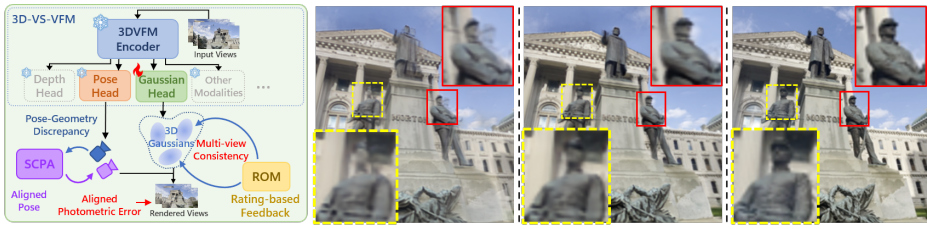
AirSplat: Alignment and Rating for Robust Feed-Forward 3D Gaussian Splatting

Minh-Quan Viet Bui*, Jaeho Moon*, and Munchurl Kim

KAIST, Republic of Korea
{bvmquan, jaeho.moon, mkimee}@kaist.ac.kr
https://kaist-viclab.github.io/airsplat-site



(a) Our AirSplat Training (b) Rendered View from DA3 (c) Rendered View from AirSplat (d) Ground-Truth

Fig. 1: (a) Our proposed AirSplat adapts 3D-VS-VFMs using Self-Consistent Pose Alignment (SCPA) and Rating-based Opacity Matching (ROM) to resolve inherent pose-geometry discrepancies and multi-view inconsistencies. Compared to (b) baseline DA3 [21] which shows floaters (red boxes) and blurred regions (dashed yellow boxes), (c) AirSplat eliminates these artifacts, faithfully recovering the scene structure shown in (d) the ground truth.

Abstract. While 3D Vision Foundation Models (3DVFMs) have demonstrated remarkable zero-shot capabilities in visual geometry estimation, their direct application to generalizable novel view synthesis (NVS) remains challenging. In this paper, we propose **AirSplat**, a novel training framework that effectively adapts the robust geometric priors of 3DVFMs into high-fidelity, pose-free NVS. Our approach introduces two key technical contributions: (1) *Self-Consistent Pose Alignment* (SCPA), a training-time feedback loop that ensures pixel-aligned supervision to resolve pose-geometry discrepancy; and (2) *Rating-based Opacity Matching* (ROM), which leverages the local 3D geometry consistency knowledge from a sparse-view NVS teacher model to filter out degraded primitives. Experimental results on large-scale benchmarks demonstrate that our method *significantly* outperforms state-of-the-art pose-free NVS approaches in reconstruction quality. Our AirSplat highlights the potential of adapting 3DVFMs to enable simultaneous visual geometry estimation and high-quality view synthesis.

Keywords: 3D Reconstruction · 3D Vision · Gaussian Splatting

* Co-first authors (equal contribution).

1 Introduction

The field of Novel View Synthesis (NVS) has undergone a paradigm shift with the advent of 3D Gaussian Splatting (3DGS) [18] and its subsequent advancements [4,11,20,31,44,51]. To overcome the bottleneck of time-intensive per-scene optimization, recent feed-forward architectures [5, 8, 45, 55] directly predict 3D scene parameters from sparse context views. However, dependence on calibrated poses limits 'in-the-wild' applicability, while several pose-free feed-forward methods [12, 17, 34, 53] remain fundamentally constrained to sparse-view inputs. To address this issue, recent works [14, 16, 32, 46, 47] leverage the robust, zero-shot depth and pose estimation capabilities of 3D Vision Foundation Models (3DVFMs) such as MASt3R [19], VGGT [38], and π^3 [42]. These approaches fine-tune or distill 3DVFMs to jointly infer scene 3D Gaussian primitives and camera parameters directly from raw input images. Nevertheless, this biases the networks toward the view synthesis objective, resulting in performance degradation on foundational geometry estimation tasks. More recent unified frameworks, which we refer to as 3DVFMs with view synthesis (3D-VS-VFMs), such as WorldMirror [25] and DepthAnything3 [21], incorporate 3DGS heads to perform both geometry estimation and NVS. Despite this integration, generating high-fidelity novel views from completely uncalibrated context images remains challenging.

We identify two fundamental obstacles that inherently hinder 3D-VS-VFMs from achieving high-fidelity NVS. First, there exists a critical *pose-geometry discrepancy* during the training process. As illustrated in Fig. 3-(a) and (b), current NVS fine-tuning strategies either fail to generalize due to a lack of direct supervision for novel viewpoints (context-only [16]) or suffer from coordinate misalignment (context-target [14]). Specifically, in the context-target setting, target camera poses are inferred using features from both the context and target images, whereas the scene geometry is conditioned strictly on the context views to maintain feed-forward generalizability [14]. However, this asymmetric information flow induces a latent coordinate misalignment between the *predicted target poses* and the *context 3D Gaussian primitives*, leading to degraded optimization (Fig. 4). Second, as shown in Fig.1, the renderings of current 3D-VS-VFMs often contain local multi-view inconsistencies. Specifically, corresponding primitives generated from different context views lack precise spatial consensus. This deficiency leads to local geometric inconsistencies and the generation of 'floaters' that severely degrade spatial stability.

To bridge these fundamental gaps, we introduce **AirSplat**, a pose-free feed-forward 3DGS training framework driven by self-consistent pose **Al**ignment and **R**ating-based feedback that adapts state-of-the-art (SOTA) 3D-VS-VFMs. First, we present **Self-Consistent Pose Alignment (SCPA)** to dynamically anchor the predicted target poses to the scene geometry derived from the context views, thereby providing geometrically-consistent reconstruction supervision. By re-estimating the camera pose from a rendered proxy image, we isolate the systematic geometric bias between the initial pose prediction and the network's 3D Gaussian primitives prediction. We then apply an inverse correction to the initial target pose, aligning it with the estimated 3D primitives, as in Fig. 3-

(c). This entirely decouples the scene reconstruction from the coordinate frame drift during training. Next, we introduce **Rating-based Opacity Matching (ROM)**, a novel optimization strategy designed to enforce multi-view consistency among predicted primitives and eliminate geometrically inconsistent artifacts. This approach is inspired by the paradigm of *learning from rating-based feedback* [1,27,28,43], where an agent's behavior is refined using positive and negative evaluations from an external teacher. In standard rating-based frameworks, an agent is optimized to ensure that its predicted ratings accurately align with the collected feedback from a human or AI evaluator. Adapting this paradigm to 3D reconstruction learning, we utilize a lightweight, sparse-view feed-forward 3DGS model as an algorithmic 'teacher oracle' to provide geometric ratings for the primitives estimated by our network. This feedback effectively partitions the predicted 3D space into preferred (geometrically consistent) and rejected (inconsistent) states. Crucially, we directly formulate a primitive's predicted rating as its opacity. By strictly matching the predicted opacity to the teacher's geometric rating, **AirSplat** implicitly prunes spatial artifacts. In summary, our core **contributions** are as follows:

- We introduce **AirSplat**, a novel framework that fine-tunes 3D-VS-VFMs to generate high-fidelity novel views, while preserving the robust geometry estimation performance.
- We propose **Self-Consistent Pose Alignment (SCPA)** to resolve pose-geometry discrepancies, thereby preventing model degradation caused by misaligned photometric supervision.
- We design **Rating-based Opacity Matching (ROM)**, an optimization strategy that learns from geometric rating-based feedback from a lightweight teacher model to seamlessly filter inconsistent and floating primitives.
- Our AirSplat achieves the state-of-the-art NVS performance *with large margins* on dense-view benchmarks, including RealEstate10K [54], DL3DV [22], and ACID [23], demonstrating robustness in pose-free settings.

2 Related Work

2.1 Optimization-based NVS

The paradigm of novel view synthesis (NVS) has seen a dramatic shift from Neural Radiance Fields (NeRF) [29]. While several NeRF variants [2, 6, 10, 30, 49] successfully compressed training and inference times, 3D Gaussian Splatting (3DGS) [18] broke new ground by modeling scenes with anisotropic primitives and a differentiable rasterizer. Subsequent research [11, 13, 51] has refined this representation to improve reconstruction quality and pose robustness. To mitigate costly per-scene optimization of 3DGS, a parallel line of work seeds the optimization with stronger priors instead of a sparse SfM point cloud: InstantSplat [9] initializes the Gaussians from the dense points of a geometric foundation model in a pose-free manner, QuickSplat [26] learns data-driven priors that jointly predict the initial Gaussians and a learned densification policy, and DAPS-AGF [48]

introduces depth-aware supervision with adaptive gradient-based densification for weakly-observed regions. However, even with such accelerated initialization, these methods still incur seconds of iterative, per-scene refinement for every new environment, preventing the immediate translation of raw pixels into 3D structures. This persistent reliance on scene-specific optimization creates a significant barrier for applications requiring low-latency and scalability, underscoring the need for generalizable, feed-forward architectures.

2.2 3D Vision Foundation Models (3DVFMs)

The rapid scaling of 2D vision models has recently extended into the 3D domain, giving rise to 3DVFMs capable of broad, zero-shot geometric reasoning. Unlike conventional Structure-from-Motion (SfM) pipelines [33, 35] which infer 3D structure through iterative and computationally expensive bundle adjustment, 3DVFMs approach geometry estimation as a feed-forward prediction task. DUSt3R [40] estimates dense corresponding point maps from unposed image pairs. MASt3R [19] further advances this by incorporating dense feature matching heads to enhance correspondence learning. VGGT [38] adopts an alternative attention mechanism to generalize across a variable number of input views. π^3 [42] explores permutation-equivariant prediction to eliminate reference coordinate bias. While the majority of 3DVFMs focus strictly on explicit geometric outputs (e.g., depth, point map, correspondence), recent variants such as WorldMirror [25] and DepthAnything3 (DA3) [21] are equipped with dedicated 3DGS predictions for NVS tasks. We formally classify this specialized subset of architectures as 3DVFMs with a view synthesis head (3D-VS-VFMs).

2.3 Feed-forward NVS

Early generalizable architectures [7, 36, 39, 50] utilized Transformer-based encoders to aggregate image features into NeRF [29] or image-based rendering models. The advent of feed-forward 3DGS [5, 8, 37, 41, 45, 55] has redefined this frontier by directly mapping pixels to 3D Gaussian primitives. Despite their efficiency, these methods typically rely on calibrated camera parameters extracted from off-the-shelf SfM pipelines [33, 35], which fundamentally restricts their applicability in spontaneous, unconstrained environments.

To bypass SfM dependency, several pose-free feed-forward NVS methods [12, 17, 34, 53] primarily focused on sparse-view reconstruction. More recently, a new paradigm has emerged that initializes directly from pre-trained 3DVFMs. Building upon MASt3R [19], NoPoSplat [47] estimates 3D Gaussian primitives in canonical space to avoid explicit pose estimation noise, while SPFSplat [14] proposes to jointly learn pose estimation and NVS, utilizing reprojection losses. AnySplat [16], building upon VGGT [38], jointly optimizes camera poses and 3DGS predictions through photometric supervision and geometric distillation from a 3DVFM teacher model. YoNoSplat [46] explores the mix-forcing training strategy for robust joint camera pose and NVS learning initialized from [42].

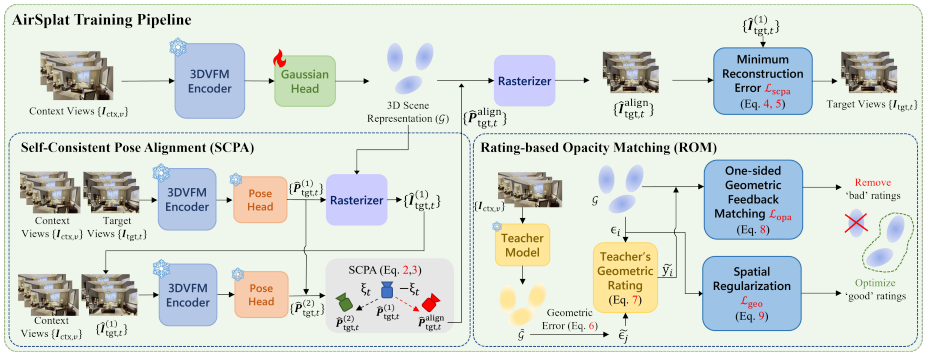


Fig. 2: Overview of our AirSplat training pipeline. During training of a Gaussian head of a 3DVFM, the encoder and its pose head are frozen. Our SCPA module corrects the predicted target pose to align the rendered target views to the GT target views. Our ROM module gathers the geometric feedback from the teacher model, and based on the feedback, it enhances the multi-view consistency of the predicted 3D primitives.

While Rayzer [15], an orthogonal approach, predicts camera and scene representations from scratch via ray-based transformers, it lacks view-count generalization, requiring retraining for varying input numbers.

Recently, 3D-VS-VFMs [21, 25] have proposed learning visual geometry estimation and NVS within a single, unified model. However, despite their zero-shot generalization, a discernible NVS performance gap remains between these unified 3D-VS-VFMs and specialized NVS pipelines [14, 46]. In this work, we identify the critical bottlenecks in the NVS quality of 3D-VS-VFMs: pose-geometry misalignment and multi-view inconsistency. To resolve these, we propose AirSplat, a 3D-VS-VFM training framework that significantly boosts high-fidelity NVS performance, while preserving the integrity of the visual geometry estimation.

3 Proposed Method

3.1 Overview of AirSplat

Given a set of V uncalibrated input context views $\mathcal{I}_{\text{ctx}} = \{\boldsymbol{I}_{\text{ctx},v}\}_{v=1}^{V}$, our model f_ϕ predicts a set of N pixel-aligned 3D Gaussian primitives $\mathcal{G} = \{g_i\}_{i=1}^{N}$ alongside the estimated context camera intrinsics $\hat{\mathcal{K}}_{\text{ctx}} = \{\hat{\boldsymbol{K}}_{\text{ctx},v}\}_{v=1}^{V}$ and extrinsics $\hat{\mathcal{P}}_{\text{ctx}} = \{\hat{\boldsymbol{P}}_{\text{ctx},v}\}_{v=1}^{V}$. Each individual Gaussian primitive g_i is explicitly parameterized by its 3D mean position $\boldsymbol{\mu}_i \in \mathbb{R}^3$, covariance matrix $\boldsymbol{\Sigma}_i$, opacity $\alpha_i \in [0, 1]$, and color $\boldsymbol{c}_i$. To adapt f_ϕ for high-fidelity NVS while preserving its foundational priors, we freeze the main 3DVFM encoder and geometry heads, optimizing only the Gaussian prediction head. Our overall training framework is then driven by two core modules: Self-Consistent Pose Alignment (SCPA), which dynamically resolves pose-geometry discrepancies during target view rendering, and Rating-based Opacity Matching (ROM), which systematically prunes hallucinated artifacts using geometric feedback from a teacher model.

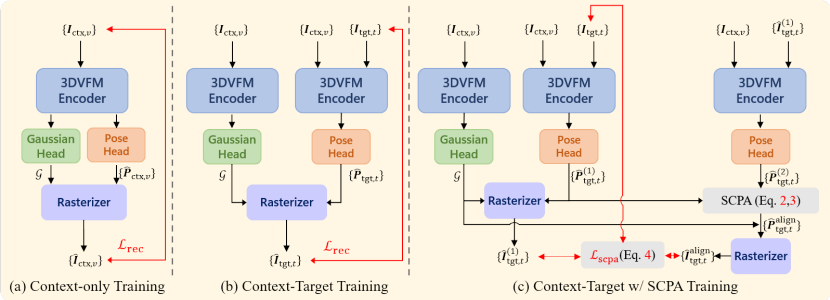


Fig. 3: Comparison of training paradigms in pose-free NVS. (a) Training with the context-only strategy, adopted in [16], leads to a lack of direct supervision for novel viewpoints. (b) The context-target strategy, following [14], results in spatial misalignment. (c) Our SCPA corrects the inherent spatial drift, enabling the network to learn both structurally consistent 3D geometry and robust novel view synthesis.

3.2 Self-Consistent Pose Alignment (SCPA)

Pose-Geometry Discrepancy. During pose-free NVS training, $\mathcal{G}$ is optimized by rasterizing it onto the image-space using a set of predicted target camera parameters ($\hat{\mathcal{P}}_{\text{tgt}} = \{\hat{\boldsymbol{P}}_{\text{tgt},t}\}_{t=1}^{T}, \hat{\mathcal{K}}_{\text{tgt}} = \{\hat{\boldsymbol{K}}_{\text{tgt},t}\}_{t=1}^{T}$) corresponding to T ground-truth target views $\mathcal{I}_{\text{tgt}} = \{\boldsymbol{I}_{\text{tgt},t}\}_{t=1}^{T}$. This rasterization process yields the synthesized target images $\hat{\mathcal{I}}_{\text{tgt}} = \{\hat{\boldsymbol{I}}_{\text{tgt},t}\}_{t=1}^{T}$. Fig. 3 illustrates two prevailing paradigms for sampling these views during training. The context-only approach, following [16], trivially restricts $\mathcal{I}_{\text{tgt}}$ to be identical to $\mathcal{I}_{\text{ctx}}$, leading to lack of direct supervision for novel viewpoints. Conversely, SPFSplat [14] introduces a context-target strategy, sampling target views $\mathcal{I}_{\text{tgt}}$ that are spatially distinct from $\mathcal{I}_{\text{ctx}}$. This necessitates an additional forward pass to estimate $(\hat{\mathcal{P}}_{\text{tgt}}, \hat{\mathcal{K}}_{\text{tgt}})$ by feeding the concatenation of $\mathcal{I}_{\text{ctx}}$ and $\mathcal{I}_{\text{tgt}}$ into the 3DVFM. While this disentanglement improves robustness to interpolated views since the 3D primitives $\mathcal{G}$ are derived solely from $\mathcal{I}_{\text{ctx}}$ and supervised by distinct target observations, it introduces a critical structural flaw. Specifically, the asymmetric information flow, where geometry extraction relies exclusively on context views while pose estimation relies on both context and target views, induces a fundamental *pose-geometry discrepancy* during training. As illustrated in Fig. 4, the error between the rendered view $\hat{\boldsymbol{I}}_{\text{tgt},t}$ ('Rendered View w/ $\hat{\boldsymbol{P}}_{\text{tgt},t}$') and $\boldsymbol{I}_{\text{tgt},t}$ (GT target View) is dominated by spatial misalignment rather than photometric degradation, which is the consequence of the *pose-geometry discrepancy*. When supervised directly via a pixel-wise reconstruction loss, this extrinsic spatial shift yields corrupted gradients, resulting in unstable optimization.

Self-Consistent Pose Alignment. To mitigate this pose-geometry discrepancy, we propose *Self-Consistent Pose Alignment* (SCPA), a self-correcting strategy that mathematically quantifies and reverses the spatial drift induced by the context-target training strategy, rather than directly applying photometric supervision on misaligned renderings. Based on the predicted $\mathcal{G}$ from $\mathcal{I}_{\text{ctx}}$ and the initial target pose estimates $\hat{\mathcal{P}}_{\text{tgt}}^{(1)}$ from the concatenation of $(\mathcal{I}_{\text{ctx}}, \mathcal{I}_{\text{tgt}})$, we first

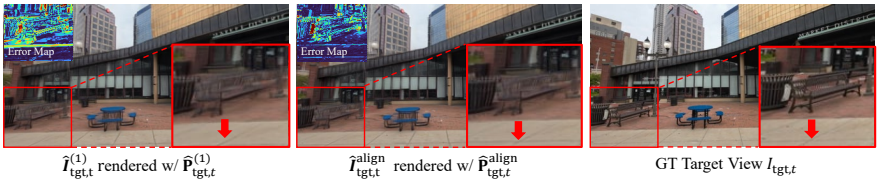


Fig. 4: Effect of Self-Consistent Pose Alignment (SCPA). We compare rendered target views using the initial predicted pose $\hat{\boldsymbol{P}}^{(1)}_{\text{tgt},t}$, our aligned pose $\hat{\boldsymbol{P}}^{\text{align}}_{\text{tgt},t}$ against the ground truth, during training. As highlighted by the red arrows, the initial pose prediction results in a noticeable spatial shift, evident in the misaligned structural lines on the ground. Our proposed SCPA corrects the inherent spatial drift and ensures that the model learn structurally consistent 3D geometry and robust NVS. In the error maps, blue indicates small errors, and red indicates large errors.

render an initial set of synthesized target images $\hat{\mathcal{I}}^{(1)}_{\text{tgt}}$. We then feed the concatenation of $\mathcal{I}_{\text{ctx}}$ and $\hat{\mathcal{I}}^{(1)}_{\text{tgt}}$ back into f_ϕ to yield a second set of pose predictions:

$$\hat{\mathcal{P}}^{(2)}_{\text{tgt}} = f_\phi(\mathcal{I}_{\text{ctx}}, \hat{\mathcal{I}}^{(1)}_{\text{tgt}}). \tag{1}$$

Empirically, rendering a second set of images $\hat{\mathcal{I}}^{(2)}_{\text{tgt}}$ from $\hat{\mathcal{P}}^{(2)}_{\text{tgt}}$ reveals that f_ϕ exhibits a systematic, repeated drift when mapping synthesized geometry back to the pose manifold (see *Suppl.* for detailed visualizations). From this observation, we propose to compute the corrected poses $\hat{\mathcal{P}}^{\text{align}}_{\text{tgt}} = \{\hat{\boldsymbol{P}}^{\text{align}}_{\text{tgt},t}\}^{T}_{t=1}$ that better align with the predicted scene geometry space by reversing the observed pose discrepancy between $\hat{\boldsymbol{P}}^{(1)}_{\text{tgt},t}$ and $\hat{\boldsymbol{P}}^{(2)}_{\text{tgt},t}$.

To approximate the aligned pose $\hat{\boldsymbol{P}}^{\text{align}}_{\text{tgt},t}$, we compute the relative transformation between the initial prediction $\hat{\boldsymbol{P}}^{(1)}_{\text{tgt},t}$ and the re-predicted pose $\hat{\boldsymbol{P}}^{(2)}_{\text{tgt},t}$. We denote this transformation as $\boldsymbol{\xi}_t$ which is a 6-vector of coordinates in the Lie algebra $\mathfrak{se}(3)$:

$$\boldsymbol{\xi}_t = \log\left(\hat{\boldsymbol{P}}^{(2)}_{\text{tgt},t}(\hat{\boldsymbol{P}}^{(1)}_{\text{tgt},t})^{-1}\right)^{\vee}, \tag{2}$$

where $\log(\cdot)^{\vee} : \mathbf{SE}(3) \to \mathfrak{se}(3)$ denotes the logarithm map [3]. To perform the self-consistent pose alignment, we back-extrapolate along the manifold by negating $\boldsymbol{\xi}_t$. This negated vector is mapped back to the $\mathbf{SE}(3)$ manifold via the exponential map $\exp(\cdot^{\wedge}) : \mathfrak{se}(3) \to \mathbf{SE}(3)$ and applied to $\hat{\boldsymbol{P}}^{(1)}_{\text{tgt},t}$:

$$\hat{\boldsymbol{P}}^{\text{align}}_{\text{tgt},t} = \exp\left((-\boldsymbol{\xi}_t)^{\wedge}\right)\hat{\boldsymbol{P}}^{(1)}_{\text{tgt},t}. \tag{3}$$

Finally, to ensure training stability and prevent degradation in cases where the re-prediction fails, we employ a minimum-error supervision strategy. We render the aligned images $\hat{\mathcal{I}}^{\text{align}}_{\text{tgt}}$ from $\hat{\mathcal{P}}^{\text{align}}_{\text{tgt}}$, and define the $\mathcal{L}_{\text{scpa}}$ as the minimum reconstruction error between the aligned $\hat{\boldsymbol{I}}^{\text{align}}_{\text{tgt},t}$ and initial $\hat{\boldsymbol{I}}^{(1)}_{\text{tgt},t}$ renderings:

$$\mathcal{L}_{\text{scpa}} = \sum_{t=1}^{T} \min\left(\mathcal{L}_{\text{rec}}(\hat{\boldsymbol{I}}^{\text{align}}_{\text{tgt},t}, \boldsymbol{I}_{\text{tgt},t}), \mathcal{L}_{\text{rec}}(\hat{\boldsymbol{I}}^{(1)}_{\text{tgt},t}, \boldsymbol{I}_{\text{tgt},t})\right). \tag{4}$$

Following standard practices [8, 45], $\mathcal{L}_{\text{rec}}$ is defined as a weighted sum of the Mean Squared Error (MSE) and the LPIPS perceptual metric [52]:

$$\mathcal{L}_{\text{rec}}(\hat{\boldsymbol{I}}_{\text{tgt}}, \boldsymbol{I}_{\text{tgt}}) = \|\hat{\boldsymbol{I}}_{\text{tgt}} - \boldsymbol{I}_{\text{tgt}}\|_2^2 + \lambda_s \text{LPIPS}(\hat{\boldsymbol{I}}_{\text{tgt}}, \boldsymbol{I}_{\text{tgt}}). \tag{5}$$

By dynamically supervising the model with the aligned viewpoint, SCPA effectively suppresses the emergence of artifacts caused by the pose-geometry discrepancy of the context-target training strategy.

3.3 Rating-based Opacity Matching (ROM)

While our SCPA successfully ensures global coordinate alignment, the predicted 3D Gaussian primitives $\mathcal{G}$ may still exhibit local geometric inconsistencies. These artifacts, known as 'floaters,' often arise from spatial discrepancies among primitives that represent the same 3D regions but are predicted from disparate views. To suppress these artifacts, we introduce **Rating-based Opacity Matching (ROM)**. We ground this module in the paradigm of *Learning from Rating-Based Feedback* [1,28,43], where an agent is refined by an external oracle that provides absolute ratings of the agent's proposed states. As in Fig. 2, ROM consists of two stages: *Teacher's Geometric Rating* and *One-sided Geometric Feedback Matching*. In the Teacher's Geometric Rating stage, we utilize a pre-trained, lightweight sparse-view NVS model [45] as an algorithmic teacher $f_{\tilde{\phi}}$. For any predicted Gaussian primitive g_i, the teacher evaluates its multi-view structural consensus to provide a geometric rating $\tilde{y}_i \in [0, 1]$. The primitives exhibiting high geometric error across views are assigned an 'bad' rating ($\tilde{y}_i \to 0$), while spatially consistent primitives are rated as 'good' ($\tilde{y}_i \to 1$). Then, in the One-sided Geometric Feedback Matching stage, we train our model f_ϕ to match its predicted ratings y_i with the target ratings $\tilde{y}_i$. Since the 'bad' primitives must be physically removed from the rendered scene, we directly formulate the predicted rating y as the primitive's opacity value α. Under this formulation, the feedback matching process naturally enables our model to prune artifacts by learning to drive their opacities to zero.

Teacher's Geometric Rating. To compute the geometric rating, we first quantify the local multi-view consistency of a Gaussian primitive g_i predicted from the context view $\boldsymbol{I}_{\text{ctx},v}$. Specifically, we project its 3D mean $\boldsymbol{\mu}_i$ onto an adjacent context view $\boldsymbol{I}_{\text{ctx},v'}$ and measure the Euclidean distance between $\boldsymbol{\mu}_i$ and the 3D mean of the primitive sampled at that corresponding projected pixel. Let $\Pi_{v \to v'}(\boldsymbol{\mu}_i)$ denote the operation of projecting $\boldsymbol{\mu}_i$ onto the image plane of $\boldsymbol{I}_{ctx,v'}$, and then spatially sampling the 3D Gaussian mean at that projected 2D location. The scale-normalized geometric error ϵ_i for g_i is formulated as:

$$\epsilon_i = \frac{\left\|\boldsymbol{\mu}_i - \Pi_{v \to v'}(\boldsymbol{\mu}_i)\right\|_2}{\text{median}(\boldsymbol{D}_i) + \eta}, \tag{6}$$

where $\boldsymbol{D}_i$ represents the projected depth value, ensuring scale-invariance of ϵ_i, and $\eta = 10^{-8}$ is a small constant to maintain numerical stability. Next, we extract the corresponding prediction from the teacher model $f_{\tilde{\phi}}$. By feeding the

view pair $(\boldsymbol{I}_{\text{ctx},v}, \boldsymbol{I}_{\text{ctx},v'})$ into $f_{\tilde{\phi}}$, we generate the teacher's sparse-view 3D primitives $\tilde{\mathcal{G}} = \{\tilde{g}_j\}_{j=1}^{\tilde{N}}$. Let $\tilde{g}_j$ (with 3D mean $\tilde{\boldsymbol{\mu}}_j$) represent the teacher's predicted primitive originating from the exact same pixel in $\boldsymbol{I}_{\text{ctx},v}$ as g_i. Following the identical protocol in Eq. 6, we compute the teacher's normalized geometric error $\tilde{\epsilon}_j$ which corresponds to our model's geometric error ϵ_i.

To formulate the teacher's rating feedback, we compute the continuous excess geometric error as $E_i^{\text{geo}} = \max(0, \epsilon_i - \tilde{\epsilon}_j)$. Then, we generate a continuous rating $\tilde{y}_i \in (0, 1]$ for each Gaussian mean $\boldsymbol{\mu}_i$ that exponentially decays as our model's structural consensus degrades relative to the teacher:

$$\tilde{y}_i = \exp\Big(- \lambda \cdot \text{sg}[E_i^{\text{geo}}]) \Big) \tag{7}$$

where λ governs the decay rate (set to 5.0), and sg$[\cdot]$ is the stop-gradient operator. In Eq. 7, if the geometric error of our model's prediction is equal to or lower than the teacher's one ($\epsilon_i \leq \tilde{\epsilon}_j$), the target rating is set to 1. As the excess geometric error increases, the rating smoothly approaches to 0.

One-sided Geometric Feedback Matching. The next step is to match our model's predicted rating y which we define as the predicted opacity α with the teacher's continuous rating $\tilde{y}$. Unlike previous rating-based models [27, 43] that enforce symmetric matching for both positive and negative feedback, we propose a *one-sided matching*. Because a geometrically 'bad' primitive should physically disappear from the scene, its opacity must be driven toward zero. Conversely, a 'good' primitive's opacity is inherently ambiguous: it may represent a semi-transparent surface (low α) or a solid object (high α). Forcing valid primitives to match a strict rating of $\tilde{y} = 1$ would cause severe 'solidification' artifacts, destroying volumetric blending. Therefore, we treat the teacher's rating $\tilde{y}$ as a strict upper bound rather than an absolute target. To conform to the teacher's geometric feedback of 'bad' primitive, our model's predicted rating (opacity) must be smaller than $\tilde{y}$. This is achieved via a pointwise margin loss $\mathcal{L}_{\text{opa}}$:

$$\mathcal{L}_{\text{opa}} = \frac{1}{N} \sum_{i=1}^{N} \Big(\max(0, \alpha_i - \tilde{y}_i) \Big)^2 \tag{8}$$

Under this formulation, primitives with 'good' feedback ($\tilde{y}_i = 1$) incur zero penalty for α, yielding optimization control entirely to the photometric rendering loss $\mathcal{L}_{\text{scpa}}$. Artifacts with 'bad' feedback ($\tilde{y}_i \to 0$), however, are aggressively penalized, naturally driving the network to prune floaters.

Spatial Regularization for Geometric Consolidation. For completeness, relying solely on opacity compression to remove artifacts can lead to overly sparse representations if mildly misaligned primitives are aggressively pruned. To prevent this, we complement the rating matching with a direct spatial regularization term, $\mathcal{L}_{\text{geo}}$. This loss explicitly minimizes the geometric error ϵ_i of the predicted primitives. Crucially, we clamp the maximum error input at $\tau = 2.0$. This ensures that massive errors do not produce exploding gradients that destabilize the entire 3D scene. Furthermore, the loss is weighted by the gradient-detached

primitive's predicted opacity, $\text{sg}[\alpha_i]$, ensuring that the network prioritizes the spatial regularization of highly visible structures over transparent background noise. The loss is formulated as:

$$\mathcal{L}_{\text{geo}} = \frac{1}{N} \sum_{i=1}^{N} \text{sg}[\alpha_i] \cdot \min(\epsilon_i, \tau). \tag{9}$$

Combining $\mathcal{L}_{\text{geo}}$ and $\mathcal{L}_{\text{opa}}$ establishes a complementary optimization framework that explicitly isolates repairable geometry from unrecoverable noise. $\mathcal{L}_{\text{geo}}$ acts as a spatial regularizer, pulling mildly deviant 3D coordinates into local multi-view consensus. However, when severe inconsistencies arise that cannot be resolved by $\mathcal{L}_{\text{geo}}$, our $\mathcal{L}_{\text{opa}}$ smoothly assumes control. It prunes these artifacts by driving their existence likelihood ($\alpha \to 0$) based on the teacher's ratings.

3.4 Loss Functions

The total loss $\mathcal{L}_{\text{total}}$ is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{scpa}} + \lambda_{\text{geo}}\mathcal{L}_{\text{geo}} + \lambda_{\text{opa}}\mathcal{L}_{\text{opa}}, \tag{10}$$

where λ_{geo} and λ_{opa} are weighting hyperparameters that balance the reconstruction fidelity against the geometric priors.

4 Experiments

4.1 Implementation Details

We adopt DA3-GIANT [21] as our baseline. We follow DA3 to freeze the main encoder and the pre-trained depth/pose heads, optimizing only the Gaussian prediction head to maintain the powerful geometric priors of the foundation model. The Gaussian head predictions include 3D refinement for Gaussian means and other Gaussian attributes such as scales, rotations, and opacities. We train our model on 252×252 images for RealEstate10K (RE10K) [54], and 252×448 images for the DL3DV dataset [22]. During each training iteration, we randomly sample 24 context views to construct the scene geometry and 8 spatially distinct target views for rendering supervision. Training is conducted on 4 NVIDIA A100 40GB GPUs with a batch size of 1 per GPU, totaling $100,000$ iterations, while testing is performed on a single A100 40GB. Further details regarding our hyperparameters and model's weights update are provided in the *Suppl.*

4.2 Evaluation Protocol

We evaluate the NVS quality via PSNR, SSIM, and LPIPS [52]. We adopt the evaluation protocol established by recent pose-free NVS literature [14, 32, 47] on the RE10K dataset. This setting stratifies testing sequences based on the visual overlap ratios between the initial and terminal frames. We utilize sequences with

Table 1: Quantitative comparison of NVS performance on the RE10K dataset [54] under various input-view settings. **Bold** indicates the best performance. OOM indicates out-of-memory inference error.

	Methods	12 Views			24 Views			36 Views		
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
w/ Pose	MonoSplat [24] (CVPR25)	18.16	0.663	0.336	16.66	0.593	0.391	15.79	0.551	0.424
	MVSplat [8] (ECCV24)	17.98	0.638	0.357	17.27	0.609	0.379	OOM	OOM	OOM
	DepthSplat [45] (CVPR25)	22.56	0.793	0.200	21.00	0.734	0.248	19.60	0.676	0.296
Pose-free	NoPoSplat [47] (ICLR25)	17.15	0.571	0.437	17.10	0.570	0.443	17.09	0.570	0.446
	AnySplat [16] (SIGGRAPHAsia25)	18.69	0.591	0.273	19.15	0.610	0.257	19.31	0.615	0.251
	WorldMirror [25] (arXiv25)	21.23	0.707	0.267	21.08	0.701	0.274	20.98	0.699	0.275
	SPFSplat [14] (ICCV25)	21.57	0.701	0.254	21.32	0.694	0.266	21.17	0.689	0.273
	YoNoSplat [46] (ICLR26)	21.62	0.679	0.229	21.63	0.679	0.227	21.60	0.681	0.226
	DA3 [21] (ICLR26)	20.78	0.715	0.250	21.06	0.710	0.254	21.11	0.684	0.274
	AirSplat(Ours)	**23.08**	**0.799**	**0.190**	**23.77**	**0.814**	**0.178**	**23.94**	**0.815**	**0.179**

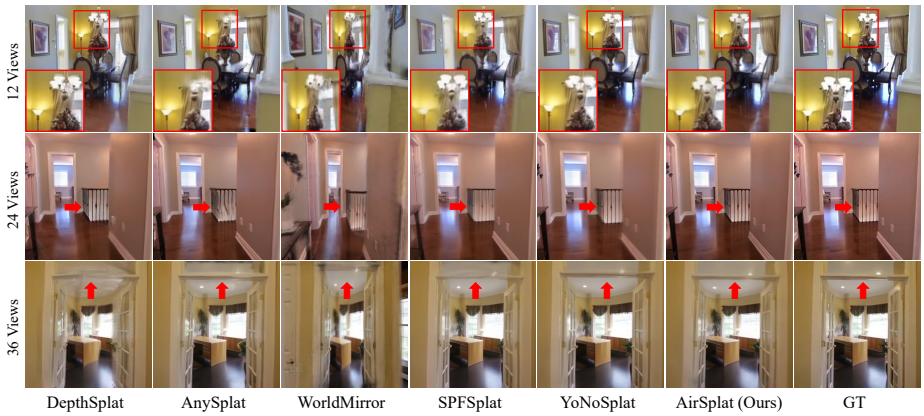


Fig. 5: Qualitative comparison of NVS performance on RE10K dataset [54].

overlap ratios of less than 10%. Furthermore, we evaluate the existing methods under wide-baseline conditions on the complex DL3DV dataset by constructing challenging context-target splits. Specifically, we constrain the frame intervals between the starts and ends of the testing sequences to 90, 150, and 150 frames for the 12-, 24-, and 36-view settings, respectively. Target views are uniformly sampled across these sequences, while context views are randomly drawn from the mutually exclusive remaining frames. To analyze models' robustness, we perform the cross-dataset generalization evaluation on the ACID dataset [23] as in [8, 14], by utilizing the testing splits from [32]. Similar to prior pose-free approaches [11, 14, 46, 47], we adopt the test-time pose optimization for ground truth alignment for all pose-free NVS baselines.

4.3 NVS Performance Evaluation

Comparison on RE10K. We evaluate our model on the RE10K dataset to verify its scalability and robustness in large-scale indoor and outdoor environ-

Table 2: Quantitative comparison of NVS performance on the DL3DV dataset [22] under various input-view settings.

	Methods	12 Views			24 Views			36 Views		
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
w/ Pose	MVSplat [8] (ECCV24)	21.55	0.729	0.239	OOM	OOM	OOM	OOM	OOM	OOM
	DepthSplat [45] (CVPR25)	22.14	0.725	0.221	19.87	0.695	0.274	18.69	0.643	0.330
Pose-free	NoPoSplat [47] (ICLR25)	16.13	0.447	0.497	14.73	0.392	0.603	13.82	0.365	0.640
	AnySplat [16] (SIGGRAPHAsia25)	18.72	0.551	0.310	18.40	0.533	0.333	18.36	0.529	0.337
	WorldMirror [25] (arXiv25)	20.44	0.625	0.278	19.78	0.594	0.312	19.67	0.589	0.316
	YoNoSplat [46] (ICLR26)	17.73	0.481	0.430	16.77	0.451	0.490	16.56	0.446	0.502
	DA3 [21] (ICLR26)	20.74	0.691	0.242	20.51	0.644	0.274	20.38	0.642	0.285
	AirSplat(Ours)	**22.50**	**0.747**	**0.207**	**22.22**	**0.735**	**0.217**	**22.07**	**0.730**	**0.225**

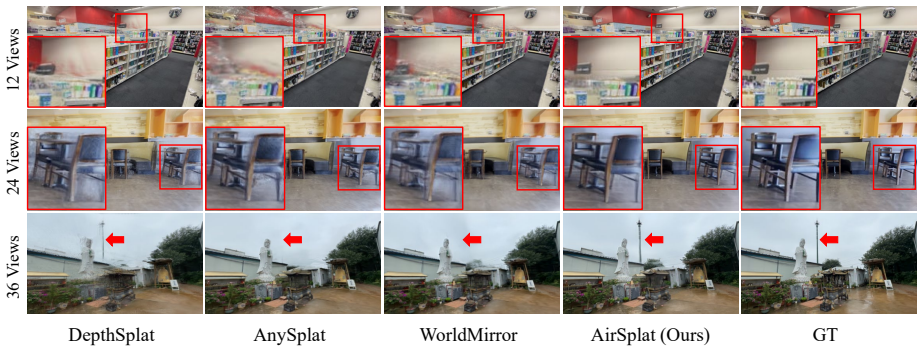


Fig. 6: Qualitative comparison of NVS performance on DL3DV dataset [22].

ments. As shown in Table 1, our proposed AirSplat achieves SOTA performance among pose-free methods, demonstrating superior NVS quality. Most notably, our method not only outperforms existing pose-free benchmarks such as YoNoSplat [46] and WorldMirror [25] by significant margins, but also remarkably exceeds the performance of several pose-required methods [24, 45]. These results indicate that our approach effectively compensates for the absence of ground-truth camera poses by leveraging powerful geometric reasoning based on our SCPA and ROM, ultimately leading to more accurate and sharper novel view renderings compared to prior pipelines. In Fig. 5, our proposed AirSplat successfully preserves challenging high-frequency details, such as thin structures, while aggressively suppressing the floater artifacts and blurry regions that frequently degrade the outputs of prior methods.

Comparison on DL3DV. Table 2 reports the quantitative NVS performance on the DL3DV dataset compared to recent SOTA methods [8, 16, 25, 45–47]. Our AirSplat consistently achieves superior performance across all evaluation settings (12, 24, 36 views) in PSNR, SSIM, and LPIPS metrics. Notably, in the 12-view setting, despite being a pose-free approach, our method surpasses the established SOTA pose-required method, DepthSplat [45]. Moreover, AirSplat outperforms the baseline DA3 by a *significant* margin of +1.76 dB in PSNR and reduces LPIPS by 14.4%. While base 3D-VS-VFMs (e.g., WorldMirror [25], DA3 [21])

Table 3: Cross-dataset generalization performance under various input-view settings. We evaluate the zero-shot performance on the ACID dataset [23].

	Methods	16 Views			20 Views			24 Views		
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
w/ Pose	MonoSplat [24] (CVPR25)	17.81	0.642	0.345	17.48	0.625	0.361	17.23	0.611	0.372
	MVSplat [8] (ECCV24)	18.19	0.502	0.376	18.12	0.500	0.379	18.09	0.500	0.379
	DepthSplat [45] (CVPR25)	20.41	0.713	0.268	19.92	0.694	0.286	19.78	0.688	0.289
Pose-free	NoPoSplat [47] (ICLR25)	22.30	0.668	0.286	22.24	0.666	0.288	22.25	0.666	0.288
	AnySplat [16] (SIGGRAPHAsia25)	21.89	0.615	0.275	22.05	0.622	0.256	21.96	0.619	0.258
	WorldMirror [25] (arXiv25)	22.15	0.646	0.275	22.20	0.650	0.275	22.34	0.658	0.273
	SPFSplat [14] (ICCV25)	24.58	0.725	0.218	24.49	0.722	0.221	24.40	0.720	0.222
	YoNoSplat [46] (ICLR26)	22.49	0.641	0.270	22.48	0.641	0.271	22.47	0.642	0.272
	DA3 [21] (ICLR26)	22.13	0.687	0.272	23.21	0.690	0.262	23.31	0.694	0.262
	AirSplat(Ours)	**25.96**	**0.796**	**0.188**	**26.21**	**0.803**	**0.178**	**26.42**	**0.813**	**0.176**

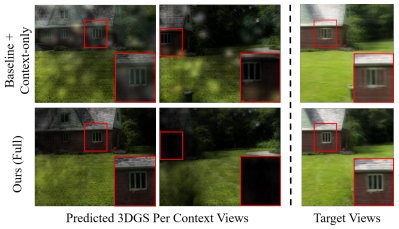


Fig. 7: Visualization of the 'floaters' compression in the predicted 3DGS per context views.

Table 4: Ablation study on our SCPA and ROM. NVS performance comparison on the RE10K dataset under 12 input views.

Methods	PSNR↑	SSIM↑	LPIPS↓	GPU Mem
Baseline (DA3 [21])	20.78	0.715	0.250	-
Baseline + Context-only Training	21.27	0.745	0.253	15.17
Baseline + Context-Target Training	21.63	0.761	0.241	19.92
Baseline + Context-Target w/ SCPA Training	22.60	0.776	0.215	23.45
Baseline + Context-Target Training + ROM	22.41	0.769	0.211	22.30
Ours (Full)	**23.08**	**0.799**	**0.190**	30.56

are robust with increased input views, applying our proposed training strategy significantly amplifies DA3's rendering quality, yielding substantial margins in all metrics. Furthermore, as shown in Fig. 6, our AirSplat synthesizes clean, floater-free novel views across varying input densities. Specifically, in the third row, while other methods fail to capture thin structures, our AirSplat accurately reconstructs the sharp geometry of the pole, further validating the robust multi-view consistency and structural integrity enforced by our pipeline.

Cross Dataset Generalization. Following [14,32], we conduct a cross-dataset evaluation at a resolution of 252×252 on the ACID dataset. All models were evaluated using 16, 20, and 24 input views without further fine-tuning. As shown in Table 3, our method achieves the highest performance across all metrics, significantly outperforming both pose-free and pose-required baselines. Notably, in this zero-shot setting, our model maintains a substantial lead over other pose-free NVS methods such as SPFSplat [14], YoNoSplat [46], and 3D-VS-VFMs [21,25]. This superior performance in unseen environments demonstrates that our AirSplat does not merely overfit to the training distribution but instead learns highly generalizable geometric representations and robust view-consistency, effectively bridging the gap between pose-free and pose-dependent NVS.

4.4 Ablation Study

In Table 4, we conduct an ablation study to evaluate the individual contributions of our proposed components: Self-Consistent Pose Alignment (SCPA) and Rating-based Opacity Matching (ROM) on the RE10K dataset [54].

Effectiveness of SCPA. The baseline model, DA3 [21], shows limited performance (PSNR: 20.78, SSIM: 0.715), which underperforms the pose-free NVS methods [25,46,47]. As discussed in Sec. 3.2, implementing a context-only training strategy yields only a marginal improvement over the baseline (reaching 21.27 dB PSNR), inherently limited by the absence of novel target views supervision. While adopting a naive context-target sampling strategy elevates this performance to 21.63 dB, the model remains fundamentally bottlenecked by the pose-geometry discrepancy. This uncorrected spatial misalignment ultimately results in structurally degraded 3D Gaussian primitives and blurry target renderings. Integrating our SCPA results in a significant performance leap, reaching 22.60 dB (+0.97 dB over context-target). The reduction in LPIPS from 0.250 (Baseline) to 0.215 indicates that SCPA effectively restores high-frequency details by ensuring a geometrically consistent rendering path during training.

Effectiveness of ROM. The integration of Rating-based Opacity Matching (ROM) independently improves the baseline to 22.41 dB PSNR. This validates the necessity of local structural consensus for high-fidelity synthesis. As illustrated in Fig. 2, ROM acts as a "geometric filter" by matching the student's predicted opacity α with the teacher's geometric rating $\tilde{y}$. The improvement in perceptual metrics confirms that ROM effectively prunes floaters via opacity compression. By explicitly penalizing spatial inconsistencies, ROM ensures that only geometrically valid structures contribute to the final rendering. As shown in Fig. 7, AirSplat learns from ROM to suppress inconsistent geometry of complex regions (e.g., the window structures highlighted in red boxes). The 'Baseline + Context-only' variant produces noisy 'floaters' at each context view's 3DGS estimation, resulting in severe blurring artifacts. On the other hand, our full model assigns near-zero opacity for these inconsistent primitives, resulting in sharp and high-quality reconstructions for target views.

Full Model. Our full model, AirSplat, which combines both SCPA and ROM, achieves the best performance across all metrics, yielding a 25% improvement in perceptual scores compared to naive fine-tuning. Although the simultaneous integration of both modules increases the peak training memory consumption to 30.56 GB, this computational footprint is strictly confined to the offline training phase, leaving the feed-forward inference speed and memory requirements entirely unaffected. The synergy between global pose alignment and local structural refinement allows our framework to synthesize sharp, consistent novel views from uncalibrated context images. These results validate our core hypothesis that decoupling the pose errors from geometry optimization, complemented by teacher-driven consistency feedback, is essential for robust feed-forward NVS.

5 Limitations

A primary limitation of our current framework, common among deterministic feed-forward NVS models, is its strict reliance on observed input context views. Since AirSplat does not explicitly hallucinate unseen structures, severe occlusions or entirely uncaptured areas may manifest as structural voids in the rendered novel views. To resolve these deterministic blind spots, future directions of this framework could integrate generative video diffusion priors as a post-processing module, allowing for the temporally consistent inpainting of geometrically plausible content within these occlusions. Furthermore, scaling to thousands of input context views is limited by the backbone's pixel-aligned primitives and self-attention cost, which could be addressed by pixel-unaligned primitives together with test-time training or causal attention mechanisms for multi-view feature extraction. Finally, AirSplat currently focuses on static scenes and hence may degrade on inputs with severe motion or 3D inconsistency.

6 Conclusion

In this paper, we presented AirSplat, a novel training framework designed for high-fidelity, pose-free novel view synthesis by effectively leveraging the geometric priors of 3D Vision Foundation Models (3DVFMs). Our approach identifies and addresses two fundamental challenges in existing pose-free NVS paradigms: (i) global pose-geometry discrepancy and (ii) local multi-view inconsistency. Through the Self-Consistent Pose Alignment (SCPA), we introduce a training-time feedback loop that dynamically anchors predicted target poses to the context-derived scene geometry, effectively decoupling coordinate drift from photometric optimization. Furthermore, we propose Rating-based Opacity Matching (ROM), which utilizes geometric feedback from a sparse-view NVS teacher model to systematically prune artifacts and floaters. Experimental results on large-scale benchmarks, including RE10K, DL3DV and ACID, demonstrate that our AirSplat achieves state-of-the-art performance with large margins, while preserving the foundational geometry estimation performance.

7 Acknowledgement

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korean Government [Ministry of Science and ICT (Information and Communications Technology)] (Project Number: RS-2022-00144444, Project Title: Deep Learning Based Visual Representational Learning and Rendering of Static and Dynamic Scenes, 100%).

References

1. Arumugam, D., Lee, J.K., Saskin, S., Littman, M.L.: Deep reinforcement learning from policy-dependent human feedback. arXiv preprint arXiv:1902.04257 (2019)
2. Bello, J.L.G., Bui, M.Q.V., Kim, M.: Pronerf: Learning efficient projection-aware ray sampling for fine-grained implicit neural radiance fields. IEEE Access (2024). https://doi.org/10.1109/ACCESS.2024.3390753
3. Blanco-Claraco, J.L.: A tutorial on **SE**(3) transformation parameterizations and on-manifold optimization. CoRR (2021)
4. Bui, M.Q.V., Park, J., Bello, J.L.G., Moon, J., Oh, J., Kim, M.: Mobgs: Motion deblurring dynamic 3d gaussian splatting for blurry monocular video. In: AAAI (2026)
5. Charatan, D., Li, S.L., Tagliasacchi, A., Sitzmann, V.: pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In: CVPR (2024)
6. Chen, A., Xu, Z., Geiger, A., Yu, J., Su, H.: Tensorf: Tensorial radiance fields. In: ECCV (2022)
7. Chen, A., Xu, Z., Zhao, F., Zhang, X., Xiang, F., Yu, J., Su, H.: Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In: ICCV (2021)
8. Chen, Y., Xu, H., Zheng, C., Zhuang, B., Pollefeys, M., Geiger, A., Cham, T.J., Cai, J.: Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In: ECCV (2024)
9. Fan, Z., Wen, K., Cong, W., Wang, K., Zhang, J., Ding, X., Xu, D., Ivanovic, B., Pavone, M., Pavlakos, G., Wang, Z., Wang, Y.: Instantsplat: Sparse-view gaussian splatting in seconds (2024)
10. Fridovich-Keil, S., Yu, A., Tancik, M., Chen, Q., Recht, B., Kanazawa, A.: Plenoxels: Radiance fields without neural networks. In: CVPR (2022)
11. Fu, Y., Liu, S., Kulkarni, A., Kautz, J., Efros, A.A., Wang, X.: Colmap-free 3d gaussian splatting. In: CVPR (2024)
12. Hong, S., Jung, J., Shin, H., Yang, J., Kim, S., Luo, C.: Coponerf: Unifying correspondence, pose and nerf for pose-free novel view synthesis from stereo pairs. In: CVPR (2024)
13. Huang, B., Yu, Z., Chen, A., Geiger, A., Gao, S.: 2d gaussian splatting for geometrically accurate radiance fields. In: ACM SIGGRAPH 2024 conference papers (2024)
14. Huang, R., Mikolajczyk, K.: No pose at all: Self-supervised pose-free 3d gaussian splatting from sparse views. In: ICCV (2025)
15. Jiang, H., Tan, H., Wang, P., Jin, H., Zhao, Y., Bi, S., Zhang, K., Luan, F., Sunkavalli, K., Huang, Q., Pavlakos, G.: Rayzer: A self-supervised large view synthesis model. In: ICCV (2025)
16. Jiang, L., Mao, Y., Xu, L., Lu, T., Ren, K., Jin, Y., Xu, X., Yu, M., Pang, J., Zhao, F., et al.: Anysplat: Feed-forward 3d gaussian splatting from unconstrained views. ACM Transactions on Graphics (TOG) **44**(6), 1–16 (2025)
17. Kang, G., Yoo, J., Park, J., Nam, S., Im, H., Shin, S., Kim, S., Park, E.: Selfsplat: Pose-free and 3d prior-free generalizable 3d gaussian splatting. In: CVPR (2025)
18. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. (2023)
19. Leroy, V., Cabon, Y., Revaud, J.: Grounding image matching in 3d with mast3r. In: ECCV (2024)
20. Lin, C.Y., Sun, C., Yang, F.E., Chen, M.H., Lin, Y.Y., Liu, Y.L.: Longsplat: Robust unposed 3d gaussian splatting for casual long videos. In: ICCV (2025)

21. Lin, H., Chen, S., Liew, J., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. arXiv preprint arXiv:2511.10647 (2025)
22. Ling, L., Sheng, Y., Tu, Z., Zhao, W., Xin, C., Wan, K., Yu, L., Guo, Q., Yu, Z., Lu, Y., et al.: Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In: CVPR (2024)
23. Liu, A., Tucker, R., Jampani, V., Makadia, A., Snavely, N., Kanazawa, A.: Infinite nature: Perpetual view generation of natural scenes from a single image. In: ICCV (2021)
24. Liu, Y., Fan, K., Yu, W., Li, C., Lu, H., Yuan, Y.: Monosplat: Generalizable 3d gaussian splatting from monocular depth foundation models. In: CVPR (2025)
25. Liu, Y., Min, Z., Wang, Z., Wu, J., Wang, T., Yuan, Y., Luo, Y., Guo, C.: Worldmirror: Universal 3d world reconstruction with any-prior prompting. arXiv preprint arXiv:2510.10726 (2025)
26. Liu, Y.C., Höllein, L., Nießner, M., Dai, A.: Quicksplat: Fast 3d surface reconstruction via learned gaussian initialization. In: ICCV (2025)
27. Luu, T.M., Lee, Y., Lee, D., Kim, S., Kim, M.J., Yoo, C.D.: Enhancing rating-based reinforcement learning to effectively leverage feedback from large vision-language models. In: ICML (2025)
28. MacGlashan, J., Ho, M.K., Loftin, R., Peng, B., Wang, G., Roberts, D.L., Taylor, M.E., Littman, M.L.: Interactive learning from policy-dependent human feedback. In: ICML (2017)
29. 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: ECCV (2020)
30. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM transactions on graphics (TOG) **41**(4), 1–15 (2022)
31. Park, J., Bui, M.Q.V., Bello, J.L.G., Moon, J., Oh, J., Kim, M.: Splinegs: Robust motion-adaptive spline for real-time dynamic 3d gaussians from monocular video. In: CVPR (2025)
32. Park, J., Bui, M.Q.V., Bello, J.L.G., Moon, J., Oh, J., Kim, M.: Ecosplat: Efficiency-controllable feed-forward 3d gaussian splatting from multi-view images. In: CVPR (2026)
33. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: CVPR (2016)
34. Smart, B., Zheng, C., Laina, I., Prisacariu, V.A.: Splatt3r: Zero-shot gaussian splatting from uncalibrated image pairs. arXiv preprint arXiv:2408.13912 (2024)
35. Snavely, N., Seitz, S.M., Szeliski, R.: Photo tourism: exploring photo collections in 3d. ACM Trans. Graph. (2006)
36. Suhail, M., Esteves, C., Sigal, L., Makadia, A.: Generalizable patch-based neural rendering. In: ECCV (2022)
37. Szymanowicz, S., Rupprecht, C., Vedaldi, A.: Splatter image: Ultra-fast single-view 3d reconstruction. In: CVPR (2024)
38. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: CVPR (2025)
39. Wang, Q., Wang, Z., Genova, K., Srinivasan, P., Zhou, H., Barron, J.T., Martin-Brualla, R., Snavely, N., Funkhouser, T.: Ibrnet: Learning multi-view image-based rendering. In: CVPR (2021)
40. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: CVPR (2024)

41. Wang, W., Chen, D.Y., Zhang, Z., Shi, D., Liu, A., Zhuang, B.: Zpressor: Bottleneck-aware compression for scalable feed-forward 3dgs. arXiv preprint arXiv:2505.23734 (2025)
42. Wang, Y., Zhou, J., Zhu, H., Chang, W., Zhou, Y., Li, Z., Chen, J., Pang, J., Shen, C., He, T.: π^3: Permutation-equivariant visual geometry learning. arXiv preprint arXiv:2507.13347 (2025)
43. White, D., Wu, M., Novoseller, E., Lawhern, V.J., Waytowich, N., Cao, Y.: Rating-based reinforcement learning. In: AAAI (2024)
44. Wu, G., Yi, T., Fang, J., Xie, L., Zhang, X., Wei, W., Liu, W., Tian, Q., Wang, X.: 4d gaussian splatting for real-time dynamic scene rendering. In: CVPR (2024)
45. Xu, H., Peng, S., Wang, F., Blum, H., Barath, D., Geiger, A., Pollefeys, M.: Depthsplat: Connecting gaussian splatting and depth. In: CVPR (2025)
46. Ye, B., Chen, B., Xu, H., Barath, D., Pollefeys, M.: Yonosplat: You only need one model for feedforward 3d gaussian splatting. In: ICLR (2026)
47. Ye, B., Liu, S., Xu, H., Li, X., Pollefeys, M., Yang, M.H., Peng, S.: No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images. arXiv preprint arXiv:2410.24207 (2024)
48. Yousaf, A., Mitra, A., Agbaje, P., Anjum, A., Olufowobi, H.: Daps-agf: Depth-aware perceptual similarity with adaptive gradient filtering for enhanced outdoor scene reconstruction. In: ICCV Workshop (2025)
49. Yu, A., Li, R., Tancik, M., Li, H., Ng, R., Kanazawa, A.: Plenoctrees for real-time rendering of neural radiance fields. In: ICCV (2021)
50. Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelnerf: Neural radiance fields from one or few images. In: CVPR (2021)
51. Yu, Z., Chen, A., Huang, B., Sattler, T., Geiger, A.: Mip-splatting: Alias-free 3d gaussian splatting. In: CVPR (2024)
52. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
53. Zhang, S., Wang, J., Xu, Y., Xue, N., Rupprecht, C., Zhou, X., Shen, Y., Wetzstein, G.: Flare: Feed-forward geometry, appearance and camera estimation from uncalibrated sparse views. In: CVPR (2025)
54. Zhou, T., Tucker, R., Flynn, J., Fyffe, G., Snavely, N.: Stereo magnification: Learning view synthesis using multiplane images. In: SIGGRAPH (2018)
55. Ziwen, C., Tan, H., Zhang, K., Bi, S., Luan, F., Hong, Y., Fuxin, L., Xu, Z.: Long-lrm: Long-sequence large reconstruction model for wide-coverage gaussian splats. In: ICCV (2025)