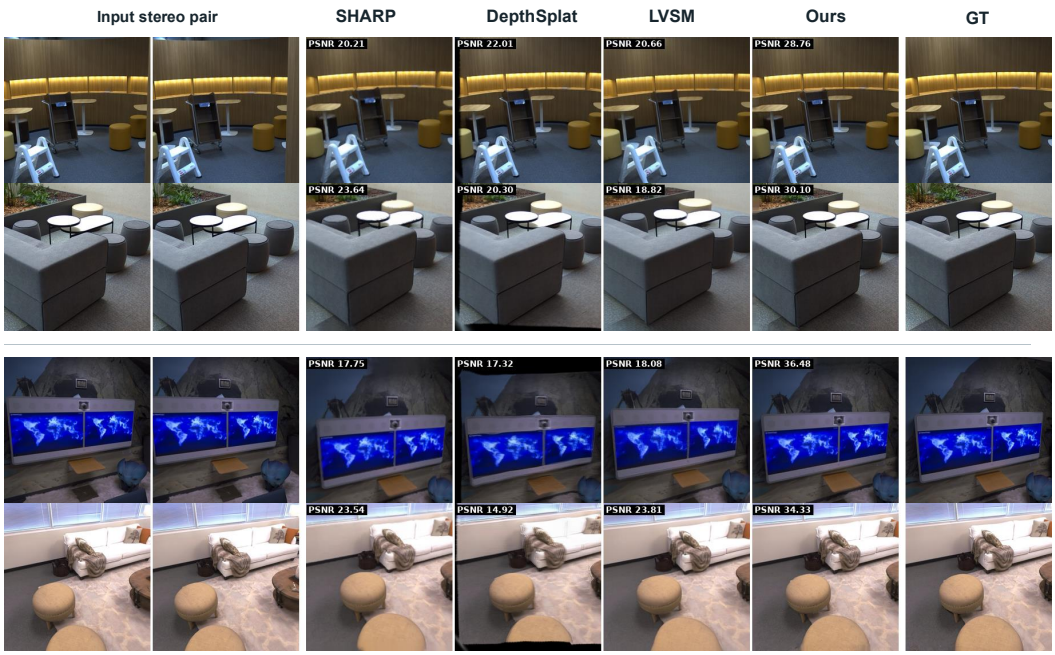

Figure 3: **Qualitative novel-view comparison** on StereoNVS (top two rows) and ReplicaGS (bottom two rows). Each row shows the input stereo pair, predictions from representative methods SHARP, DepthSplat, and LVSM, StereoGaussians (Ours), and ground truth (GT). Prediction panels report PSNR in dB.

Table 3: StereoGaussians trained on different datasets. Columns group test datasets. StereoNVS-to-StereoNVS is in-domain; all others test zero-shot cross-dataset transfer.

	StereoNVS			ReplicaGS		
Training dataset	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
StereoNVS	**22.707**	**0.7585**	0.2024	23.929	0.8914	0.1935
IRS	19.656	0.6546	0.2412	24.614	0.8421	0.1873
SceneSplat-Stereo	21.261	0.7549	**0.1653**	**30.266**	**0.9447**	**0.0818**

5.3 Contribution of training data

We train StereoGaussians on three datasets (Table 3). StereoNVS uses its training split; IRS preparation is in Appendix F. On ReplicaGS, SceneSplat-Stereo improves PSNR by 6.337 dB over StereoNVS training and 5.652 dB over IRS training, with better SSIM and LPIPS. Without StereoNVS training images, SceneSplat-Stereo achieves lower StereoNVS LPIPS than in-domain training (0.1653 versus 0.2024), which attains higher PSNR/SSIM (22.707/0.7585 versus 21.261/0.7549). SceneSplat-Stereo also outperforms IRS on all three StereoNVS metrics, including a 1.605 dB PSNR gain. These results demonstrate its effectiveness for cross-dataset generalisation and perceptual quality in novel-view rendering on both evaluation benchmarks.

5.4 Transferring frozen stereo representations

Table 4 and Appendix G.2 compare variants trained separately with frozen encoders under the common schedule; unmentioned settings remain unchanged.

Feature pathway and RGB-D conditioning. Removing the DPT descriptor retains RGB and stereo geometry but reduces PSNR by 0.509 dB on StereoNVS and 1.982 dB on ReplicaGS, with worse SSIM and LPIPS on both (block A). The highlighted tabletop distortions and foreground-

Table 4: Ablations of input conditioning, frozen stereo encoders, Gaussian prediction, and FoV expansion under a shared training and zero-shot evaluation protocol. Each block varies the default in §5.1; variant definitions are given in §5.4–5.5.

Variant	StereoNVS PSNR ↑	StereoNVS SSIM ↑	StereoNVS LPIPS ↓	ReplicaGS PSNR ↑	ReplicaGS SSIM ↑	ReplicaGS LPIPS ↓
Shared default	21.261	0.7549	0.1653	**30.266**	0.9447	0.0818
(A) Input conditioning						
Without descriptor	20.752	0.7417	0.1871	28.284	0.9362	0.1045
Without RGB-D conditioning	21.269	0.7594	0.1717	29.953	0.9391	0.0915
(B) Frozen stereo encoder						
Fast-FoundationStereo	20.876	0.7409	0.1737	29.404	0.9368	0.0876
(C) Gaussian prediction design						
Without back layer	21.081	0.7506	0.1698	26.440	0.9255	0.0988
Hard edge mask	21.351	0.7573	0.1643	27.421	0.9314	0.0951
Soft edge mask	21.125	0.7527	0.1662	30.256	**0.9452**	**0.0816**
Without composer	21.875	0.7510	0.1887	30.029	0.9445	0.0845
(D) FoV padding per side						
without padding	18.093	0.7238	0.1956	21.534	0.8884	0.1393
32 pixels	22.400	0.7600	0.1639	30.188	0.9445	0.0823
64 pixels	**23.189**	**0.7689**	**0.1612**	30.165	0.9443	0.0823

boundary artifacts in Figure 4 also illustrate the feature pathway's contribution to Gaussian prediction. RGB-D conditioning further improves LPIPS on both benchmarks and ReplicaGS PSNR by 0.313 dB, with little change in StereoNVS PSNR. This ablation removes RGB-D inputs from the decoder and downstream predictor while retaining stereo features and metric depth anchors.

Different frozen encoders. Our framework also transfers pretrained representations from Fast-FoundationStereo (Wen et al., 2026). With an adapted feature interface and a separately trained predictor, this variant reaches 20.876/29.404 dB on StereoNVS/ReplicaGS (block B), exceeding the evaluated external baselines in PSNR and LPIPS on both sets. These results demonstrate effective transfer from both frozen stereo encoders; implementation details are in Appendix A.

5.5 Gaussian prediction design

Second-layer capacity. The single-layer variant retains the front layer anchored at stereo depth and removes the additional back-layer Gaussians. This reduces PSNR by 0.180 dB on StereoNVS and 3.826 dB on ReplicaGS (Table 4, block C). The highlighted desk/chair regions in Figure 4 show missing content and disrupted occlusion boundaries without the back layer; the full model renders these regions more completely.

Hard and soft edge masks. We test back-layer opacity masks: hard masks zero opacity outside depth-edge regions, while soft masks attenuate it (Appendix A.2). Qualitatively, both promote clearer front/back separation and back-layer specialisation in occluded content. Rendering gains vary across datasets (Table 4, block C), so we use the simpler ungated design as the default.

Attribute composition. The geometry- and colour-based attribute priors in §4.3 improve LPIPS on both datasets and all three ReplicaGS metrics. On StereoNVS, composition reduces LPIPS from 0.1887 to 0.1653 and improves SSIM; the composer-free variant has 0.614 dB higher PSNR. On ReplicaGS, composition also increases PSNR by 0.237 dB. We retain the composer for its consistent improvements in perceptual quality across both evaluation benchmarks.

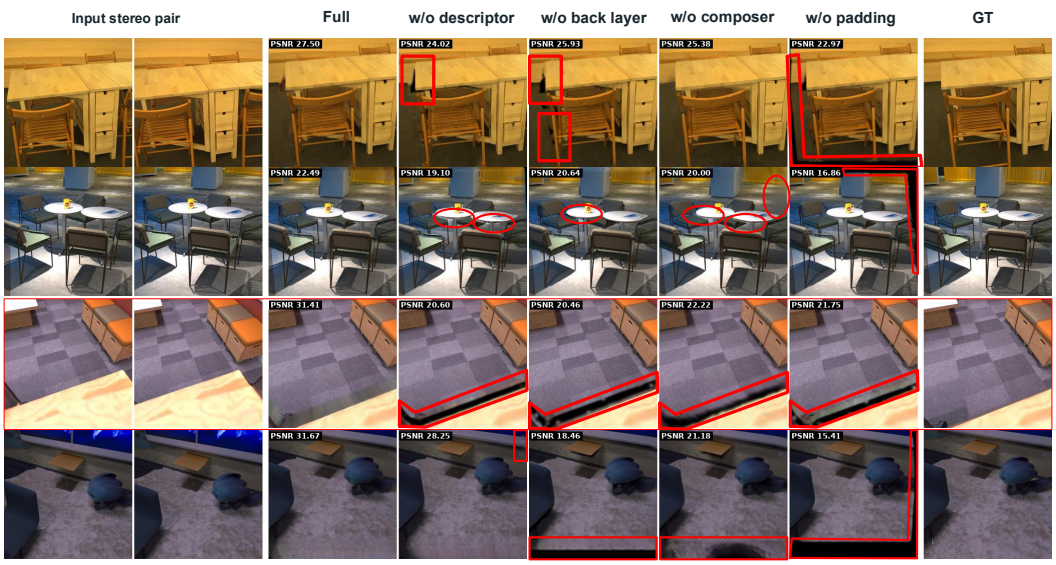

Figure 4: **Qualitative ablations.** Each row shows the input stereo pair, the full model, variants without the feature descriptor, back layer, composer, or FoV padding, and ground truth (GT). Red outlines highlight missing content and distortions near occlusion boundaries and in newly exposed image-boundary regions; corresponding full-model and GT regions provide visual references. Prediction panels report PSNR in dB.

Field-of-view expansion. Block D varies per-side padding at fixed 256^2 input/target resolution, changing the prediction canvas and Gaussian count. Removing the default 16-pixel padding lowers PSNR by 3.168/8.732 dB on StereoNVS/ReplicaGS, with worse SSIM and LPIPS on both, supporting canvas expansion. Increasing padding from 16 to 64 pixels instead improves StereoNVS PSNR by 1.928 dB and LPIPS from 0.1653 to 0.1612, while lowering ReplicaGS PSNR by 0.101 dB. We use 16-pixel padding as the shared default, reducing canvas size and Gaussian count relative to the 64-pixel variant. In Figure 4, the no-padding variant leaves empty target-border bands that the full model fills more completely, illustrating the benefit of prediction beyond the reference field of view.

6 Conclusion and Discussion

StereoGaussians transfers frozen stereo features into metric 3DGS prediction from a single calibrated pair. Two-layer prediction and an expanded canvas support nearby-view extrapolation, while SceneSplat-Stereo supplies consistent target-view supervision from static 3DGS teachers. The model achieves state-of-the-art performance among evaluated methods on StereoNVS and ReplicaGS. Experiments demonstrate the effectiveness of SceneSplat-Stereo for cross-dataset generalisation, while ablations establish the benefits of stereo feature transfer and prediction capacity for newly exposed regions.

Transfer from both stereo encoders demonstrates their value as representation backbones for rendering. Combined with teacher-generated supervision, this provides a practical route from calibrated stereo to reusable Gaussian scenes.

Limitations and directions. Fixed front-layer depth limits stereo-error correction, and teacher artifacts can persist despite reconstruction-quality filtering. Our evaluation focuses on nearby views. Future work includes geometry refinement, supervision from real target photographs, and extending evaluation to larger viewpoint changes and region-specific measures of rendering quality in disoccluded and newly exposed outside-FoV regions.

AI USE STATEMENT

We used generative AI tools to assist with manuscript preparation, including improving grammar, wording, clarity, and organisation. We also used these tools to create selected illustrative components in Figures 1 and 2, which serve explanatory purposes and do not represent experimental outputs. In addition, AI tools assisted with experimental code development and optimisation, as well as the execution of experiments. The reported numerical results were obtained by running the experimental code. We reviewed the AI-assisted code changes and verified the reported experimental results against the corresponding experimental outputs. We take full responsibility for the final manuscript, experimental methodology, implementation, and reported results.

REPRODUCIBILITY STATEMENT

Section 3 and Appendix B describe the construction of SceneSplat-Stereo, including source-scene selection, data splits, camera sampling, and rendering settings. Section 4 and Appendix A specify the model architecture, frozen stereo features, Gaussian parameterisation, and training hyperparameters. Section 5.1 and Appendices C–D describe the evaluation datasets, metrics, and baseline configurations. Appendices E–F provide additional details on higher-resolution experiments and IRS data preparation.

REFERENCES

Wei Bao, Wei Wang, Yuhua Xu, Yulan Guo, Siyu Hong, and Xiaohu Zhang. InStereo2K: A large real dataset for stereo matching in indoor scenes. *Science China Information Sciences*, 2020.

David Charatan, Sizhe Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelSplat: 3D Gaussian splats from image pairs for scalable generalizable 3D reconstruction. In *CVPR*, 2024.

Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. MVSplat: Efficient 3D Gaussian splatting from sparse multi-view images. In *ECCV*, 2024.

Ziwen Chen, Hao Tan, Kai Zhang, Sai Bi, Fujun Luan, Yicong Hong, Fuxin Li, and Zexiang Xu. Long-LRM: Long-sequence large reconstruction model for wide-coverage Gaussian splats. In *ICCV*, 2025. URL https://arxiv.org/abs/2410.12781.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *CVPR*, 2012.

Xiufeng Huang, Ka Chun Cheung, Runmin Cong, Simon See, and Renjie Wan. Stereo-GS: Multi-view stereo vision model for generalizable 3D Gaussian Splatting reconstruction. In *ACM Multimedia*, 2025. URL https://arxiv.org/abs/2507.14921.

Xiaosong Jia, Yihang Sun, Junqi You, Songbur Wong, Zichen Zou, Junchi Yan, Zuxuan Wu, and Yu-Gang Jiang. Efficient-LVSM: Faster, cheaper, and better large view synthesis model via decoupled co-refinement attention. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2602.06478.

Lihan Jiang, Yucheng Mao, Linning Xu, Tao Lu, Kerui Ren, Yichen Jin, Xudong Xu, Mulin Yu, Jiangmiao Pang, Feng Zhao, Dahua Lin, and Bo Dai. AnySplat: Feed-forward 3D Gaussian Splatting from unconstrained views. *ACM Transactions on Graphics (SIGGRAPH Asia)*, 2025. URL https://arxiv.org/abs/2505.23716.

Haian Jin, Hanwen Jiang, Hao Tan, Kai Zhang, Sai Bi, Tianyuan Zhang, Fujun Luan, Noah Snavely, and Zexiang Xu. LVSM: A large view synthesis model with minimal 3d inductive bias. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=QQBPWtvtcn.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d Gaussian splatting for real-time radiance field rendering. In *ACM Transactions on Graphics (SIGGRAPH)*, 2023.

Haechan Lee, Wonjoon Jin, Seung-Hwan Baek, and Sunghyun Cho. Generalizable novel-view synthesis using a stereo camera. In *CVPR*, 2024.

Yue Li, Qi Ma, Runyi Yang, Huapeng Li, Mengjiao Ma, Bin Ren, Nikola Popovic, Nicu Sebe, Ender Konukoglu, Theo Gevers, et al. SceneSplat: Gaussian splatting-based scene understanding with vision-language pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Lu Ling et al. DL3DV-10K: A large-scale scene dataset for deep learning-based 3D vision. In *CVPR*, 2024.

Lahav Lipson, Zachary Teed, and Jia Deng. RAFT-Stereo: Multilevel recurrent field transforms for stereo matching. In *3DV*, 2021.

Tianqi Liu, Guangcong Wang, Shoukang Hu, Liao Shen, Xinyi Ye, Yuhang Zang, Zhiguo Cao, Wei Li, and Ziwei Liu. MVSGaussian: Fast generalizable Gaussian Splatting reconstruction from multi-view stereo. In *ECCV*, 2024. URL https://arxiv.org/abs/2405.12218.

Yifan Liu, Zhiyuan Min, Zhenwei Wang, Junta Wu, Tengfei Wang, Yixuan Yuan, Yawei Luo, and Chunchao Guo. WorldMirror: Universal 3D world reconstruction with any-prior prompting. In *ICML*, 2026. URL https://arxiv.org/abs/2510.10726.

Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *CVPR*, 2016.

Lars Mescheder, Wei Dong, Shiwei Li, Xuyang Bai, Marcel Santos, Peiyun Hu, Bruno Lecouat, Mingmin Zhen, Amaël Delaunoy, Tian Fang, Yanghai Tsin, Stephan R. Richter, and Vladlen Koltun. SHARP: Monocular view synthesis in less than a second. In *arXiv preprint arXiv:2512.10685*, 2025.

Daniel Scharstein, Heiko Hirschmuller, York Kitajima, Greg Krathwohl, Nera Nesic, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *GCPR*, 2014.

Brandon Smart, Chuanxia Zheng, Iro Laina, and Victor Adrian Prisacariu. Splatt3r: Zero-shot Gaussian splatting from uncalibrated image pairs. In *arXiv*, 2024.

Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3D reconstruction. In *CVPR*, 2024.

Stanislaw Szymanowicz, Eldar Insafutdinov, Chuanxia Zheng, Dylan Campbell, João F. Henriques, Christian Rupprecht, and Andrea Vedaldi. Flash3D: Feed-forward generalisable 3D scene reconstruction from a single image. In *3DV*, 2025.

Qiang Wang, Shaohuai Zheng, Qing Yan, Fei Deng, Kaiyong Zhao, and Xiaowen Chu. IRS: A large naturalistic indoor robotics stereo dataset to train deep models for disparity and surface normal estimation. *arXiv preprint arXiv:1912.09678*, 2019a.

Yingqian Wang, Longguang Wang, Jungang Yang, Wei An, and Yulan Guo. Flickr1024: A large-scale dataset for stereo image super-resolution. In *ICCV Workshops*, 2019b.

Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, and Stan Birchfield. FoundationStereo: Zero-shot stereo matching. In *CVPR*, 2025.

Bowen Wen, Shaurya Dewan, and Stan Birchfield. Fast-FoundationStereo: Real-time zero-shot stereo matching. In *CVPR*, 2026.

Gangwei Xu, Xianqi Wang, Xiaohuan Ding, and Xin Yang. Iterative geometry encoding volume for stereo matching. In *CVPR*, 2023.

Haofei Xu, Daniel Barath, Andreas Geiger, and Marc Pollefeys. ReSplat: Learning recurrent Gaussian splatting. *arXiv preprint arXiv:2510.08575*, 2025a. URL https://arxiv.org/abs/2510.08575.

Haofei Xu, Songyou Peng, Fangjinhua Wang, Hermann Blum, Daniel Barath, Andreas Geiger, and Marc Pollefeys. DepthSplat: Connecting Gaussian splatting and depth. In *CVPR*, 2025b.

Guorun Yang, Xiao Song, Chaoqing Huang, Zhidong Deng, Jianping Shi, and Bolei Zhou. DrivingStereo: A large-scale dataset for stereo matching in autonomous driving scenarios. In *CVPR*, 2019.

Botao Ye, Boqi Chen, Haofei Xu, Daniel Barath, and Marc Pollefeys. YoNoSplat: You only need one model for feedforward 3D Gaussian Splatting. *arXiv preprint arXiv:2511.07321*, 2025a. URL https://arxiv.org/abs/2511.07321.

Botao Ye, Sifei Liu, Haofei Xu, Xueting Li, Marc Pollefeys, Ming-Hsuan Yang, and Songyou Peng. No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=P4o9akekdf.

Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. GS-LRM: Large reconstruction model for 3D Gaussian Splatting. In *ECCV*, 2024. URL https://arxiv.org/abs/2404.19702.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *ACM Transactions on Graphics*, 37(4), 2018.