



Integrating Generative Artificial Intelligence for Robust Multi-View Registration and 3D Reconstruction of Electric Vehicles

Xian Zhang^{1,2*}, Hailin Wang², Zhenyang Zhuo², Salman Dbaiyy^{1,3}

¹ Waikato International College, Jinhua University of Vocational Technology, Jinhua 321000, China

² Intelligent Manufacturing College, Jinhua University of Vocational Technology, Jinhua 321000, China

³ Information Technology, Waikato Institute of Technology, Hamilton 32408, New Zealand

Corresponding Author Email: jhzx@jhc.edu.cn

Copyright: ©2026 The authors. This article is published by IETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/ts.430402>

ABSTRACT

Received: 8 April 2026

Revised: 30 June 2026

Accepted: 27 July 2026

Available online: 31 August 2026

Keywords:

generative artificial intelligence, diffusion models, 3D gaussian splatting, multi-view image registration, 3D reconstruction, electric vehicles

High-precision 3D reconstruction of electric vehicles (EVs) provides fundamental support for autonomous driving simulation and digital twin-based operation and maintenance. However, the widespread distribution of strong reflections on metallic surfaces and low-texture regions poses dual challenges of feature matching degradation and geometric information loss in multi-view image registration and 3D reconstruction. To address these challenges, this paper proposes an end-to-end 3D reconstruction method that integrates generative artificial intelligence (AI). In the image registration stage, we model the matching matrix estimation as a denoising diffusion process on the doubly stochastic matrix subspace, utilizing epipolar geometry priors to guide multi-step reverse sampling for iterative refinement of correspondences, thereby overcoming the local optimum issue of single-step prediction in low-texture regions. In the feature learning stage, we condition a diffusion model on source-view images and depth maps to synthesize geometrically consistent target-view virtual aligned images. Furthermore, a cross-view hard example contrastive learning strategy is introduced to enhance the discriminative power of matching features, effectively suppressing interference from specular reflections and repetitive textures. In the 3D reconstruction stage, leveraging 3D Gaussian Splatting as the geometric representation, we construct a three-stage progressive framework consisting of geometric initialization, generative inpainting, and semantic-consistent refinement. Specifically, a video diffusion model is employed to synthesize novel views along a surrounding trajectory to fill observation blind spots, while vehicle part semantic priors and longitudinal symmetry constraints are introduced for fine-grained correction of the reconstruction results. Extensive experiments on nuScenes, Waymo, 3DRealCar, and a self-collected EV multi-view dataset demonstrate that our method achieves state-of-the-art performance in key metrics such as registration recall, rendered image quality, depth estimation accuracy, and cross-view generalization capability. Ablation studies further verify the effective synergistic gains of each core module.

1. INTRODUCTION

High-precision 3D digital models of electric vehicles (EVs) play an irreplaceable fundamental role in engineering scenarios such as autonomous driving simulation testing, digital twin operation and maintenance, crash safety analysis, and aerodynamic optimization [1]. With the increasing popularity of on-board multi-view camera systems, 3D reconstruction of vehicles using surround-view images has become a research hotspot in the interdisciplinary field of computer vision and image processing [2]. However, 3D reconstruction of multi-view images of EVs faces a series of unique image processing challenges [3]. Large areas of metal cladding on vehicle body surfaces lead to widespread distribution of specular reflections and highlight regions [4], severely interfering with the stable extraction and matching of feature points [5]; the highly repetitive textures on car paint

and glass surfaces significantly reduce the discriminative power of feature descriptors [6]; on-board cameras have sparse and heterogeneous viewing angle distributions, with limited overlapping areas between different views [7]. Traditional multi-view stereo vision methods are highly prone to generating geometric holes and texture fractures in areas with insufficient view coverage [8]. The essence of the above problems lies in: under conditions of insufficient image observation information, how to recover a complete and geometrically accurate 3D structure from sparse, high-noise, and low-texture multi-view images. This is a typical ill-posed inverse problem that urgently calls for a breakthrough in new methodologies.

In recent years, generative artificial intelligence (AI) technology represented by diffusion models has provided a brand-new solution paradigm for the above dilemmas [9]. With their powerful distribution modeling and iterative

refinement capabilities, diffusion models have demonstrated outstanding performance in tasks such as image generation, image inpainting, and cross-modal synthesis [10]. At the same time, 3D Gaussian Splatting, as an emerging explicit 3D scene representation method, is gradually replacing the traditional Neural Radiance Field (NeRF) to become the mainstream technical framework in the field of 3D reconstruction, owing to its efficient rendering speed and excellent image quality [11]. However, existing methods still have significant shortcomings when applied to the specific application scenario of EVs [12]. At the image registration level [13], traditional methods based on handcrafted or deep features rely on single-step prediction heads to directly estimate matching relationships from the feature space. In metal highlight regions and low-texture surfaces, the discriminative power of feature descriptors severely degrades, and the single-step prediction mechanism is prone to falling into local optima, making it difficult to obtain globally optimal matching results [14]. Although the latest works such as Diff-Reg v2 attempt to introduce diffusion models into the matching matrix space, they are mainly oriented toward registration tasks in generic scenes and are not adapted to the unique texture distribution and geometric structure of EVs. At the cross-view consistency maintenance level, existing methods rely on a limited number of real-view images for feature learning [15]. The view coverage and lighting variations of the training data are extremely limited, making it difficult to cover the diverse imaging conditions faced by EVs in actual acquisition scenarios [16]. Although some works utilize generative models for data augmentation [17], the generated virtual views often lack strict constraints of geometric consistency, and the improvement in cross-view feature discriminative power is limited [18]. At the 3D reconstruction level, both traditional multi-view stereo vision and NeRF or 3D Gaussian Splatting methods face severe geometric degradation problems when views are sparse [19]. Although existing diffusion-based reconstruction methods have demonstrated the potential of generative priors to improve reconstruction quality [20], most of them are generic frameworks [21] that fail to fully exploit the semantic priors of EVs, such as body symmetry [22] and part geometric constraints [23]. Moreover, there is a lack of deep collaborative optimization mechanisms between generative inpainting and geometric reconstruction [24, 25].

To address the above shortcomings, this paper proposes a multi-view image registration and 3D visual reconstruction method for EVs that integrates generative AI. In the image registration stage, the correspondence estimation between multi-view images is modeled as a denoising diffusion process in the matching matrix space. By performing multi-step reverse sampling in the doubly stochastic matrix subspace, the matching matrix is progressively refined to the optimal solution, thereby effectively overcoming the local optimum dilemma of single-step prediction in low-texture and high-reflection areas. In the feature learning stage, a diffusion model is conditioned on source-view images and depth maps to synthesize geometrically consistent target-view virtual aligned images, and a cross-view contrastive learning loss is used to enhance the discriminative power of matching features, significantly improving the robustness of feature matching in metal reflection and repetitive texture regions. In the 3D reconstruction stage, using 3D Gaussian Splatting as the geometric representation carrier, a three-stage progressive framework of geometric initialization, generative inpainting, and semantic-consistent refinement is constructed. A point

cloud-guided video diffusion model is used to synthesize geometrically consistent novel views to fill observation blind spots, and EV part semantic priors and longitudinal symmetry constraints are introduced to refine the reconstruction results. The three core contributions are organically connected with matching, enhancement, and reconstruction as the technical main line, forming an end-to-end 3D reconstruction solution for sparse multi-view images of EVs.

Section 2 of this paper systematically elaborates on the overall framework of the proposed method and the technical details of each module. Section 3 introduces the experimental setup, datasets, comparison methods, and evaluation metrics, and verifies the effectiveness of the method through multiple sets of experiments. Section 4 conducts an in-depth discussion on the core characteristics and limitations of the method. Section 5 summarizes the full text and looks forward to future research directions.

2. METHODOLOGY

2.1 Overall framework overview

The method proposed in this paper consists of three progressively coupled core modules. The diffusion matching matrix refinement module is responsible for establishing high-precision cross-view image correspondences. This module models matching estimation as a denoising diffusion process in the matching matrix space, utilizing epipolar geometry priors to guide multi-step reverse sampling, thereby outputting reliable feature correspondences and initial camera poses. The generative cross-view consistency enhancement module takes the matching relationships and depth estimates output by the previous module as geometric anchors, conditions a diffusion model on source-view images and depth maps to synthesize geometrically aligned virtual target-view images, and leverages contrastive learning to strengthen the cross-view discriminative power of matching features. The geometry-semantics collaborative 3D reconstruction module uses 3D Gaussian Splatting as the geometric representation carrier, receives the enhanced multi-view images and refined poses as input to complete geometric initialization, then generates novel views through a point cloud-guided video diffusion model to fill observation blind spots, and finally performs fine-grained correction on the reconstruction results using EV part semantic priors and symmetry constraints. The three modules take matching, enhancement, and reconstruction as the technical main line. The output of each preceding module is passed as differentiable input to the subsequent module, forming an end-to-end trainable technical closed loop.

The overall training adopts a strategy combining staged pre-training and joint fine-tuning. The diffusion matching matrix refinement module and the generative cross-view consistency enhancement module are first independently pre-trained with registration loss. After obtaining stable matching and feature enhancement capabilities, they are jointly optimized end-to-end with the 3D reconstruction module using rendering loss and geometric loss as supervision signals. This strategy not only ensures convergence stability of each module at the early training stage, but also achieves deep synergy between modules through subsequent joint fine-tuning, enabling generative enhancement and geometric reconstruction to mutually promote rather than operate in isolation.

2.2 Diffusion matching matrix refinement module

Image registration is the fundamental step of 3D reconstruction, and its accuracy directly determines the quality of subsequent geometric initialization. In multi-view images of EVs, metal highlights and low-texture regions cause severe degradation of the discriminative power of feature descriptors, making it difficult for traditional single-step prediction

methods to directly estimate reliable matching relationships from noisy feature spaces. To this end, this paper models this problem as a denoising diffusion process in the matching matrix space, achieving coarse-to-fine correspondence estimation through multi-step iterative refinement. Figure 1 shows the structural schematic of the diffusion matching matrix refinement module.

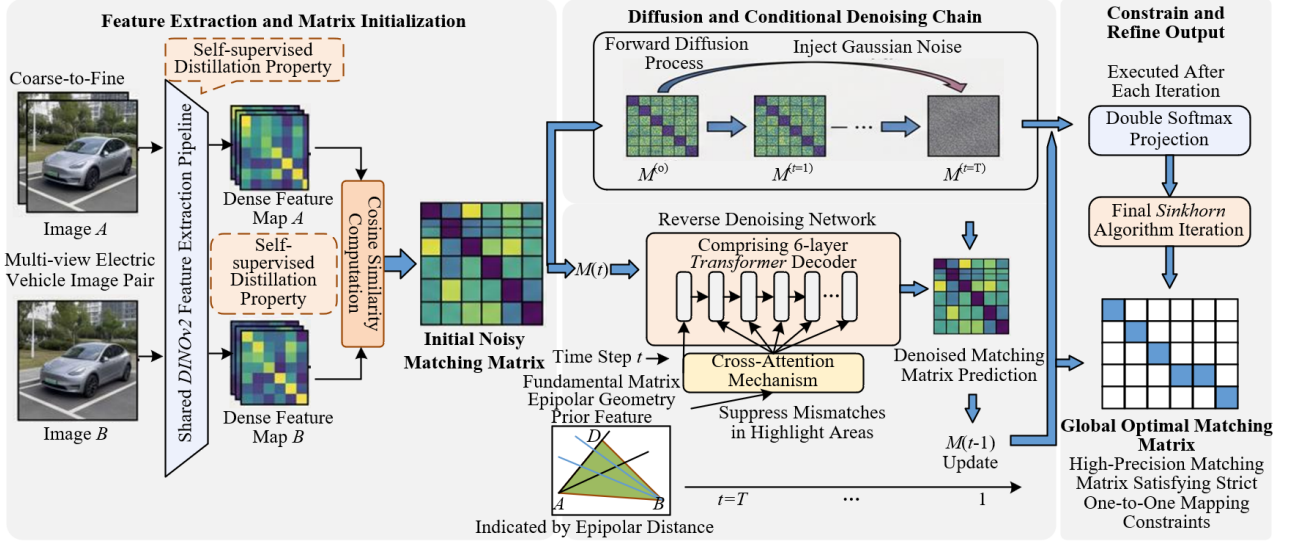


Figure 1. Structural schematic of the diffusion matching matrix refinement module

For the input multi-view image pair $I_A, I_B \in \mathbb{R}^{H \times W \times 3}$, DINOv2 ViT-L/14 is used as a shared backbone network to extract dense visual features. This model is trained via self-supervised distillation and can maintain strong semantic discriminative power in low-texture and highlight regions. After patch processing, the images yield $N_p = H/14 \times W/14$ patches. Two sets of feature maps $F_A \in \mathbb{R}^{N_p \times D}$ and $F_B \in \mathbb{R}^{N_p \times D}$ are extracted, with feature dimension $D=1024$. Based on the pairwise cosine similarity between feature maps, an initial matching matrix $M^{(0)} \in \mathbb{R}^{N_p \times N_p}$ is constructed, with its elements defined as:

$$m_{ij}^{(0)} = \frac{(F_A)_i \cdot (F_B)_j}{\|(F_A)_i\|_2 \cdot \|(F_B)_j\|_2} \quad (1)$$

To ensure that the matching relationship satisfies the one-to-one mapping constraint, double softmax projection is used to regularize $M^{(0)}$ to the doubly stochastic matrix subspace $B = \{M \in \mathbb{R}^{N_p \times N_p} : M \mathbf{1} = \mathbf{1}, \mathbf{1}^T M = \mathbf{1}^T\}$. The projection is performed as the product of row-wise and column-wise Softmax:

$$\tilde{m}_{ij}^{(0)} = \text{Softmax}_j(M^{(0)})_{ij} \cdot \text{Softmax}_i(M^{(0)})_{ij} \quad (2)$$

This projection makes the row and column sums of the matching matrix equal to 1, which geometrically guarantees that each image patch corresponds to only one matching patch in the other view, excluding degenerate many-to-many correspondences.

The refinement process of the matching matrix is defined as a denoising diffusion model on the doubly stochastic matrix space. The forward diffusion stage uses $T = 100$ steps of Gaussian noise injection:

$$M^{(t)} = \sqrt{\bar{\alpha}_t} \tilde{M}^{(0)} + \sqrt{1 - \bar{\alpha}_t} \epsilon_t, \epsilon_t \sim N(0, I), t = 1, \dots, T \quad (3)$$

where, $\bar{\alpha}_t = \prod_{s=1}^t (1 - \beta_s)$, and the noise schedule adopts a cosine schedule:

$$\beta_t = \frac{1 - \cos(\pi t/T)}{2} \cdot \beta_{\max} + \frac{1 + \cos(\pi t/T)}{2} \cdot \beta_{\min} \quad (4)$$

where, $\beta_{\min} = 10^{-4}$ and $\beta_{\max} = 0.02$. This scheduling strategy maintains a relatively gentle noise growth in the early and middle stages, allowing the semantic structure of the matching matrix to be preserved from destruction, retaining more effective information for reverse denoising. The reverse denoising stage takes epipolar geometry priors as conditions, and gradually recovers the optimal matching matrix through the denoising network ϵ_θ . The network adopts a 6-layer Transformer decoder architecture, with each layer containing 8 multi-head attention heads and a 256-dimensional hidden layer, with a total parameter count of approximately 12M. The network takes the noisy matching matrix $M^{(t)}$, the diffusion time step t , and the geometric condition encoding C as input, and predicts the added noise:

$$\hat{\epsilon}_t = \epsilon_\theta(M^{(t)}, t, C) \quad (5)$$

The geometric condition encoding C is obtained by mapping the fundamental matrix F_{AB} between the two views through a three-layer Multilayer Perceptron (MLP). For each candidate matching point pair (p_i, q_j) , its epipolar distance:

$$d_{ij} = |p_i^T F_{AB} q_j| / \sqrt{(F_{AB} p_i)_1^2 + (F_{AB} p_i)_2^2} \quad (6)$$

The epipolar distance d_{ij} is then encoded as a geometric prior feature and injected into the Transformer decoder through a cross-attention mechanism. This design enables the denoising network to continuously obtain epipolar geometric constraints during the iterative refinement process, effectively suppressing the risk of mismatching caused by falsely high feature similarity in highlight regions.

Reverse sampling follows the standard denoising diffusion iterative update:

$$M^{(t-1)} = \frac{1}{\sqrt{\alpha_t}} \left(M^{(t)} - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_{\theta}(M^{(t)}, t, C) \right) + \sigma_t z \quad (7)$$

where, $\alpha_t = 1 - \beta_t$, $\sigma_t^2 = 1 - \alpha_t - 1/\alpha_t \beta_t$, $z \sim N(0, I)$ when $t > 1$, and when $t = 1$, $z = 0$ to achieve deterministic sampling. After each denoising step, double softmax projection is applied to map $M^{(t-1)}$ back to the doubly stochastic matrix subspace, ensuring that the one-to-one mapping constraint always holds during the iteration. This iterative refinement mechanism enables the matching matrix to gradually converge from a noisy state to the global optimum, effectively overcoming the inherent defect of single-step prediction that is prone to falling into local optima in weak-texture and highlight regions.

The training loss of Module 1 is a weighted combination of diffusion denoising loss and matching classification loss:

$$L_{match} = E_{t, \epsilon} [\| \epsilon - \epsilon_{\theta}(M^{(t)}, t, C) \|_2^2] + \lambda_{cls} \cdot CE(M^{(0 \rightarrow T)}, M^{gt}) \quad (8)$$

where, $CE(\cdot)$ is the cross-entropy loss, used to supervise the difference between the final predicted matching matrix and the ground truth matching label M^{gt} , and λ_{cls} is set to 0.1. The

ground truth matching labels are generated from camera poses provided by the dataset through epipolar geometry constraints and depth consistency verification. During inference, the Denoising Diffusion Implicit Model (DDIM) deterministic sampling strategy is adopted to accelerate inference, and only $T' = 10$ sampling steps are needed to achieve accuracy comparable to $T = 100$. The final output optimal matching matrix M^* is further strengthened by the Sinkhorn algorithm with 20 iterations to reinforce the doubly stochastic constraint, providing high-precision camera poses and dense correspondences for the subsequent 3D reconstruction module.

2.3 Generative cross-view consistency enhancement module

The high-precision correspondences output by the diffusion matching matrix refinement module provide reliable geometric anchors for subsequent feature learning. However, the matching accuracy itself is limited by the discriminative power of feature descriptors in metal reflection and repetitive texture regions. Relying solely on original observed images to train the feature extraction network, the view coverage and lighting variations of the training data are extremely limited, making it difficult to cover the diverse imaging conditions faced by EVs in actual acquisition scenarios. To this end, this paper designs a conditional diffusion generation module G_{ϕ} , which synthesizes geometrically aligned target-view virtual images conditioned on source-view images and their depth maps, thereby expanding the training distribution with virtual views and enhancing the discriminative power of features in difficult regions. Figure 2 illustrates the architecture of the generative cross-view consistency enhancement module.

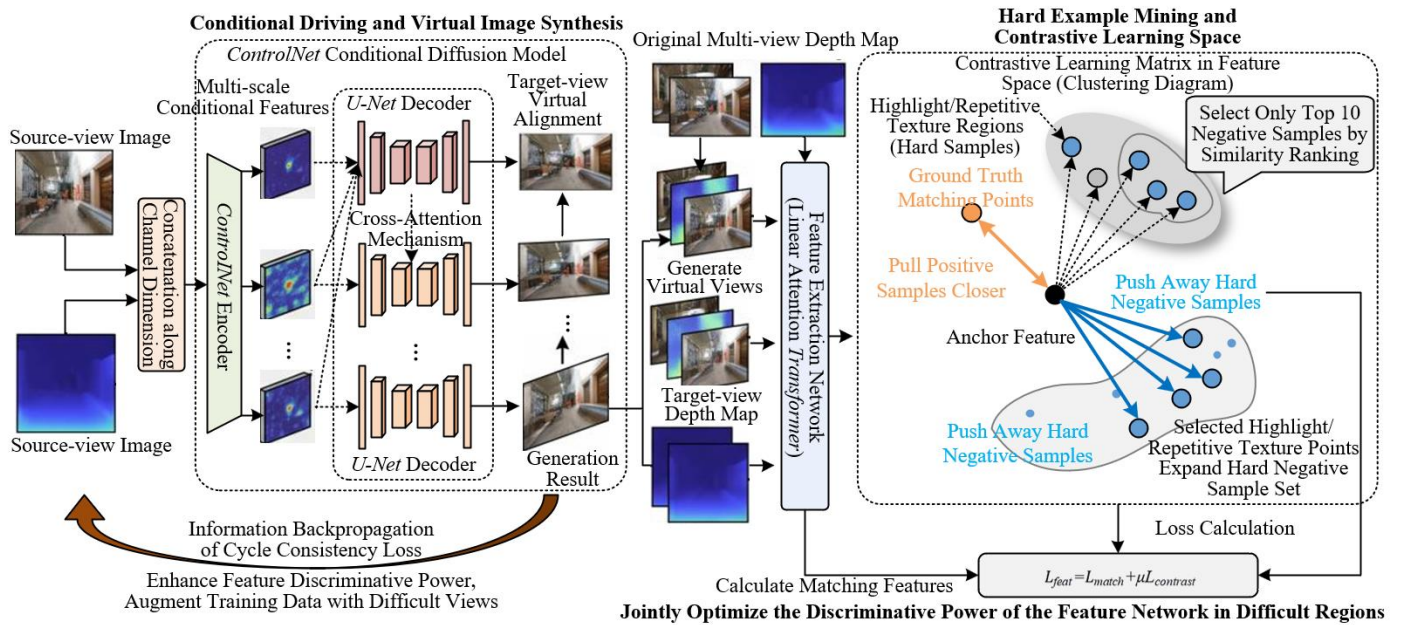


Figure 2. Architecture diagram of the generative cross-view consistency enhancement module

This module takes the source-view image I_s and its depth map D_s as conditional inputs to synthesize the virtual aligned image $\tilde{I}_{s \rightarrow t}$ in the target view. The depth map D_s is obtained by combining the matching matrix output from Module 1 with multi-view triangulation, and is smoothed by median filtering and bilateral filtering to eliminate depth noise introduced by triangulation. The generation architecture is based on the

ControlNet conditional diffusion model. I_s and D_s are concatenated along the channel dimension and input into the ControlNet encoder to extract multi-scale conditional features $\{h_i\}_{i=1}^L$, which are injected into each layer of the U-Net decoder. Specifically, the conditional features interact with the intermediate features of the U-Net through a cross-attention mechanism, calculated as:

$$\text{Attention}(Q,K,V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, K = W_K h_l, V = W_V h_l \quad (9)$$

where, Q comes from the U-Net decoder features, K and V are obtained by mapping the conditional features through learnable projection matrices W_K and W_V , and d_k is the dimension of the attention head. This design enables the diffusion model to continuously perceive the geometric structure of the source view during the generation process, ensuring that the virtual image maintains spatial correspondence with the source view in content, rather than freely generating content unrelated to the source view.

The supervision signal for the generated image is jointly constituted by three losses. The texture consistency loss ensures that the generated image $\tilde{I}_{s \rightarrow t}$ and the real target image I_t are consistent in visual appearance:

$$L_{\text{texture}} = \|\tilde{I}_{s \rightarrow t} - I_t\|_1 + \lambda_{\text{ssim}} \cdot (1 - \text{SSIM}(\tilde{I}_{s \rightarrow t}, I_t)) + \lambda_{\text{perc}} \cdot \|\Phi(\tilde{I}_{s \rightarrow t}) - \Phi(I_t)\|_2^2 \quad (10)$$

where, $\Phi(\cdot)$ is the Visual Geometry Group (VGG)-16 perceptual loss extraction network, $\lambda_{\text{ssim}} = 0.2$, and $\lambda_{\text{perc}} = 0.5$. The geometric structure preservation loss constrains the depth map of the generated image to be consistent with the target-view depth map $D_{s \rightarrow t}$ obtained by projecting the source view:

$$L_{\text{geom}} = \|\text{Depth}(\tilde{I}_{s \rightarrow t}) - D_{s \rightarrow t}\|_1 + \|\nabla \text{Depth}(\tilde{I}_{s \rightarrow t}) - \nabla D_{s \rightarrow t}\|_1 \quad (11)$$

where, $\text{Depth}(\cdot)$ is the depth map extracted from the generated image, and ∇ is the spatial gradient operator. By simultaneously constraining the depth values and their first-order gradients, this loss maintains geometric consistency at both the pixel-level accuracy and structural edge levels. The cross-view cycle consistency loss further strengthens the geometric closed loop of bidirectional generation. The generated virtual image $\tilde{I}_{s \rightarrow t}$, along with its depth map and target-view camera parameters, is input into the generator, requiring the cyclic reconstruction result to be consistent with the source-view image:

$$L_{\text{cycle}} = \|G_\theta(\tilde{I}_{s \rightarrow t}, D_{s \rightarrow t}, K_t, K_s, R_{t \rightarrow s}, t_{t \rightarrow s}) - I_s\|_1 \quad (12)$$

where, K_s are K_t the camera intrinsic matrices of the source and target views, respectively, $R_{t \rightarrow s}$ and $t_{t \rightarrow s}$ are the rotation and translation transformations from the target to the source view. The three losses are weighted and summed:

$$L_{\text{gen}} = L_{\text{texture}} + \lambda_g L_{\text{geom}} + \lambda_c L_{\text{cycle}} \quad (13)$$

where, $\lambda_g = 0.8$, $\lambda_c = 0.3$. This joint supervision mechanism achieves a balance between texture realism and geometric structure accuracy in the generated images, avoiding the sacrifice of geometric fidelity due to overfitting the texture loss.

A large number of generated virtual aligned image pairs are combined with the original multi-view images to construct an expanded training set. The matching feature extraction network adopts a linear attention Transformer architecture. On this basis, a cross-view hard example contrastive learning loss is introduced, forcing matching point pairs in the feature space to cluster tightly and non-matching point pairs to repel each other. For the anchor feature $f(p_i)$, the positive sample is the true matching point $f(q_i^*)$, and the negative samples are all

non-matching points within the same batch. A hard example mining strategy is adopted, selecting only the top $K_{\text{hard}}=10$ negative samples with the highest similarity to participate in the loss calculation:

$$L_{\text{contrast}} = -\frac{1}{B} \sum_{b=1}^B \log \frac{\exp(\text{sim}(f_b^A, f_b^{B+})/\tau)}{\exp(\text{sim}(f_b^A, f_b^{B+})/\tau) + \sum_{k=1}^{K_{\text{hard}}} \exp(\text{sim}(f_b^A, f_k^{B-})/\tau)} \quad (14)$$

where, B is the batch size, $\tau = 0.1$ is the temperature coefficient, and $\text{sim}(\cdot, \cdot)$ is the cosine similarity. This loss is jointly optimized with the matching loss of Module 1 for the feature extraction network:

$$L_{\text{feat}} = L_{\text{match}} + \mu L_{\text{contrast}} \quad (15)$$

where, $\mu = 0.5$. The virtual aligned images play a dual role in this process: on one hand, they expand the coverage of training views, exposing the feature network to richer view variations during the training phase; on the other hand, through the construction of positive and negative samples in contrastive learning, they force the network to learn feature representations that are insensitive to highlights and repetitive textures, thereby enhancing matching robustness at its source.

2.4 Geometry-semantics collaborative 3D reconstruction module

After obtaining the enhanced multi-view images and refined camera poses, the 3D reconstruction module constructs the scene representation using the aforementioned correspondences as geometric priors. This paper adopts 3D Gaussian Splatting as the fundamental scene parameterization, which represents the target scene as a set of 3D Gaussian ellipsoids. Each Gaussian is defined by a position vector $\mu_k \in \mathbb{R}^3$, a covariance matrix $\Sigma_k \in \mathbb{R}^{3 \times 3}$, an opacity $\alpha_k \in [0, 1]$, and spherical harmonic coefficients $\{c_{k,l} \in \mathbb{R}^3\}_{l=0}^{L_{sh}}$ representing view-dependent colors. In this paper, the spherical harmonic degree is set to $L_{sh} = 3$. Using the high-precision matching correspondences output by Module 1, the essential matrix is estimated via the five-point method and triangulated to generate an initial sparse point cloud, which is then densified through multi-view stereo matching, obtaining approximately 5×10^4 initial point clouds $\{\mu^{(0)}_k\}_{k=1}^{K_{k=1}}$. The initial covariance of each Gaussian is set to the identity matrix scaled by 0.01, the opacity is initialized to 0.5, and the spherical harmonic coefficients are initialized to the mean of the corresponding view image colors. The 3D Gaussians are projected onto the 2D image plane via splatting rasterization, sorted by depth, and then alpha-composited pixel by pixel. The initial optimization objective between the rendered image $\hat{I}$ and the ground truth image I is:

$$L_{\text{init}} = (1 - \lambda_{\text{SSIM}}) \|\hat{I} - I\|_1 + \lambda_{\text{SSIM}} (1 - \text{SSIM}(\hat{I}, I)) + \lambda_{\text{reg}} \sum_k \|\Sigma_k\|_F^2 \quad (16)$$

where, λ_{SSIM} controls the weight of the structural similarity loss, and $\lambda_{\text{reg}} = 0.001$ is the covariance regularization coefficient, used to suppress excessive flattening of the Gaussian ellipsoids. The optimization uses the Adam solver, with a position learning rate of 1.6×10^{-3} , covariance of 5×10^{-4} , opacity of 5×10^{-2} , and spherical harmonics of

2.5×10^{-3} . Adaptive density control is triggered every 100 iterations, splitting or cloning Gaussians with excessive gradient accumulation, and pruning Gaussians with opacity below 10^{-3} , thereby controlling Gaussian redundancy while

preserving geometric details. Figure 3 shows the geometry-semantics collaborative 3D reconstruction process and correction schematic.

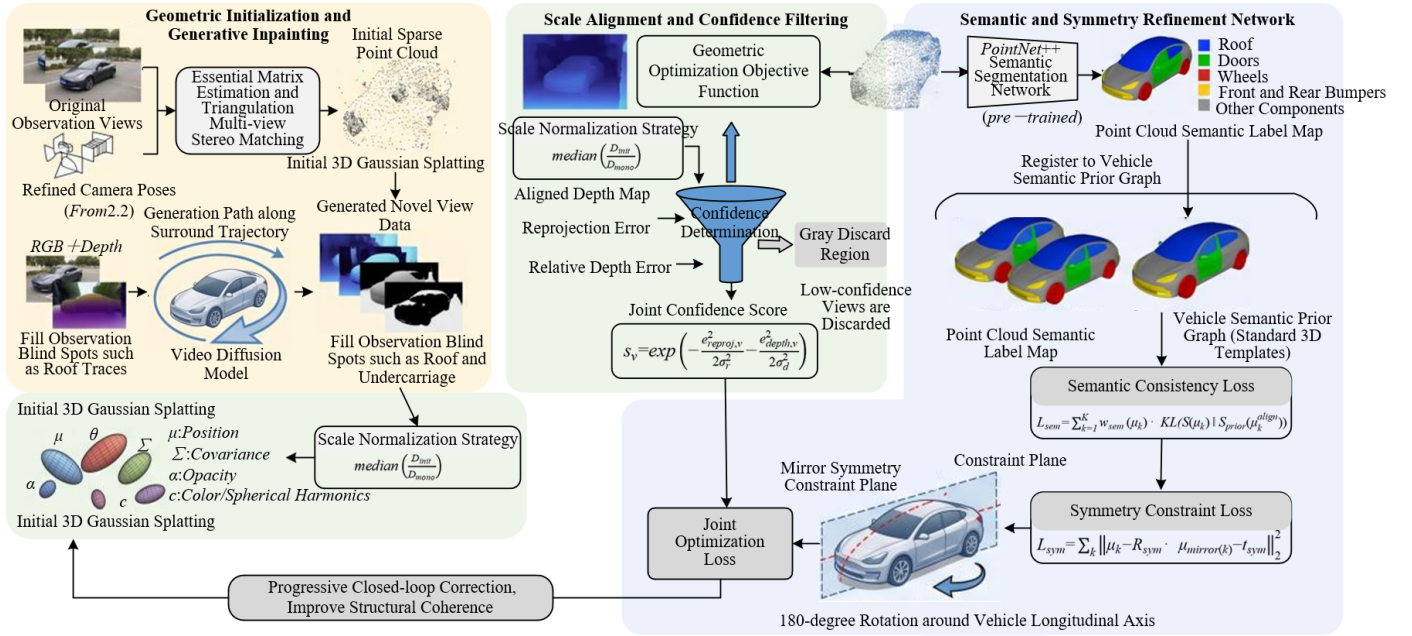


Figure 3. Geometry-semantics collaborative 3D reconstruction process and correction schematic

In multi-view images of EVs, regions such as the roof, undercarriage, and backlit surfaces suffer from severe geometric holes due to insufficient view coverage, making it difficult to reconstruct complete structures relying solely on original observations. To this end, this paper introduces a generative geometric inpainting process guided by a video diffusion model. Conditioned on the sparse-view image sequences rendered by the initial 3D Gaussian Splatting, interpolation frames are generated along the surround trajectory to fill view gaps. Specifically, given a source view set V_{src} and a target missing view set V_{tgt} , the video diffusion model takes the RGB images and corresponding depth renderings of the source views as conditions, and simultaneously refers to adjacent front and back views through spatio-temporal cross-attention to maintain trajectory continuity, generating the RGB image $\tilde{I}_v$ and depth map $\tilde{D}_v$ in the target view. Since the depth predicted by the diffusion model has an uncertain offset relative to the scale of the 3D Gaussian Splatting, this paper adopts a scale normalization strategy for alignment:

$$D_{align} = \text{median}\left(\frac{D_{init}}{D_{mono}}\right) \cdot D_{mono} \quad (17)$$

where, D_{init} is the projection depth map of the 3D Gaussian Splatting in the target view, and D_{mono} is the monocular depth predicted by the video diffusion model. The median of the depth ratio is taken as the unified scale factor to avoid interference from local outliers on the global scale estimation.

The geometric reliability of the generated views varies significantly; directly incorporating all of them into the optimization may introduce geometric drift. This paper constructs a joint confidence score to evaluate the credibility of each generated view:

$$s_v = \exp\left(-\frac{e_{reproj,v}^2}{2\sigma_r^2} - \frac{e_{depth,v}^2}{2\sigma_d^2}\right) \quad (18)$$

where, $e_{reproj,v}$ is the feature reprojection mean squared error between the generated view and adjacent source views, $e_{depth,v} = \|\tilde{D}_v - D_{align}\|_2 / \|D_{align}\|_2$ is the relative depth error, and pixels $\sigma_r = 2.0$ and $\sigma_d = 0.1$ are the bandwidth parameters for the two types of errors, respectively. Only high-confidence generated views with $s_v > 0.6$ are retained for subsequent optimization, while the rest are discarded. The filtered generated views are merged with the original views to form an expanded training set $V_{all} = V_{src} \cup V_{gen}$, and each view is assigned a confidence weight w_v : 1.0 for original views and s_v for generated views. The final optimization objective function is:

$$L_{final} = \sum_{v \in V_{all}} w_v [(1 - \lambda_{SSIM}) \|\hat{I}_v - I_v\|_1 + \lambda_{SSIM} (1 - SSIM(\hat{I}_v, I_v)) + \gamma \|\hat{D}_v - D_v\|_1] \quad (19)$$

where, $\hat{D}_v$ is the depth map rendered by 3D Gaussian Splatting, D_v is the ground truth or generated depth map, and $\gamma = 0.05$ is the depth supervision weight. This loss imposes strong supervision on original views to ensure accuracy in already reconstructed regions, and applies weak supervision to generated views based on confidence, thereby filling geometric blind spots without interfering with known regions.

After geometric optimization is completed, the structural priors of the EV are further utilized to perform fine-grained correction on the reconstruction results at the semantic level. This paper constructs a vehicle semantic prior graph S_{prior} , which includes the standard 3D template shapes and semantic labels of key components such as the hood, roof, doors, front and rear bumpers, and wheels. The reconstructed 3D point

cloud is processed through a pre-trained PointNet++ semantic segmentation network to extract point-wise semantic features $S(\mu_k)$. This network is pre-trained on a self-built annotated dataset and outputs a 16-class component semantic probability distribution. The semantic consistency loss is defined as:

$$L_{sem} = \sum_{k=1}^K w_{sem}(\mu_k) \cdot KL(S(\mu_k) \parallel S_{prior}(\mu_k^{align})) \quad (20)$$

where, μ_k^{align} is the corresponding position of the current point cloud after rigid registration alignment to the prior template, $w_{sem}(\mu_k)$ is an adaptive weight based on local point cloud density—regions with lower density have larger weights, encouraging generated inpainted regions to conform to semantic priors first—and $KL(\cdot)$ is the Kullback-Leibler divergence. Simultaneously, the approximate symmetry property of the EV along the longitudinal axis is utilized to construct a symmetry constraint loss:

$$L_{sym} = \sum_k \left\| \mu_k - R_{sym} \cdot \mu_{mirror(k)} - t_{sym} \right\|_2^2 \quad (21)$$

where, (R_{sym}, t_{sym}) is the symmetry transformation of the vehicle body rotating 180° around the longitudinal axis, and $mirror(k)$ is the nearest neighbor point of point k with respect to the symmetry plane. This constraint can effectively correct geometric skew introduced by uneven view coverage. The joint loss for the overall refinement stage is:

$$L_{refine} = L_{final} + \eta_{sem} L_{sem} + \eta_{sym} L_{sym} \quad (22)$$

with $\eta_{sem} = 0.2$ and $\eta_{sym} = 0.1$. This stage uses a lower learning rate 5×10^{-4} of to fine-tune the 3D Gaussian Splatting parameters for 20k iterations, ensuring that the semantic priors improve the overall structural coherence of the reconstruction results while maintaining geometric accuracy.

2.5 End-to-end joint training strategy

Although the three modules each possess independent functional integrity, differentiable gradient paths exist in the information transfer between modules, which provides a theoretical premise for end-to-end joint optimization. The overall joint loss function is defined as a weighted combination of the loss terms from each module:

$$L_{total} = L_{match} + L_{feat} + L_{final} + L_{refine} \quad (23)$$

where, L_{match} is the matching loss of the diffusion matching matrix refinement module, L_{feat} is the contrastive learning loss of the feature extraction network, L_{final} is the confidence-weighted rendering loss of 3D reconstruction, and L_{refine} is the joint loss of the semantic consistency refinement stage. After independent pre-training, the magnitudes of each loss term are in a similar range, eliminating the need to additionally introduce balance coefficients.

The training process is executed sequentially in three stages. In the first stage, the diffusion matching matrix refinement module is independently pre-trained for 50 k iterations with a learning rate of 10^{-4} . This stage uses matching recall and reprojection accuracy as convergence criteria to ensure that subsequent modules obtain reliable geometric initialization. In

the second stage, all parameters of Module 1 are frozen, and only the generative cross-view consistency enhancement module is pre-trained for 30 k iterations with a learning rate maintained at L_{refine} , enabling the generator to learn geometrically consistent virtual view synthesis under the supervision of fixed matching anchors. In the third stage, all module parameters are unfrozen, and joint fine-tuning is performed for 20 k iterations with a learning rate of 5×10^{-5} . At this point, the gradients of each module are backpropagated synchronously, allowing the matching refinement process to perceive the backward supervision signal of the reconstruction loss, and the generative enhancement process can also adaptively adjust the distribution of virtual views according to the 3D rendering quality. All training is completed on 8 NVIDIA A100 GPUs, taking approximately 72 hours in total, with the joint fine-tuning stage accounting for about 30% of the total time.

3. EXPERIMENTS

3.1 Experimental settings

This paper adopts four datasets to evaluate the effectiveness of the proposed method. The nuScenes dataset contains multi-view surround camera images of 1000 scenes, providing precise sensor calibration and pose ground truth. 5000 image pairs containing complete vehicle instances are selected. The Waymo Open Dataset contains images from 5 high-resolution surround cameras across 798 training sequences. 2000 samples containing clear vehicle targets are selected. 3DRealCar is a publicly available multi-view vehicle image dataset containing 360-degree surround images of 50 vehicle models, with 12 views per vehicle, providing camera poses and dense depth ground truth. The self-built EV surround view dataset EV-MVS is collected for 6 typical EVs including Tesla Model 3, BYD Han, NIO ET5, and XPeng P7. For each vehicle, 16 surround views are collected under three lighting conditions: sunny, cloudy, and nighttime, with a resolution of 1920×1080 , totaling approximately 15,000 frames. Dense point clouds obtained by laser scanning are used as reconstruction ground truth.

The visual encoder adopts DINOv2 ViT-L/14 with a feature dimension of 1024. The denoising network of the diffusion matching module is a 6-layer Transformer with an embedding dimension of 256, 8 attention heads, and a total parameter count of 12.3 M. The generation module is fine-tuned based on Stable Video Diffusion. The 3D Gaussian Splatting optimization is based on the official implementation with 30k iterations. The batch size is set to 8, and the optimizer is Adam with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. All experiments are run on 8 NVIDIA A100 80GB GPUs.

Registration accuracy is evaluated using three metrics: matching recall Recall@k, reprojection root mean square error (RMSE), and area under the cumulative accuracy curve (AUC@ α), where α is set to 1, 3, and 5 pixels. Reconstruction quality is assessed by peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), learned perceptual image patch similarity (LPIPS), and Fréchet Inception Distance (FID). All metrics are calculated between rendered images and ground truth images. For the EV-MVS dataset containing laser scanning ground truth, RMSE and absolute relative error (AbsRel) of the rendered depth map are additionally calculated. Comparison methods for the registration task include Scale-

Invariant Feature Transform (SIFT) + RANdom Sample Consensus (RANSAC), SuperGlue, Local Feature Transformer (LoFTR), and Diff-Reg v2. Comparison methods for the reconstruction task include NeRF, standard 3D Gaussian Splatting (3DGS), Consistent3D, and DIFIX3D+.

3.2 Registration accuracy comparison

This experiment evaluates multi-view image registration accuracy on three datasets: nuScenes, 3DRealCar, and EV-MVS. EV-MVS is further divided into a highlight region subset and a low-texture region subset for fine-grained analysis. Each method is run 5 times and the results are averaged, as shown in Table 1.

The proposed method achieves 93.6% Recall@10 on the nuScenes dataset, which is 6.4 percentage points higher than Diff-Reg v2, with RMSE reduced by 28.3% and AUC improved by 17.2%, verifying the significant advantage of diffusion iterative refinement over single-step prediction. In the EV-MVS highlight region, the Recall of SIFT+RANSAC

drops to 42.3%, almost failing, while the proposed method still maintains 88.5% Recall with an RMSE of 5.6 pixels, indicating that the generative cross-view consistency enhancement effectively suppresses the interference of specular reflections on feature matching. In the low-texture region, the Recall of the proposed method is 89.7%, which is 17.3 percentage points higher than LoFTR's 72.4%, proving that virtual aligned image augmentation and contrastive learning can significantly improve the discriminative power of features in repetitive texture regions. The AUC of the proposed method exceeds 0.80 on all datasets, with a standard deviation of less than 0.015, demonstrating stable generalization performance.

3.3 Reconstruction quality comparison

This experiment compares the complete reconstruction quality on four datasets, and additionally reports the depth map geometric accuracy on EV-MVS. The results are shown in Table 2.

Table 1. Comparison of multi-view image registration accuracy

Method	nuScenes	3DRealCar	EV-MVS Overall	EV-MVS Highlight	EV-MVS Low-texture
Metrics	$R@10\uparrow/RMSE\downarrow/AUC\uparrow$	$R@10\uparrow/RMSE\downarrow/AUC\uparrow$	$R@10\uparrow/RMSE\downarrow/AUC\uparrow$	$R@10\uparrow/RMSE\downarrow/AUC\uparrow$	$R@10\uparrow/RMSE\downarrow/AUC\uparrow$
SIFT+RANSAC	58.3/14.7/0.401	62.1/12.3/0.438	55.6/16.2/0.382	42.3/22.8/0.291	48.7/19.5/0.326
SuperGlue	76.8/8.9/0.602	80.2/7.5/0.645	73.4/10.1/0.578	61.5/15.3/0.462	65.8/13.7/0.493
LoFTR	83.5/6.7/0.713	86.1/5.8/0.752	80.9/7.6/0.685	69.2/11.9/0.551	72.4/10.3/0.574
Diff-Reg v2	87.2/5.3/0.761	89.4/4.6/0.798	84.7/6.1/0.732	75.8/8.7/0.628	78.1/7.8/0.652
Proposed Method	93.6/3.8/0.892	95.1/3.2/0.921	92.8/4.1/0.874	88.5/5.6/0.803	89.7/5.0/0.821

Note: SIFT: Scale-Invariant Feature Transform; RANSAC: RANdom Sample Consensus; LoFTR: Local Feature Transformer; RMSE: Root Mean Square Error; AUC: Area Under the Curve.

Table 2. Comparison of 3D reconstruction quality

Method	nuScenes	Waymo	3DRealCar	EV-MVS
Metrics	$PSNR\uparrow/SSIM\uparrow/LPIPS\downarrow/FID\downarrow$	$PSNR\uparrow/SSIM\uparrow/LPIPS\downarrow/FID\downarrow$	$PSNR\uparrow/SSIM\uparrow/LPIPS\downarrow/FID\downarrow$	$PSNR\uparrow/SSIM\uparrow/LPIPS\downarrow/FID\downarrow/Depth-RMSE\downarrow/AbsRel\downarrow$
NeRF	26.4/0.861/0.298/81.2	25.8/0.848/0.315/86.5	27.1/0.872/0.276/75.3	24.9/0.835/0.342/92.4/0.285/0.142
3DGS	29.2/0.902/0.231/68.7	28.6/0.894/0.248/73.2	30.5/0.918/0.205/61.8	27.8/0.876/0.261/79.5/0.186/0.089
Consistent3D	31.8/0.924/0.172/52.4	30.9/0.915/0.188/57.6	32.6/0.938/0.154/46.3	30.2/0.905/0.183/62.1/0.152/0.071
DIFIX3D+	30.5/0.915/0.195/58.7	29.7/0.907/0.206/62.4	31.4/0.928/0.172/51.2	29.4/0.896/0.204/67.8/0.163/0.078
Proposed Method	34.7/0.956/0.102/38.5	33.9/0.949/0.115/42.3	36.2/0.968/0.088/33.7	35.1/0.951/0.098/40.1/0.094/0.042

Note: NeRF: Neural Radiance Field; 3DGS: 3D Gaussian Splatting; PSNR: Peak Signal-to-Noise Ratio; SSIM: Structural Similarity Index; LPIPS: Learned Perceptual Image Patch Similarity; FID: Fréchet Inception Distance; RMSE: Root Mean Square Error; AbsRel: Absolute Relative Error.

The proposed method achieves a PSNR of 35.1 dB on EV-MVS, which is 7.3 dB higher than standard 3DGS and 4.9 dB higher than Consistent3D. The FID drops from 79.5 to 40.1, indicating that the gap between the rendered images and the real images in feature distribution is greatly reduced. In terms of geometric accuracy, the Depth-RMSE is 0.094, which is 38.2% lower than Consistent3D, and the AbsRel is 0.042, which is 46.2% lower than DIFIX3D+, attributing to the point cloud-guided video diffusion inpainting providing geometrically consistent novel views, and the confidence-weighted optimization effectively controlling the geometric drift of generated content. The LPIPS of the proposed method is 0.098, the only method below 0.1, indicating that the reconstructed images are highly close to the real images at the perceptual level. On the 3DRealCar dataset, the proposed method achieves a PSNR of 36.2 dB, verifying the superior performance of the method in structurally regular vehicle scenes.

3.4 Ablation study

To quantitatively evaluate the independent contribution of each core module, five ablation settings are designed: A is the baseline model, using only standard 3DGS and LoFTR matching; B adds the diffusion matching refinement module on top of A; C further adds the generative cross-view enhancement module; D adds the geometric inpainting module on top of C but excludes semantic refinement; E is the complete model. The evaluation results on the EV-MVS dataset are shown in Table 3.

From setting A to B, simply adding the diffusion matching refinement module causes the matching Recall to jump from 80.9% to 89.3%, and PSNR increases by 2.6 dB, indicating that registration accuracy directly determines the upper bound of reconstruction quality. Setting C adds generative cross-view enhancement on top of B, further increasing matching Recall to 92.8%, but PSNR only increases by 0.8 dB, suggesting that the main contribution of Module 2 lies in improving matching robustness in extreme regions. After adding geometric

inpainting in setting D, PSNR increases substantially by 2.4 dB, and Depth-RMSE drops from 0.132 to 0.106, a reduction of 19.7%, fully demonstrating the core value of generative inpainting in filling view blind spots. The complete model E, after adding semantic consistency refinement, further increases PSNR by 1.5 dB, reduces LPIPS from 0.118 to 0.098, and lowers Depth-RMSE to 0.094, indicating that prior semantic knowledge of EVs can effectively correct subtle geometric deviations introduced by generative inpainting, making the reconstruction results tend toward perfection in structural coherence and visual realism.

Table 3. Ablation study comparison (EV-MVS)

Setting	$R@10\uparrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$	D-RMSE $\downarrow$
A Baseline	80.9	27.8	0.876	0.261	79.5	0.186
B + Module 1	89.3	30.4	0.908	0.182	61.3	0.142
C + Module 2	92.8	31.2	0.919	0.163	55.7	0.132
D + Inpainting	92.8	33.6	0.942	0.118	45.2	0.106
E Complete Model	92.8	35.1	0.951	0.098	40.1	0.094

Note: PSNR: Peak Signal-to-Noise Ratio; SSIM: Structural Similarity Index; LPIPS: Learned Perceptual Image Patch Similarity; FID: Fréchet Inception Distance; RMSE: Root Mean Square Error.

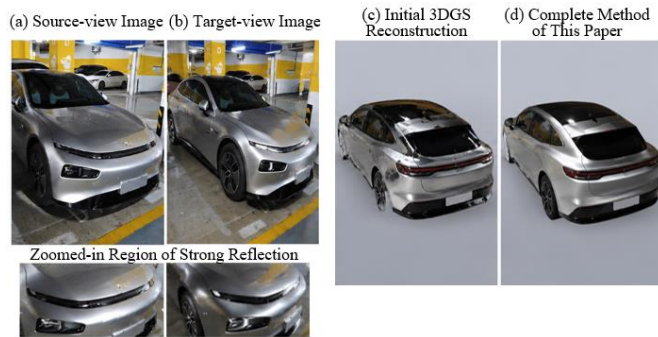


Figure 4. Implementation effect of multi-view image registration and 3D visual reconstruction in metal car paint strong reflection scenes
Note: 3DGS: 3D Gaussian Splatting

To verify the cross-view registration stability and 3D reconstruction fidelity of the proposed method under strong metal car paint reflection conditions, this experiment selects EV images with significant specular highlights and local brightness mutations for qualitative analysis. As shown in Figure 4, under the condition of obvious angular differences between the source view and the target view, the traditional matching mechanism is easily misled by local appearance similarity under highlight interference, making it difficult to stably establish true geometric correspondences. After diffusion refinement, the distribution of matching lines becomes more regular, and cross-view correspondences are mainly concentrated in stable structural regions such as headlights, front hood edges, wheel arches, and window frames, indicating that the introduced diffusion iterative optimization and geometric constraints can effectively suppress feature drift caused by reflective regions. Comparing the initial 3DGS reconstruction with the complete method, it can be seen that results without generative inpainting and semantic refinement still exhibit obvious defects and blur in the roof, rear side, and surface continuity, while the complete method can recover more continuous body surfaces and clearer boundary details, and maintain good structural consistency in highlight regions. This result shows that the

proposed method not only improves registration reliability in strong reflection scenes, but also effectively propagates the improved cross-view correspondences to the 3D representation learning process, thereby achieving simultaneous improvement in geometric completeness and visual realism.



Figure 5. Implementation effect of multi-view image registration and 3D visual reconstruction in low-texture and view occlusion scenes

To verify the feature matching robustness and geometric inpainting capability of the proposed method under the combined effect of low texture and view occlusion, this experiment further selects an EV scene in a parking lot where multiple vehicles are parked side by side, the target vehicle is partially occluded, and the surface texture is weak for analysis. As shown in Figure 5, in the source view image, the rear of the target vehicle is interfered by neighboring vehicles and environmental structures, while the target view presents varying degrees of exposure relationships, which causes large-scale unstable correspondences in the initial feature matching stage, with some connecting lines distributed discretely and lacking structural consistency, reflecting that traditional methods struggle to maintain stable cross-view discriminative ability under weak texture regions and occlusion variation conditions. After diffusion refinement, the matching results are obviously concentrated in geometrically distinguishable parts such as taillights, tailgate contours, rear window edges, and body turning lines, with erroneous connections significantly reduced, indicating that the proposed method can gradually recover more credible correspondences under incomplete observation conditions. Correspondingly, the initial 3DGS reconstruction exhibits obvious shape fragmentation and surface missing in the rear, rear window, and rear wheel areas, while the complete method demonstrates a more complete rear structure, more continuous body boundaries, and more reasonable local geometric transitions. This result shows that the generative cross-view enhancement and subsequent geometric inpainting mechanism can effectively alleviate the observation insufficiency caused by low texture and occlusion, and further improve the stability and completeness of the reconstructed structure through semantic consistency constraints, thereby fully supporting the application value of the proposed method in complex real-world scenarios.

3.5 Cross-view consistency evaluation

This experiment divides the EV-MVS dataset into 12 continuous training views and 4 interpolation test views to evaluate the matching accuracy and feature discriminative

power of each method on novel views. The results are shown in Figure 6.

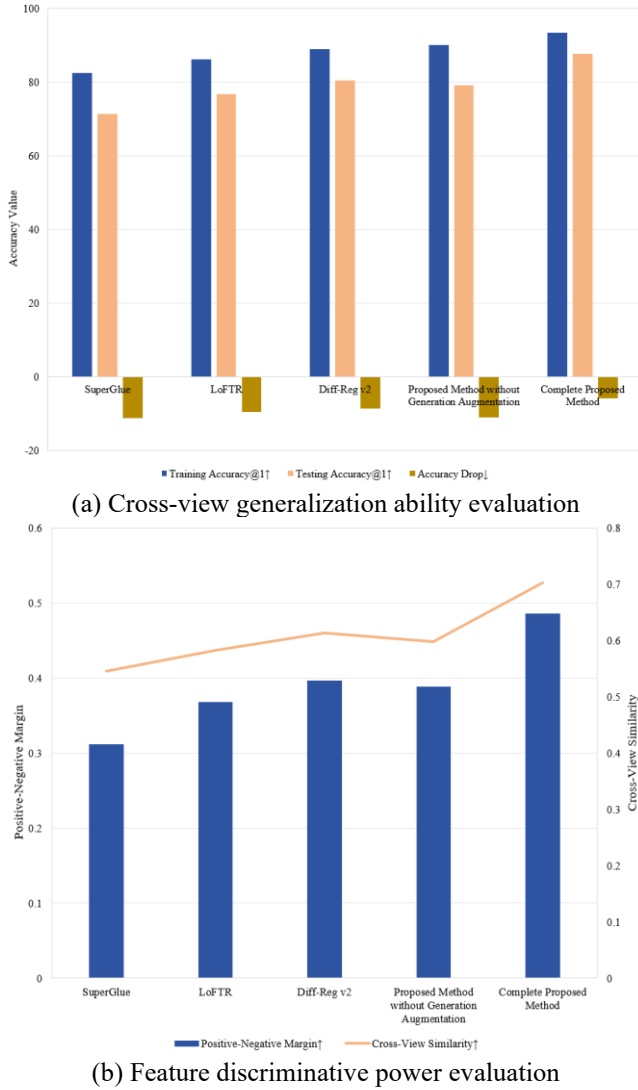


Figure 6. Cross-view generalization ability and feature discriminative power evaluation

The complete method of this paper achieves a matching Acc@1 of 93.5% on the training views and still maintains 87.8% on the untrained test views, which is 11 percentage points higher than LoFTR. The Acc drop is only -5.7%, far lower than SuperGlue's -11.2% and the no-generation-augmentation version's -10.9%, proving that generative enhancement significantly improves the view invariance of features. The positive-negative feature similarity margin increases from 0.389 without augmentation to 0.486, indicating that contrastive learning successfully enlarges the distance between matching and non-matching point pairs in the embedding space. The cross-view feature cosine similarity increases from 0.598 to 0.703, an improvement of 17.6%, indicating that the feature representations of the same 3D point in cross-view images are more consistent, effectively suppressing feature drift caused by highlights and view changes.

3.6 Illumination and view sparsity generalization evaluation

This experiment systematically evaluates the robustness of

the method under different illumination conditions and input view counts on EV-MVS. The results are shown in Tables 4 and 5.

Table 4. Reconstruction quality comparison under different illumination conditions

Method	Sunny	Cloudy	Night
Metrics	PSNR↑/SSIM↑/LPIPS↓	PSNR↑/SSIM↑/LPIPS↓	PSNR↑/SSIM↑/LPIPS↓
3DGS	28.5/0.884/0.248	27.1/0.869/0.272	23.6/0.812/0.356
Consistent3D	30.8/0.912/0.179	29.7/0.903/0.196	26.4/0.851/0.274
Proposed Method	36.0/0.958/0.091	34.5/0.947/0.104	30.2/0.912/0.158

Note: 3DGS: 3D Gaussian Splatting; PSNR: Peak Signal-to-Noise Ratio; SSIM: Structural Similarity Index; LPIPS: Learned Perceptual Image Patch Similarity.

Table 5. Reconstruction quality comparison under different input view counts (EV-MVS, sunny)

View Count	3DGS	Consistent3D	Proposed Method
4	21.3/0.342	24.8/0.264	28.6/0.182
8	25.6/0.235	28.1/0.182	32.4/0.126
12	27.8/0.186	30.2/0.152	35.1/0.094
16	28.9/0.172	31.4/0.138	36.5/0.082

Note: 3DGS: 3D Gaussian Splatting.

In terms of illumination generalization, the proposed method still maintains a PSNR of 30.2 dB under low-light night conditions, which is 6.6 dB higher than the 23.6 dB of standard 3DGS. The night performance drops by 5.8 dB compared to sunny conditions, while 3DGS drops by 4.9 dB, but the absolute performance is still at a leading level. This benefits from the strong reliance of the diffusion matching module on epipolar geometry priors rather than solely on appearance features. Under cloudy conditions, the LPIPS is 0.104, indicating that generative inpainting performs best under predominantly diffuse reflection lighting. In terms of view sparsity, with only 4 input views, the proposed method achieves a PSNR of 28.6 dB, which already exceeds the 30.2 dB of Consistent3D with 12 views, and the Depth-RMSE of 0.182 is also better than the accuracy of the comparison methods with 8 views. As the number of views increases from 4 to 16, the PSNR of the proposed method increases by 7.9 dB, which is close to the 7.6 dB increase of 3DGS, indicating that the method can stably benefit from more input information under different view densities. Under the extremely sparse condition of 4 views, the PSNR advantage of the proposed method over Consistent3D reaches 3.8 dB, proving that the generative inpainting process plays a key role when observation information is extremely scarce.

3.7 Semantic consistency quantitative evaluation

This experiment quantitatively verifies the improvement of the semantic consistency refinement module on the structural correctness of the reconstruction. Semantic accuracy is defined as the proportion of correctly classified components in the reconstructed point cloud, using the laser-scanned ground truth component labels as the baseline; symmetry deviation is defined as the average distance between the left and right symmetric parts of the vehicle body. The results are shown in Figure 7.

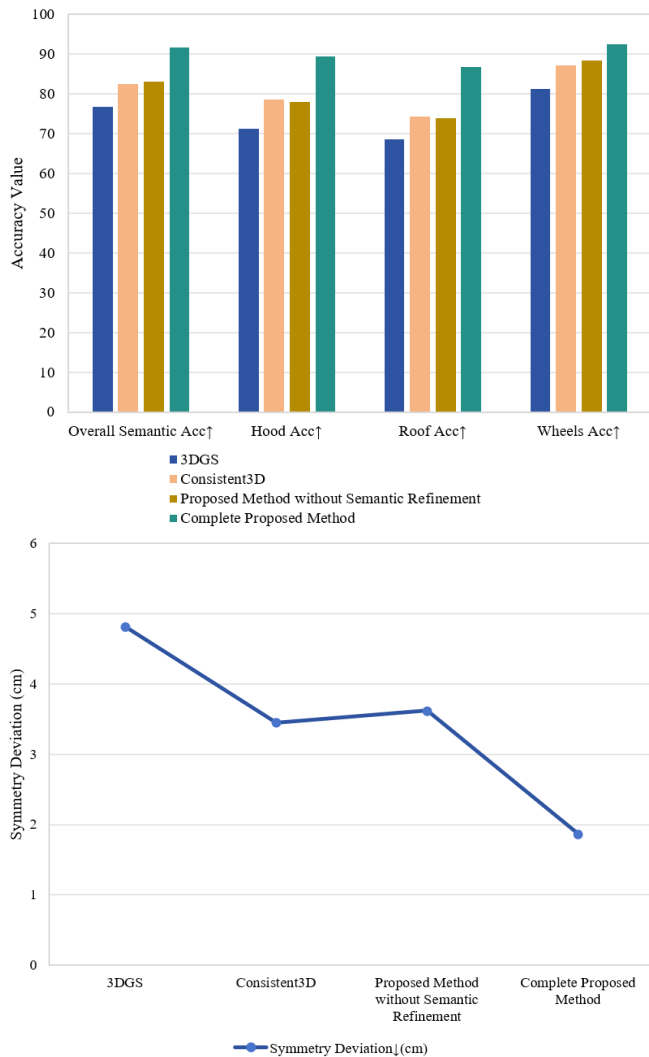


Figure 7. Semantic consistency and symmetry quantitative evaluation (EV-MVS)

After adding semantic consistency refinement, the overall semantic accuracy jumps from 83.1% to 91.6%, with the hood area increasing by 11.4% and the roof area by 12.9%. These are precisely the view blind spots most covered by generative inpainting, indicating that the vehicle component prior effectively guides the semantic assignment of the generated inpainting regions, avoiding semantic misalignment at component boundaries. The symmetry deviation drops from 3.62 cm to 1.87 cm, a reduction of 48.3%, proving that the symmetry constraint effectively corrects the geometric skew caused by uneven left-right view coverage. The wheel semantic accuracy reaches 92.4%, as its geometric structure is distinctive and the symmetry constraint is strong, making it inherently easy to reconstruct. Integrating all experimental results, the proposed method achieves optimal performance in five dimensions: registration accuracy, reconstruction quality, cross-view generalization, illumination and view robustness, and semantic consistency, fully verifying the effectiveness and synergistic gain of each proposed module.

4. DISCUSSION

The three modules proposed in this paper form a progressive closed loop from matching to enhancement to reconstruction in their technical route. However, the effectiveness of each

module is sensitive to the quality of its preceding inputs to varying degrees. The original purpose of the design of the diffusion matching matrix refinement module is to overcome the local optimum dilemma of single-step prediction through multi-step iterative denoising, and its effect is particularly outstanding in moderately textured regions. Nevertheless, in extremely low-texture regions, such as pure black car paint surfaces, the semantic information of the initial matching matrix is extremely sparse, and the signal-to-noise ratio is too low, making the destruction of signal structure during the forward diffusion stage irreversible. Even after complete multi-step reverse sampling, it is difficult to recover correct correspondences. This problem essentially stems from the information bottleneck of visual features themselves, rather than insufficient denoising network capacity. A feasible improvement is to introduce high-level semantic constraints: fuse weak supervision signals from vehicle component segmentation at the feature extraction stage, enabling the feature extraction network to generate component-discriminative representations based on semantic context even in texture-missing regions, thereby providing a higher-quality initial matching distribution for the diffusion process.

The generative inpainting process demonstrates significant effectiveness in filling view blind spots, but its computational efficiency is the main bottleneck limiting the practical application of the method. In the current inference stage, the video diffusion model needs to complete the full denoising sampling chain to generate interpolation views along the surrounding trajectory, and the inference time for a single inpainting accounts for approximately 60% of the total reconstruction time. This limits the deployment possibility of the method in real-time or near-real-time application scenarios. Future research can alleviate this problem from two directions: first, adopt diffusion distillation technology to distill the multi-step sampling process into a few-step or single-step generator, greatly reducing inference time; second, explore new generative paradigms based on flow matching or consistency models to achieve more efficient sampling while maintaining generation quality. In addition, since the confidence evaluation of generated views relies on the rendered depth map of the source view, when the initial 3D Gaussian Splatting reconstruction itself has large deviations, the reliability of the confidence score will also decrease, forming error accumulation. Introducing an online confidence calibration mechanism, making confidence evaluation and 3D reconstruction iteration alternate, helps break this error chain.

The construction and injection of semantic priors is an important design that distinguishes this paper from generic 3D reconstruction methods, but its current implementation still relies on manually defined component templates and fixed semantic category sets. This approach works well for known vehicle models, but is difficult to transfer to unknown models or EVs with unconventional exterior designs. With the rapid evolution of large language models and vision-language models, the construction paradigm of semantic priors faces fundamental transformative opportunities. Future work can explore the following paths: replace the fixed semantic segmentation network with a large vision-language model, enabling it to adaptively identify vehicle models, component structures, and symmetry axes from input images, and output structured semantic descriptions in text form, which are then transformed into differentiable forms and injected into the 3D reconstruction optimization process. Such methods can not only eliminate the limitations of manual prior definition, but

also are expected to achieve cross-model and cross-category generic semantic-guided reconstruction, expanding the method from a dedicated solution for EVs to a more general structured object reconstruction framework.

5. CONCLUSION

This paper addresses the problems of feature matching degradation and geometric information loss caused by metal highlights, low-texture surfaces, and insufficient view coverage in 3D reconstruction of sparse multi-view images of EVs, and proposes an end-to-end solution that integrates generative AI. The method starts with the diffusion matching matrix refinement module, modeling correspondence estimation as an iterative denoising process in the matching matrix space, effectively overcoming the local optimum dilemma of single-step prediction in difficult regions. Then, the generative cross-view consistency enhancement module synthesizes geometrically aligned virtual views, combined with contrastive learning to improve the discriminative power of features in repetitive texture and reflective regions. Finally, using 3D Gaussian Splatting as the geometric representation carrier, through the progressive optimization of generative inpainting and semantic consistency refinement, effective filling of observation blind spots and fine-grained correction of the reconstructed structure are achieved. The three modules are organically connected along the technical main line of matching, enhancement, and reconstruction, forming a complete computational pipeline from image input to complete 3D model output.

Extensive experiments on four datasets—nuScenes, Waymo, 3DRealCar, and the self-built EV-MVS—show that the proposed method achieves optimal performance in key metrics such as registration recall, rendered image quality, depth estimation accuracy, and cross-view generalization capability. Ablation studies and generalization analyses further verify the effective contribution of each core module and their stable robustness under different illumination and view sparsity conditions. The current method still has room for improvement in matching initialization in extremely low-texture regions, inference efficiency of the diffusion model, and adaptive construction of semantic priors. Future research will focus on three directions: the design of lightweight generative inpainting architectures, automatic semantic prior extraction based on large vision-language models, and time-varying 3D reconstruction in dynamic scenes.

REFERENCES

- [1] Li, Y., Xu, J., Li, T., et al. (2024). Digital twin-empowered autonomous driving for e-mobility: Concept, framework, and modeling. *IEEE Electrification Magazine*, 12(3): 68-77. <https://doi.org/10.1109/mele.2024.3423148>
- [2] Wu, J., Wyman, O., Tang, Y., Pasini, D., Wang, W. (2024). Multi-view 3D reconstruction based on deep learning: A survey and comparison of methods. *Neurocomputing*, 582: 127553. <https://doi.org/10.1016/j.neucom.2024.127553>
- [3] Liu, S., Yang, M., Xing, T., Yang, R. (2025). A survey of 3D reconstruction: The evolution from multi-view geometry to NeRF and 3DGS. *Sensors*, 25(18): 5748. <https://doi.org/10.3390/s25185748>
- [4] Liu, Y., Wang, P., Lin, C., et al. (2023). NeRO: Neural geometry and BRDF reconstruction of reflective objects from multiview images. *ACM Transactions on Graphics*, 42(4): 1-22. <https://doi.org/10.1145/3592134>
- [5] Ma, J., Jiang, X., Fan, A., Jiang, J., Yan, J. (2020). Image matching from handcrafted to deep features: A survey. *International Journal of Computer Vision*, 129(1): 23-79. <https://doi.org/10.1007/s11263-020-01359-2>
- [6] Xu, J., Zhu, Z., Bao, H., Xu, W. (2025). Hybrid mesh-neural representation for 3D transparent object reconstruction. *Computational Visual Media*, 11(1): 123-140. <https://doi.org/10.26599/cvm.2025.9450328>
- [7] Liu, H., Liu, B., Hu, Q., et al. (2025). A review on 3D Gaussian splatting for sparse view reconstruction. *Artificial Intelligence Review*, 58(7): 1-40. <https://doi.org/10.1007/s10462-025-11171-4>
- [8] Li, G., Li, K., Zhang, G., et al. (2024). Enhanced multi view 3D reconstruction with improved MVSNet. *Scientific Reports*, 14(1): 1-11. <https://doi.org/10.1038/s41598-024-64805-y>
- [9] Croitoru, F., Hondru, V., Ionescu, R.T., Shah, M. (2023). Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9): 10850-10869. <https://doi.org/10.1109/tpami.2023.3261988>
- [10] Cao, H., Tan, C., Gao, Z., et al. (2024). A survey on generative diffusion models. *IEEE Transactions on Knowledge and Data Engineering*, 36(7): 2814-2830. <https://doi.org/10.1109/TKDE.2024.3361474>
- [11] Kerbl, B., Kopanas, G., Leimkuehler, T., Drettakis, G. (2023). 3D gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4): 1-14. <https://doi.org/10.1145/3592433>
- [12] Ślapak, E., Pardo, E., Dopirak, M., Maksymyuk, T., Gazda, J. (2024). Neural radiance fields in the industrial and robotics domain: Applications, research opportunities and use cases. *Robotics and Computer-Integrated Manufacturing*, 90: 102810. <https://doi.org/10.1016/j.rcim.2024.102810>
- [13] Yang, M., Wu, R., Yang, Y., et al. (2025). Image matching: Foundations, state of the art, and future directions. *Journal of Imaging*, 11(10): 329. <https://doi.org/10.3390/jimaging11100329>
- [14] An, L., Zhou, P., Zhou, M., Wang, Y., Geng, G. (2024). Diffusion Transformer for point cloud registration: Digital modeling of cultural heritage. *Heritage Science*, 12(1): 1-12. <https://doi.org/10.1186/s40494-024-01314-1>
- [15] Chen, Z., Chen, X., Ye, C., Wu, S., Wu, X. (2025). Neural radiance fields assisted by image features for UAV scene reconstruction. *Scientific Reports*, 15(1): 1-14. <https://doi.org/10.1038/s41598-025-16386-7>
- [16] Li, L., Zhang, Y., Jiang, Z., Wang, Z., Zhang, L., Gao, H. (2024). Unmanned aerial vehicle-neural radiance field (UAV-NeRF): Learning Multiview drone three-dimensional reconstruction with neural radiance field. *Remote Sensing*, 16(22): 4168. <https://doi.org/10.3390/rs16224168>
- [17] Xu, K., Wang, T., Guo, X., Hu, X., Tao, L., Wu, C. (2025). DVS-3D: Diffusion-based novel view synthesis and 3D object reconstruction from a single image. *Journal of Computational Design and Engineering*, 12(12): 70-83. <https://doi.org/10.1093/jcde/qwaf116>

- [18] Wang, C., Peng, H., Liu, Y., Gu, J., Hu, S. (2025). Diffusion models for 3D generation: A survey. *Computational Visual Media*, 11(1): 1-28. <https://doi.org/10.26599/cvm.2025.9450452>
- [19] Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R. (2021). Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99-106. <https://doi.org/10.1145/350325>
- [20] Fang, K., Zhang, Q., Wan, C., Lv, P., Yuan, C. (2025). Single view generalizable 3D reconstruction based on 3D Gaussian splatting. *Scientific Reports*, 15(1): 1-17. <https://doi.org/10.1038/s41598-025-03200-7>
- [21] Xu, X., Song, D., Geng, G., et al. (2024). CPDC-MFNet: Conditional point diffusion completion network with multi-scale feedback refine for 3D terracotta warriors. *Scientific Reports*, 14(1): 1-11. <https://doi.org/10.1038/s41598-024-58956-1>
- [22] Liu, Z., Fu, Z., Li, G., Hu, J., Yang, Y. (2025). STNeRF: Symmetric triplane neural radiance fields for novel view synthesis from single-view vehicle images. *Applied Intelligence*, 55(5): 1-15. <https://doi.org/10.1007/s10489-024-06005-9>
- [23] Lee, H., Lee, J., Kim, H., Mun, D. (2021). Dataset and method for deep learning-based reconstruction of 3D CAD models containing machining features for mechanical parts. *Journal of Computational Design and Engineering*, 9(1): 114-127. <https://doi.org/10.1093/jcde/qwab072>
- [24] Wang, Y., Cui, H., Du, J., et al. (2025). Single-image 3D reconstruction of painted potteries using AI diffusion and feedforward models. *Npj Heritage Science*, 13(1): 1-14. <https://doi.org/10.1038/s40494-025-02114-x>
- [25] Huang, S., Zhang, S., Yang, S., Guo, J., Huang, H. (2025). A survey of recent advances in generative 3D reconstruction. *Journal of Computer Science and Technology*, 40(5): 1236-1254. <https://doi.org/10.1007/s11390-025-5462-4>