GNFM: Generalizable NeRF with View-Aware Feature Modulation for Reflective Surgical Instrument Rendering toward Robot-assisted Surgery

Lifei Xiao
School of Computer Science, Nanjing University of Information Science and Technology
Nanjing, China
xlf@nuist.edu.cn

Zhenjie Zhao*
Academy for Advanced Interdisciplinary Studies, College of Artificial Intelligence, Nankai University
Tianjin, China
zzhaoao@nankai.edu.cn

Abstract

Accurate 3D reconstruction of small, highly reflective surgical instruments underpins robust perception and safe operation in robot-assisted surgery. Traditional multi-view reconstruction methods often fail to capture subtle geometry and view-dependent reflections. Motivated by recent advances in Neural Radiance Fields (NeRF), we explore their potential for reconstructing surgical instruments, where photorealistic rendering serves as an indicator of 3D surface fidelity. However, conventional NeRFs require dense per-scene training, lack cross-scene generalization capability, and struggle with specular highlights and fine structures, posing limitations that are particularly problematic in robot-assisted surgery. To address these challenges, we propose Generalizable NeRF with View-Aware Feature Modulation (GNFM), which extends the Generalizable NeRF Transformer (GNT) with two novel modules: a Joint Multiresolution Hash Encoder (JMHE) for joint spatial and directional encoding, which enhances surface detail and fine-structure recovery, and a View-Conditional Feature-wise Linear Modulation (V-FiLM) that adaptively modulates features according to viewing direction to improve the rendering of specular and view-dependent effects. We also introduce the Reflective Surgical Instrument Dataset (RSID), which consists of 8 synthetic categories with 2,400 multi-view images and 6 real surgical instruments. Experiments demonstrate that GNFM improves the recovery of fine structures and the rendering of specular surfaces, achieving approximately 5% improvements in evaluation metrics such as PSNR, SSIM, and LPIPS over competitive baselines while maintaining comparable computational efficiency. This provides a practical framework for robust perception in robot-assisted surgery. The dataset and code are available at https://h-ci.github.io/GNFM/.

*Corresponding author

Keywords: Neural radiance fields, Transformer, Surgical instruments, Specular rendering

1. Introduction

Robot-assisted surgery requires accurate perception of surgical scenes to ensure safety and enable intelligent decision-making. Among these requirements, 3D reconstruction of surgical instruments plays a central role by offering essential data support for precise manipulation [43, 48] and intraoperative guidance [3, 14]. However, unlike common reconstruction targets such as architectural or natural scenes, surgical instruments present unique challenges. First, their small volume and fine structures include thin edges, sharp tips, and grooves that are difficult to recover with conventional approaches [28]. Second, their highly reflective surfaces, often stainless steel or polished alloys, produce strong view-dependent appearances that violate Lambertian assumptions [39, 22]. Third, clinical constraints such as complex illumination and limited viewpoints further aggravate the difficulty [44].

Traditional multi-view stereo (MVS) [46, 38, 5] and voxel-based reconstruction methods [42, 24] have been widely used for general 3D reconstruction tasks. However, these approaches rely on Lambertian surface assumptions and uniform lighting, which often fail when dealing with small instruments containing intricate structures and strong specular reflections. In particular, fine geometrical details such as grooves or thin edges are difficult to recover, and view-dependent appearances are often completely lost. Recent non-neural efforts partially alleviate these issues via multi-view photometric-stereo and its deep

variants [53, 11], or by polarization and structured-light sensing that helps disambiguate specular reflections [10]. However, such approaches usually require controlled lighting or extra hardware, limiting their practicality in robot-assisted surgery.

Neural rendering [30, 6, 31], and specifically Neural Radiance Fields (NeRF) [18], offers a promising alternative by representing 3D scenes implicitly and learning continuous volumetric radiance from images. NeRF and its variants can capture view-dependent effects and produce high-fidelity novel views. Mip-NeRF [1] introduces a scale-aware formulation that employs conical frustums and integrated positional encoding to reduce aliasing and better represent multi-scale detail. Ref-NeRF [32] structures view-dependent appearance and explicitly decomposes diffuse and specular components, thereby improving reflective-object rendering. More recent works [15, 9] address challenging specularities through specialized directional encodings or reflection-aware rendering strategies, such as Gaussian directional encoding and reflection tracing.

However, most NeRF-based methods remain scene-specific and require retraining for each new object [36, 34], which severely limits their generalization across instruments or surgical scenarios. Recent studies on generalizable NeRF variants, such as PixelNeRF [50], IBRNet [36], and the Generalizable NeRF Transformer [34], have achieved a certain degree of generalization to unseen scenes. However, these methods are mainly evaluated on large objects with low reflectivity. When applied to small surgical instruments with intricate geometry and strong specular reflections, they suffer from significant performance degradation. The main causes include limited ability to recover fine-grained geometry, inadequate modeling of view-dependent reflections, and the lack of validation on small, highly reflective instruments [32, 15, 34].

To enable NeRF generalization in this small, specular object regime, two key challenges must be addressed. First, cross-scene reconstruction needs a compact yet expressive code to preserve tiny structures and specular cues, but conventional sinusoidal encodings often lose high-frequency details. Second, viewpoint-dependent appearance must adapt without distorting geometry. Simple concatenation of position and direction features tends to entangle geometry and appearance, while explicit reflectance decomposition relies on brittle assumptions [32].

To deal with these problems, in this paper, we propose Generalizable NeRF with View-Aware Feature Modulation (GNFM), a framework tailored for reflective surgical instrument rendering. GNFM follows a two-stage generalizable NeRF pipeline and augments the feature encoding stage with two lightweight modules: the Joint Multiresolution Hash Encoder (JMHE) module and the View-Conditional Feature-wise Linear Modulation (V-FiLM) module. The JMHE module jointly hashes positions and viewing directions to form an efficient multi-scale embedding that preserves fine geometry and specular cues across scenes. The V-FiLM module modulates the positional hash features based on viewing direction, producing smooth and consistent responses across viewpoints without altering the sampling or interpolation process. By integrating directional information directly into positional features during encoding, GNFM eliminates the need to explicitly concatenate directional encodings, while strengthening NeRF's ability to handle small, highly reflective objects; as a result, it reconstructs accurate geometry and renders realistic novel views for previously unseen instruments.

For comprehensive evaluation, we construct the Reflective Surgical Instrument Dataset (RSID), a high-fidelity benchmark covering eight synthetic instruments and six real-world surgical tools with diverse geometries, illuminations, and camera poses. We adopt Peak Signal-to-Noise Ratio(PSNR), Structural Similarity Index Measure (SSIM)[40], and the Learned Perceptual Image Patch Similarity (LPIPS)[51] as evaluation metrics. On the unseen-instrument setting without finetuning, GNFM achieves an average improvement of +3.65 dB in PSNR, +0.010 in SSIM, and a reduction of about 32% in LPIPS compared with the strongest generalizable baseline GNT[34]. These results highlight the robustness of GNFM in cross-scene generalization.

The main contributions of this work are summarized as follows:

- We propose a generalizable NeRF framework(GNFM) tailored for rendering small, highly reflective surgical instruments, which enables accurate 3D reconstruction and realistic novel view synthesis across diverse scenes.
- We propose a joint feature modulation mechanism that integrates the Joint Multiresolution Hash Encoder (JMHE) and the View-Conditional Feature-wise Linear Modulation (V-FiLM), which captures fine geometrical details and robustly represents view-dependent reflections.
- We propose the Reflective Surgical Instrument Dataset (RSID), a high-quality dataset designed for benchmarking reflective surgical instrument reconstruction, which facilitates quantitative evaluation of cross-scene generalization.
- We conduct comprehensive experiments, which demonstrate superior performance and cross-scene generalization compared to existing methods.

2. Related Works

2.1. Classical 3D Reconstruction Methods

Classical geometry-based approaches, such as multi-view stereo (MVS) [46, 38, 5] and voxel-based reconstruction [42, 24], have formed the foundation of 3D reconstruction for decades. These methods typically assume Lambertian reflectance and rely on photometric consistency across views. While effective for diffuse or large-scale objects, they often fail to capture fine geometrical structures, for example, thin edges and grooves, and struggle with specular surfaces, which are ubiquitous in surgical instruments. Recent variants, such as photometric-enhanced MVS [37] and hybrid SfM–MVS pipelines [19], attempt to mitigate these limitations by leveraging shading cues, regularization, or more flexible matching schemes. However, their robustness under severe view-dependent reflection remains insufficient, limiting their applicability in surgical scenarios.

Beyond MVS, other traditional techniques have been explored to improve robustness against reflective or poorly textured regions. Shape-from-shading [52] and photometric stereo [41] exploit illumination cues to refine geometry, but these methods often require carefully controlled lighting setups that are unrealistic in clinical contexts. Polarization-based imaging [13, 20] and structured-light scanning [7] have also been employed to disambiguate specular highlights or capture fine surface detail. Nevertheless, such approaches generally depend on specialized hardware, which limits their adoption in robotic surgery where space and equipment are constrained.

Recent years have also seen attempts to integrate deep learning into traditional pipelines. For example, deep-MVS frameworks [46, 8] replace handcrafted matching with CNN-based cost volumes, achieving stronger generalization on textured scenes. Uncertainty-aware MVS [11] and transformer-based stereo [45] further enhance robustness by modeling confidence or global context. However, despite these improvements, such methods remain sensitive to specular reflections and cannot fully recover the fine-grained details of surgical instruments.

2.2. NeRF-based Models for Reconstruction

The emergence of neural implicit representations has introduced a new paradigm for joint geometry and appearance modeling. Neural Radiance Fields (NeRF) [18] learns a volumetric radiance field that maps spatial positions and viewing directions to density and color, enabling high-fidelity novel view synthesis. Extensions such as Mip-NeRF [1] reduce aliasing and enhance multi-scale representations, whereas efficiency oriented approaches like Instant-NGP [21], PlenOctrees [49], Plenoxels [4], and KiloNeRF [26] accelerate training and inference using voxel grids, explicit sparsity, or decomposed MLPs. Despite their practicality, these methods remain scene specific and fail to address the reflective and small scale characteristics of surgical instruments.

2.3. Specularity-Aware and Hybrid NeRF Extensions

Another major direction targets reflectance modeling. Reflection-aware methods such as Ref-NeRF [32] and Spec-NeRF [15] explicitly decompose diffuse and specular components, while NeRF-W [17] and Urban-NeRF [27] separate appearance variations due to illumination in outdoor scenes. More advanced designs include Normal-NeRF [29], which leverages transmittance-gradient normals for disambiguation on reflective surfaces, TraM-NeRF [9], which integrates reflection tracing, and NeRF-Casting [33], which explicitly casts reflection rays to enforce view-dependent consistency. Hybrid neural field strategies such as UNISURF [23], VolSDF [47], and NeuS [35] unify implicit fields with explicit surface priors to refine geometry. While these methods advance reflectance handling, they are mostly validated on macroscopic or moderately reflective scenes, rather than the highly specular, fine-grained instruments studied here. Moreover, these reflection oriented and hybrid approaches remain scene specific and require retraining for each new object, limiting their applicability to generalizable reconstruction tasks.

2.4. 3D Gaussian Splatting and Explicit Representations

Apart from NeRF-based implicit fields, recent progress in explicit neural scene representations has introduced 3D Gaussian Splatting (3DGS) [12]. By representing a scene as a set of anisotropic gaussian primitives, 3DGS achieves impressive real-time rendering efficiency and high visual fidelity in large-scale, mostly diffuse environments. However, its discrete point-based formulation struggles with small-scale or highly reflective objects, where unstable optimization and fragmented highlights often occur. Moreover, explicit gaussian primitives are less effective at modeling continuous specular variations or subtle reflectance transitions. In contrast, implicit volumetric fields as used in NeRF inherently provide a continuous representation of geometry and appearance, which better accommodates the complex specular effects found in surgical instruments. Nevertheless, both implicit and explicit neural field formulations remain constrained by scene specific training and lack the ability to generalize across different objects or lighting conditions, motivating research on generalizable NeRF architectures.

2.5. Generalizable NeRF Variants

To overcome scene specificity, generalizable NeRF families such as PixelNeRF [50], IBRNet [36], and the Generalizable NeRF Transformer (GNT) [34] condition radiance field learning on image features extracted from multiple

views, enabling cross-scene inference without retraining. More recent efforts such as GANESH [16] enhance generalization via meta-learning, while others focus on specularity-aware priors or reflection tracing. However, these generalizable models are predominantly validated on large-scale or moderately reflective objects, and remain insufficient for reconstructing small surgical instruments with intricate geometry and strong specularities.

In summary, existing approaches have substantially advanced 3D reconstruction across both classical and neural paradigms. Classical pipelines provide a strong foundation but are fundamentally limited by Lambertian assumptions and controlled illumination requirements. NeRF-based methods achieve high visual fidelity but remain scene-specific, while specularity-aware and hybrid extensions improve reflectance modeling but at high computational cost or only for large-scale scenes. Generalizable NeRFs reduce retraining costs yet lack robustness in the small and specular domain. This gap motivates our GNFM framework, which explicitly targets the dual challenges of fine-grained geometry recovery and robust cross-scene generalization for reflective surgical instruments.

3. Methods

3.1. Preliminary and Problem Formulation

3.1.1 Generalizable NeRF Transformer (GNT)

The Generalizable NeRF Transformer (GNT) [34] is a NeRF-based framework that enables generalization to unseen scenes through a pre-trained model. It operates through two key stages. First, for a given target view, GNT selects relevant source views and extracts epipolar-aligned point features using a trainable U-Net–like image encoder. These features are subsequently aggregated by the *View Transformer* to construct a 3D coordinate-aligned feature field. Second, the *Ray Transformer* aggregates point-wise features sampled along each ray in the target view and directly predicts the corresponding pixel color. This feed-forward design allows GNT to generalize across scenes without per-scene retraining, making it a strong baseline for building our framework.

View Transformer. The View Transformer in the Generalizable NeRF Transformer (GNT) framework constructs a coordinate-aligned 3D feature field from multiple input views. Unlike the vanilla NeRF, which parameterizes a radiance field via an MLP and optimizes it separately for each scene, GNT formulates a feed-forward mapping from input images to a latent 3D feature field:

$$F(\mathbf{x}) = V(\mathbf{x}; I_1, \ldots, I_N), \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^3$ is a 3D location, $(I_1, I_2, ..., I_N)$ are the source images, and $F(\mathbf{x}) \in \mathbb{R}^d$ denotes the latent feature vector at position $\mathbf{x}$. This feed-forward formulation maps multi-view image evidence into a unified scene representation and maintains permutation invariance with respect to the order of input views. This property ensures that the constructed feature field depends solely on the visual content, rather than the arbitrary arrangement of source views.

Concretely, each source image I_i is processed by an image encoder to extract multi-scale feature maps:

$$F_i = \mathrm{ImageEncoder}(I_i) \in \mathbb{R}^{H \times W \times d}, \tag{2}$$

where H and W denote the spatial dimensions, and d represents the number of feature channels. For a 3D sample point $\mathbf{x}$ on a target ray $r = (o, \mathbf{d})$, we project $\mathbf{x}$ onto the i-th image plane using the camera intrinsics and extrinsics, denoted as $\pi_i(\mathbf{x})$, and extract the corresponding per-view feature through bilinear interpolation:

$$f_i = F_i\big(\pi_i(\mathbf{x})\big) \in \mathbb{R}^d. \tag{3}$$

The View Transformer aggregates the sampled per-view features $\{f_1, f_2, ..., f_N\}$ into a single coordinate-aligned feature vector using a transformer encoder:

$$F(\mathbf{x}, \mathbf{d}) = \mathrm{ViewTransformer}\big(\{f_1, f_2, ..., f_N\}, \mathrm{PE}(\mathbf{x}, \mathbf{d})\big), \tag{4}$$

where $\mathrm{PE}(\mathbf{x}, \mathbf{d})$ denotes the positional encoding that incorporates the 3D coordinate, viewing direction, and relative directions of source views. Two implementation details are important: (i) the transformer is made epipolar-aware by restricting attention to pixels lying on corresponding epipolar lines across source views, which substantially reduces memory cost and injects geometric priors; (ii) the transformer can implicitly detect occlusion and selectively weight visible views, behaving similarly to stereo-matching aggregation. The output $F(\mathbf{x}, \mathbf{d})$ provides a geometrically consistent, multi-view conditioned latent descriptor for each 3D location.

Ray Transformer. Rendering proceeds after obtaining coordinate-aligned point-wise features $F(\mathbf{x})$, where the goal is to compose feature samples along each camera ray into the final pixel color representation. The Generalizable NeRF Transformer (GNT) replaces the classical volumetric integration with an attention-based aggregation mechanism implemented by the Ray Transformer. For each camera ray $r = (o, \mathbf{d})$, a set of m sample points is drawn uniformly between the near and far planes:

$$\mathbf{x}_i = o + t_i\, \mathbf{d}, \quad i = 1, \ldots, m, \tag{5}$$

Each sampled point $\mathbf{x}_i$ yields a feature token $F(\mathbf{x}_i)$, which is concatenated with the corresponding positional and view encodings before being fed into the Ray Transformer. The

Ray Transformer processes the ordered sequence of point-wise features and models inter-sample dependencies using multi-head self-attention:

$$\mathbf{Z}_{1:m} = \mathrm{RayTransformer}\big(F(\mathbf{x}_1), \ldots, F(\mathbf{x}_m)\big), \quad (6)$$

where $\mathbf{Z}_{1:m}$ represents the transformed feature tokens along the ray. A pooled ray representation, obtained via mean aggregation of the output tokens, is passed through a lightweight MLP to predict the final RGB color:

$$C(r) = \mathrm{MLP}\Big(\mathrm{Mean}\big(\mathbf{Z}_{1:m}\big)\Big). \quad (7)$$

Eqs. (6–7) summarize the feature-space rendering process. Conceptually, the attention weights learned by the Ray Transformer act analogously to transmittance or blending coefficients in classical volume rendering, adaptively determining how point-wise features contribute to the final color. The learned aggregation implicitly captures occlusion reasoning, surface smoothness, and view-dependent illumination, allowing the model to represent complex light transport effects, including specular highlights, directly within the latent feature composition. The entire pipeline remains fully differentiable and supports end-to-end optimization of both the image encoder and transformer modules.

3.1.2 Problem Formulation

Given a set of N input images $\{I_i\}_{i=1}^N$ of a reflective surgical instrument, each associated with known camera parameters $\{P_i\}_{i=1}^N$, our objective is to synthesize photorealistic novel views of unseen instruments under arbitrary camera poses.

This problem is particularly challenging because reflective surgical instruments are characterized by small physical scales, intricate geometric details, and strong view-dependent specular reflections. These factors make both traditional geometry based reconstruction and scene specific neural rendering methods inadequate for achieving reliable performance.

3.2. Method Overview

An overview of the proposed **Generalizable NeRF with View-Aware Feature Modulation (GNFM)** is presented in Fig. 1. The architecture retains the two-stage paradigm of GNT: (a) the *View Transformer* aggregates coordinate-aligned epipolar features X from source views under geometric constraints, and (b) the *Ray Transformer* composes point-wise features along each ray into aggregated ray descriptors X_0 for predicting target pixel colors. To better preserve the fine structural details of reflective surgical instruments, we introduce two lightweight yet effective modules: (i) the *Joint Multiresolution Hash Encoder (JMHE)* and (ii) the *View-Conditional Feature-wise Linear Modulation (V-FiLM)*, which jointly enhance feature encoding by integrating spatial and directional features cues. These modules are seamlessly integrated into the feature encoding stage, resulting in enhanced robustness and rendering fidelity without sacrificing computational efficiency.

Specifically, the JMHE explicitly couples spatial positions and viewing directions within a multi-resolution hash encoding, enabling the model to capture both fine-grained geometric details and view-dependent reflective effects. The V-FiLM further applies a smooth, feature-wise modulation conditioned on the view direction, allowing continuous adaptation of feature responses to effectively model reflective and anisotropic surfaces. Together, these modules substantially enhance the representational power of the baseline, particularly when reconstructing small-scale, highly specular surgical instruments.

3.3. Joint Multiresolution Hash Encoder

We propose the **Joint Multiresolution Hash Encoder (JMHE)**, a compact and efficient multi-resolution hash encoding that *jointly* embeds spatial positions and viewing directions through cross-dimensional feature coupling. Unlike encodings designed for large-scale scenes, JMHE is tailored for small, fine-structured, and highly specular objects such as surgical instruments, effectively capturing both high-frequency geometric details and view-dependent appearance while maintaining low computational overhead.

As illustrated in Fig. 2, for a 3D sample point $\mathbf{x} \in \mathbb{R}^3$ along a target ray $r = (o, \mathbf{d})$, where o is the ray origin and $\mathbf{d} \in \mathbb{S}^2$ denotes the normalized viewing direction, we convert $\mathbf{d}$ into spherical coordinates and discretize it into $T \times P$ angular bins. With polar and azimuth indices $i_\theta \in \{0, \ldots, T-1\}$ and $i_\phi \in \{0, \ldots, P-1\}$, the quantized direction bin is expressed as:

$$b(\mathbf{d}) = i_\theta P + i_\phi. \quad (8)$$

At each resolution level $\ell \in \{0, \ldots, L-1\}$, the grid resolution grows geometrically as:

$$N_\ell = \left\lfloor N_0\, b^{\,\ell} \right\rfloor, \qquad b = \exp\left(\frac{\ln N_{\max} - \ln N_0}{L-1}\right), \quad (9)$$

where N_0 and $N_{\max}$ denote the base and maximum resolutions, respectively, and L is the total number of levels. To couple spatial and directional information, each voxel corner with integer coordinates (c_x, c_y, c_z) is concatenated with the direction bin $b_d := b(\mathbf{d})$ to form a joint 4D hash key $k = (c_x, c_y, c_z, b_d)$. The hash index h is computed as:

$$h = \big(p_x c_x \oplus p_y c_y \oplus p_z c_z \oplus p_b b_d\big) \bmod H, \quad (10)$$

where H is the hash table size, $\oplus$ denotes bitwise XOR, and (p_x, p_y, p_z, p_b) are fixed prime constants used to decorrelate dimensions.

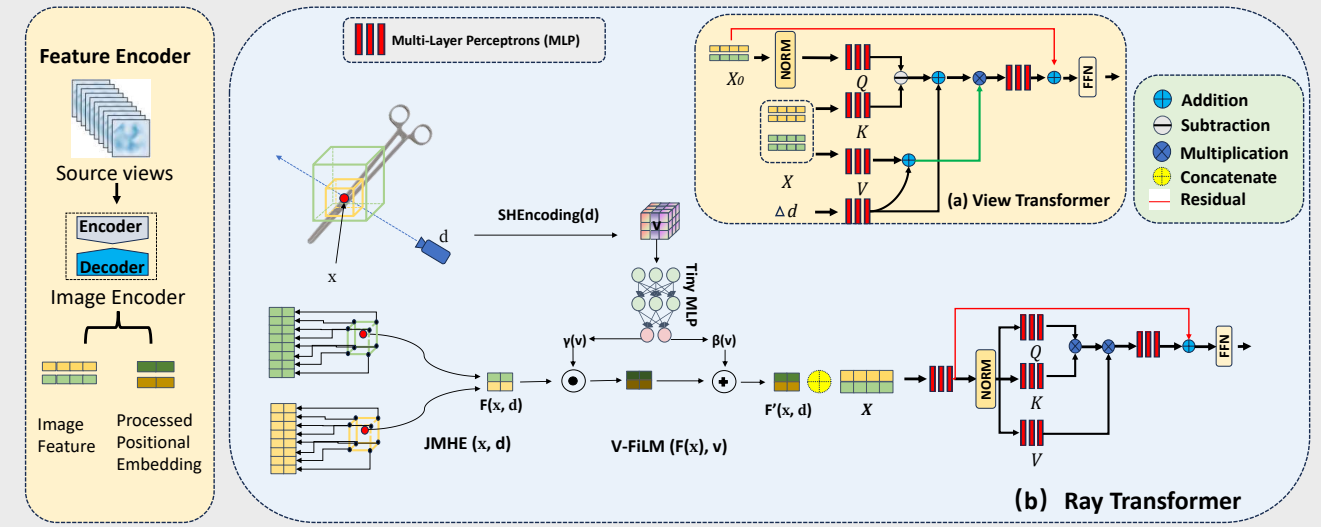

Figure 1. **Overview of GNFM framework.** The pipeline follows the Generalizable NeRF Transformer (GNT) paradigm and consists of two stages: (a) the **View Transformer**, which aggregates coordinate-aligned epipolar features X from source views under geometric constraints, and updates the initial aggregated ray features X_0; (b) the **Ray Transformer**, which refines point-wise features along each ray using the joint spatial and directional features, encoded by our JMHE and V-FiLM modules, together with contextual features from the View Transformer, to predict the target pixel color. GNFM augments the encoding stage with JMHE and V-FiLM to enrich intermediate representations with spatial and directional interactions and smooth view-dependent modulation.

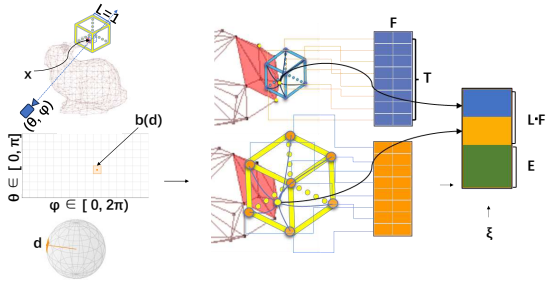

Figure 2. Illustration of the Joint Multiresolution Hash Encoding. (1) For a sample point x and its viewing direction d, we convert d into spherical coordinates and quantize it into angular bins. The spatial voxel corners around x with coordinates (c_x, c_y, c_z) are combined with the direction bin b_d to form a 4D joint key. (2) Using hash indices, we retrieve F-dimensional feature vectors from learnable hash tables and apply trilinear interpolation to obtain continuous features. (3) The outputs across all resolution levels are concatenated into the final multi-resolution position–direction embedding.

For each level ℓ, we maintain a learnable embedding table $E_\ell \in \mathbb{R}^{H \times F}$, and fetch corner embeddings as:

$$\mathrm{emb}(u) = E_\ell[h(\mathrm{corner}(u))], \quad u \in \{0,1\}^3, \tag{11}$$

where u indexes one of the eight voxel corners. We then apply trilinear interpolation to obtain a continuous direction-conditioned feature:

$$y_\ell = \sum_{u \in \{0,1\}^3} w(u)\,\mathrm{emb}(u), \tag{12}$$

where $w(u)$ are interpolation weights determined by the fractional offset of $\mathbf{x}$ within the voxel cell. Finally, concatenating features across all resolution levels yields the multi-resolution embedding:

$$y = [\,y_0;\, y_1;\, \ldots;\, y_{L-1}\,], \qquad \dim(y) = L\,F. \tag{13}$$

This hierarchical design allows coarse levels to capture global spatial context, while finer levels preserve high-frequency geometric details. The joint position and direction hashing guarantees distinct embeddings for the same spatial position under different viewpoints, enabling effective modeling of specular and anisotropic surface effects. Notably, the direction component participates only in the indexing process, ensuring computational efficiency. Both lookup and interpolation are fully differentiable, and all hash tables are optimized end-to-end through backpropagation. In practice, JMHE enhances spatial and directional interactions with negligible memory and latency overhead, making it highly suitable for reconstructing small, highly reflective surgical instruments.

3.4. View-Conditional Feature-wise Linear Modulation

This section introduces **View-Conditional Feature-wise Linear Modulation (V-FiLM)**, a feature-wise modulation mechanism that conditions positional hash features

on viewing direction through Feature-wise Linear Modulation (FiLM) [25]. Instead of directly concatenating view and positional features, V-FiLM applies view-conditioned affine transformations to each feature channel. This design produces smooth, continuous, and fully differentiable responses to view-dependent effects, including highlights, specular reflections, and anisotropy, while keeping the parameter count low. The overall structure of V-FiLM is illustrated in Fig. 3.

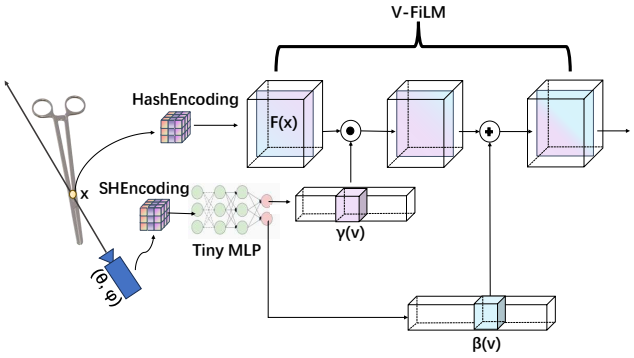

Figure 3. **V-FiLM.** A sample position x is encoded by *HashEncoding* into $F(x)$, while the view direction d is encoded via *SHEncoding* and fed into a lightweight MLP to generate FiLM parameters $\gamma(v)$ and $\beta(v)$. These parameters are then applied element-wise to $F(x)$, yielding view-dependent modulated features.

For each sampled point, the positional feature $F(\mathbf{x})$ is derived from the multi-resolution hash encoder, and the normalized view direction is encoded via spherical harmonics into $\mathbf{v}$. A two-layer MLP maps $\mathbf{v}$ to per-channel scale and bias parameters $\gamma(\mathbf{v})$ and $\beta(\mathbf{v})$, which modulate $F(\mathbf{x})$ as:

$$\mathbf{y} = \gamma(\mathbf{v}) \odot F(\mathbf{x}) + \beta(\mathbf{v}), \tag{14}$$

where $\odot$ denotes element-wise multiplication. The modulation is applied only to the color branch to avoid ambiguities in geometry estimation, while the density branch remains view-independent to preserve stable shape reconstruction.

V-FiLM introduces only a lightweight computational overhead: two linear layers generate twice the dimension of the positional encoding parameters for per-channel affine modulation. This design is substantially more efficient than concatenating direction and position features and increasing network width.

V-FiLM complements JMHE: while JMHE captures discrete position–direction interactions but may suffer from binning artifacts, V-FiLM provides continuous, differentiable modulation that enhances the smoothness and fidelity of view-dependent reflections. In combination, JMHE captures localized high-frequency effects, whereas V-FiLM provides smooth feature recalibration across viewpoints.

In summary, V-FiLM provides a parameter-efficient and smoothly responsive mechanism for view-dependent feature modulation. By applying conditional affine transformations to positional features, it significantly enhances the rendering fidelity of reflective and anisotropic materials while introducing negligible computational overhead, and integrates seamlessly with JMHE to further improve robustness and cross-view reconstruction quality.

3.5. Integration and Training

JMHE and V-FiLM are integrated into the feature encoding and representation construction stages, jointly enhancing the interactions between spatial and directional features and ensuring smooth view-dependent modulation with negligible computational overhead. All components, including the image encoder, transformers, JMHE, V-FiLM, and MLP heads, are trained end-to-end using standard reconstruction and perceptual losses. Importantly, the overall pipeline preserves GNT's two stage design while benefiting from enhanced feature representations that better capture high-frequency details and view-dependent appearance.

Furthermore, during feature generation, directional information is already embedded into positional features through JMHE and V-FiLM. As a result, the resulting feature representations inherently encode view-dependent cues. Consequently, our GNFM framework no longer requires explicit concatenation of directional encodings, yet it still achieves significant improvements over prior methods. This advantage is particularly evident in our dedicated dataset of highly reflective surgical instruments, where the model captures reflection cues more accurately and stably. Moreover, this design substantially reduces the parameter count without compromising performance.

4. Experiments

4.1. Dataset Construction

Accurate 3D reconstruction of reflective surgical instruments is a critical challenge in robot-assisted surgery, where safe manipulation and intelligent perception depend on reliable visual representations. However, existing neural rendering benchmarks are dominated by toy scenes or household objects, which fail to capture the small scale, intricate geometry, and strong specular reflections of real surgical tools. To address this gap, we construct the *Reflective Surgical Instrument Dataset (RSID)*, a high-fidelity benchmark tailored for novel view synthesis of reflective surgical instruments.

Synthetic Data. The synthetic subset of RSID is created in Blender [2], with each instrument modeled and rendered as an independent scene. It covers eight representative surgical instruments, including *Haemostatic Clamp*, *Curved Needle Holder*, *DeBakey–Cooley Forcep*, *Dissecting Forcep*, *O-ring Forcep*, *Scalpel*, *Surgical Blades*, and *Umbil-*

ical Cord Scissor. For every instrument, we generate 100 training views, 100 validation views, and 100 test views at 800×800 resolution, yielding a total of 2,400 multi-view images with corresponding camera pose files.

Real Data. The real-world subset of RSID consists of six physical surgical instruments captured under realistic imaging conditions. Images are acquired using a handheld camera, with each image recorded at a resolution of 1280×720. This subset complements the synthetic data by introducing real-world illumination variations, sensor noise, and complex specular reflections that are difficult to accurately model in simulation.

Unlike conventional NeRF datasets that adopt a fixed-radius camera setting unrelated to real-world scale, RSID preserves the true physical dimensions of instruments and positions cameras at clinically realistic working distances ranging from $0.3\,\mathrm{m}$ to $0.5\,\mathrm{m}$, consistent with robot-assisted surgical environments. This design ensures that the dataset stresses both geometric fidelity and photometric complexity while remaining aligned with practical requirements of robot-assisted surgery.

4.2. Evaluation Metrics

To comprehensively evaluate the performance of GNFM on reflective surgical instruments, we adopt Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM)[40], and the Learned Perceptual Image Patch Similarity (LPIPS)[51], which together capture complementary aspects of novel view synthesis quality. PSNR provides a pixel-level measure of reconstruction fidelity, indicating how closely the synthesized views match the ground truth. However, this metric alone is insufficient in our context, as strong specularities often cause large pixel discrepancies that do not necessarily reflect perceptual quality. SSIM addresses this limitation by evaluating structural similarity in terms of luminance, contrast, and spatial consistency, making it particularly relevant for assessing the fine grooves and sharp edges of surgical tools. Since realistic rendering of reflections is central to our task, we further employ LPIPS, which leverages deep neural network features to approximate human perceptual judgments of visual realism. By jointly considering these metrics, we show that GNFM not only achieves higher numerical accuracy but also delivers reconstructions with greater structural fidelity and perceptual plausibility compared to existing methods.

4.3. Experimental Setup

All experiments were conducted on a workstation equipped with a single NVIDIA RTX 4060 Ti GPU (16GB VRAM), an Intel Core i5-12400F CPU, and 32GB of RAM. The software environment was Ubuntu 22.04 with CUDA 12.4 installed. We compared our GNFM framework against

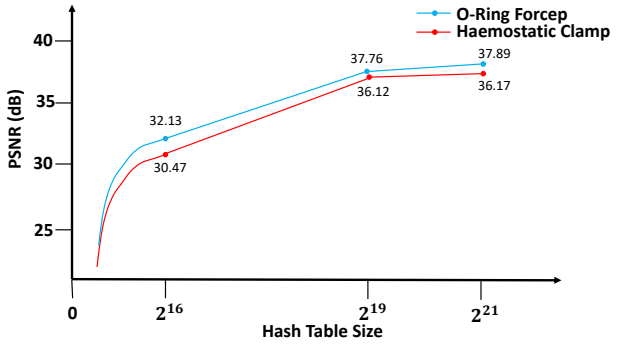

Figure 4. Ablation on hash table size T for the Joint Multiresolution Hash Encoder (JMHE). Reconstruction quality improves notably up to $T{=}2^{19}$ and saturates thereafter.

five representative baselines: Mip-NeRF, Ref-NeRF, 3DGS, IBRNet, and GNT. These methods were chosen for their demonstrated performance in novel view synthesis and their relevance to specular surface rendering. For fairness, we adopted the official implementations of all baselines and retained default hyperparameter configurations.

To ensure consistent evaluation, all methods were trained with 1,024 rays per iteration. For the baselines, each ray was sampled with a standard coarse-to-fine strategy, whereas our GNFM simplified the process by sampling only 64 points per ray without hierarchical refinement. Despite the reduced sampling budget, this design leverages our feature modulation mechanism to achieve both efficiency and accuracy. For the Joint Multiresolution Hash Encoder (JMHE), we set the hash table capacity to $T = 2^{19}$, following an internal ablation similar to Fig. 4, where we compared $T \in \{2^{16}, 2^{19}, 2^{21}\}$ on the *Haemostatic Clamp* and *O-Ring Forcep* scenes. We observed that reconstruction quality PSNR improved sharply up to $T = 2^{19}$ but saturated thereafter, while GPU memory increased substantially beyond this point. Thus, $T = 2^{19}$ offers the best balance between fidelity and efficiency on our RTX 4060 Ti GPU. The initial learning rate was set to 5×10^{-4} with a cosine decay schedule, and optimization was performed using Adam. Each scene was trained for 250,000 iterations to ensure convergence, while cross-scene generalization experiments were trained for 300,000 iterations to account for the increased variability across unseen instruments.

We evaluate GNFM on our RSID dataset. Experiments are conducted in two regimes: (1) single-scene training, where the model is trained on one instrument and evaluated on unseen viewpoints of the same object to assess novel view synthesis quality; and (2) cross-scene generalization, where four instruments are randomly selected for training and the remaining ones are used as unseen test objects. For cross-scene experiments, we report both direct generalization performance and results after lightweight finetuning on the target instrument. Quantitative and qualitative evalua-

Table 1. PSNR Comparison on Reflective Surgical Instruments Dataset in the Single-Scene Setting.

	PSNR↑				
	Mip-NeRF	Ref-NeRF	GNT	3DGS	GNFM(Ours)
Haemostatic Clamp	34.08	33.22	33.87	34.43	**36.12**
Curved Needle Holder	33.85	32.86	33.24	35.04	**37.23**
DeBakey-Cooley Forcep	32.43	31.30	32.43	33.68	**34.13**
Dissecting Forcep	34.11	32.06	33.83	33.68	**35.61**
O-Ring Forcep	33.21	32.14	36.51	34.08	**37.76**
Scalpel	37.23	33.96	34.11	36.64	**38.23**
Surgical Blades	32.35	28.68	33.76	31.27	**36.21**
Umbilical Cord Scissor	31.04	31.09	32.15	31.91	**34.19**

Table 2. SSIM Comparison on Reflective Surgical Instruments Dataset in the Single-Scene Setting.

	SSIM↑				
	Mip-NeRF	Ref-NeRF	GNT	3DGS	GNFM(Ours)
Haemostatic Clamp	0.9879	0.9891	0.9887	0.9632	**0.9941**
Curved Needle Holder	0.9880	0.9897	0.9863	0.9640	**0.9957**
DeBakey-Cooley Forcep	0.9822	0.9832	0.9821	0.9611	**0.9920**
Dissecting Forcep	0.9821	0.9800	0.9865	0.9594	**0.9963**
O-Ring Forcep	0.9819	0.9831	0.9903	0.9608	**0.9907**
Scalpel	0.9922	0.9897	0.9862	0.9655	**0.9928**
Surgical Blades	0.9757	0.9605	0.9817	0.9563	**0.9904**
Umbilical Cord Scissor	0.9719	0.9713	0.9819	0.9564	**0.9876**

Table 3. LPIPS Comparison on Reflective Surgical Instruments Dataset in the Single-Scene Setting.

	LPIPS↓				
	Mip-NeRF	Ref-NeRF	GNT	3DGS	GNFM(Ours)
Haemostatic Clamp	0.0235	0.0162	0.0192	0.0239	**0.0075**
Curved Needle Holder	0.0241	0.0155	0.0151	0.0226	**0.0059**
DeBakey-Cooley Forcep	0.0315	0.0246	0.0237	0.0271	**0.0137**
Dissecting Forcep	0.0286	0.0247	0.0319	0.0298	**0.0094**
O-Ring Forcep	0.0302	0.0217	0.0223	0.0272	**0.0183**
Scalpel	0.0198	0.0263	0.0245	0.0210	**0.0174**
Surgical Blades	0.0493	0.0469	0.0225	0.0412	**0.0217**
Umbilical Cord Scissor	0.0460	0.0450	0.0389	0.0351	**0.0183**

tions were conducted to assess reconstruction fidelity and visual realism, including both numerical metrics and rendered visual comparisons from challenging viewpoints. We also perform ablation studies to analyze the independent contributions of JMHE and V-FiLM to the overall performance.

4.4. Single Scene Results

We first evaluate model performance under the single-scene setting, where each method is trained on a single instrument and evaluated on novel viewpoints of the same object. Unless otherwise specified, the results reported in this subsection are obtained on the synthetic subset of RSID.

Synthetic Single-Scene Results. Quantitative results on the synthetic subset of RSID are summarized in Table 1, Table 2, and Table 3. GNFM achieves consistent improvements across PSNR, SSIM, and LPIPS compared with all baselines. Notably, the gains on LPIPS are substantially larger than those on PSNR and SSIM. This behavior is expected in reflective scenes: PSNR and SSIM are dominated by pixel-wise alignment and low-frequency errors, so even small shifts in specular peaks can penalize them when the perceived appearance is already correct. In contrast, LPIPS measures perceptual similarity in a learned feature space that correlates better with human visual perception, and is more sensitive to coherent reproduction of high-frequency cues such as highlights and micro-structures.

By jointly encoding spatial and directional information with JMHE and applying view-conditioned feature modulation via V-FiLM, GNFM renders specular highlights and fine details more consistently across viewpoints, leading to markedly lower LPIPS while delivering stable improvements in PSNR and SSIM. Unlike existing approaches, which often struggle with view-dependent reflections and fine-scale geometry, GNFM effectively captures specular highlights while maintaining structural fidelity and perceptual realism.

To further illustrate these advantages, Fig. 5 presents visual comparisons on representative instruments from the synthetic subset. GNFM produces sharper edges, more faithful reflections, and reduced artifacts compared with Mip-NeRF, Ref-NeRF, and GNT, highlighting its robustness under challenging reflective conditions. While GNT occasionally achieves competitive results in certain cases, its performance is less stable across different instruments. In contrast, GNFM delivers consistent gains, confirming the effectiveness of integrating JMHE and V-FiLM for handling specular surfaces in surgical instrument rendering.

Real-World Single-Scene Results. We further evaluate GNFM on the real-world subset of RSID under the single-scene setting. Each model is trained and evaluated on images captured from a single real surgical instrument. Quantitative results are reported in Table 4, and qualitative visualizations in Fig. 6 demonstrate improved reconstruction fidelity and view-dependent reflection modeling.

4.5. Cross-scene Generalization to Unseen Instruments

To comprehensively evaluate the generalization capability of GNFM, we perform cross-scene experiments across surgical instruments on both synthetic and real-world datasets. The synthetic setting enables controlled evaluation under consistent geometry, material properties, and illumination conditions, while the real-world setting further assesses robustness under practical challenges such as sensor noise, calibration inaccuracies, and complex lighting variations.

Synthetic Cross-scene Setting. In the synthetic experiments, four instruments, namely *Haemostatic Clamp*, *Curved Needle Holder*, *DeBakey–Cooley Forcep*, and *Dissecting Forcep*, are used for training. The remaining four instruments, including *O-Ring Forcep*, *Scalpel*, *Surgical Blades*, and *Umbilical Cord Scissor*, are reserved as unseen test cases. All models, including GNFM and two represen-

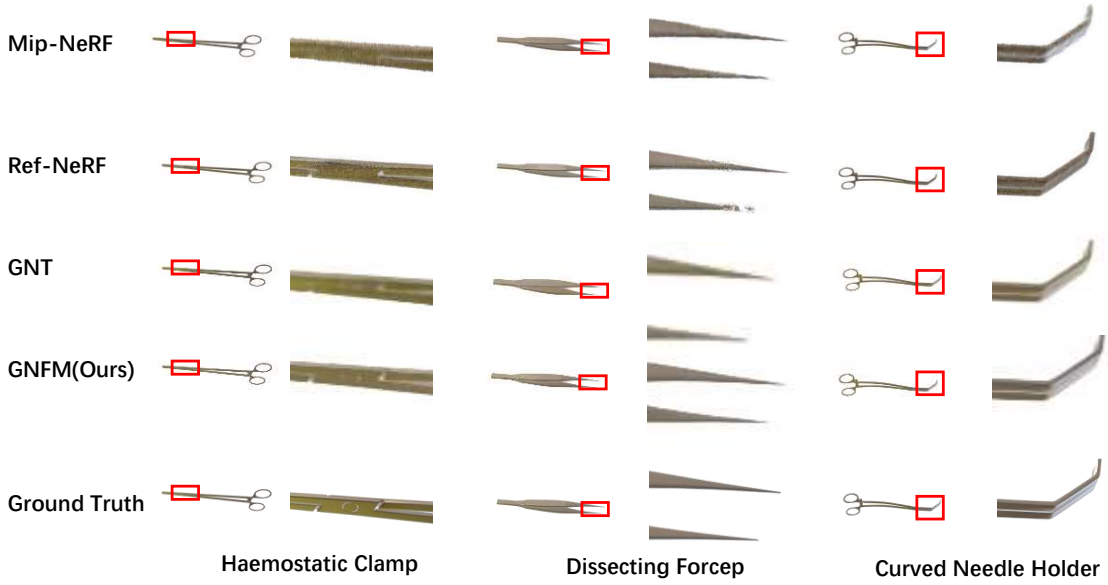

Figure 5. Visualization results of representative instruments under synthetic single-scene setting.

Table 4. Quantitative results on the real-world RSID dataset. Higher PSNR and SSIM indicate better performance, while lower LPIPS is preferred.

Scene	GNT			3DGS			GNFM (Ours)		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Adson Forceps	30.50	0.9337	0.1989	32.39	0.9483	0.1643	**33.41**	**0.9621**	**0.1532**
Bayonet Forceps	36.47	0.9540	0.2345	**38.52**	0.9737	**0.1842**	37.97	**0.9801**	0.1876
Hemostat	32.56	0.9683	0.2175	27.47	0.9545	0.2604	**34.73**	**0.9798**	**0.1556**
Scalpel	27.68	0.9291	0.2378	31.95	0.9618	0.1755	**33.42**	**0.9765**	**0.1632**
Scissors	33.77	0.9709	0.1562	34.34	0.9716	0.1447	**36.91**	**0.9804**	**0.1108**
Thumb Forceps	25.11	0.8775	0.2533	28.83	0.9552	0.2340	**32.45**	**0.9679**	**0.1973**

tative baselines, IBRNet and GNT, are trained under identical protocols to ensure a fair comparison.

The evaluation considers three complementary aspects: (i) *cross-scene training performance*, which measures reconstruction accuracy on instruments seen during training; (ii) *generalization to unseen instruments without finetuning*, which directly evaluates transferability to novel instruments without any adaptation; and (iii) *generalization to unseen instruments with finetuning*, which assesses the adaptability of the model when limited additional data are provided.

As shown in Table 5, GNFM consistently outperforms the baselines across all three evaluation settings. In particular, the performance gains are most significant in the unseen-without-finetuning scenario, highlighting the robustness of GNFM in handling reflective surgical instruments without relying on scene-specific optimization.

Qualitative Analysis on Synthetic Data. Fig. 7 and Fig. 8 further provide qualitative comparisons under the cross-scene setting. Without finetuning, GNFM reconstructs sharper geometry and preserves specular highlights more faithfully, while finetuning further improves detail fidelity and reflection consistency.

Real-world Cross-scene Generalization. We further evaluate GNFM on real-world RSID data following the same cross-scene protocol. Quantitative comparisons with GNT are reported in Table 6. In addition, qualitative visualizations in Fig. 9 show that GNFM achieves more stable geometry reconstruction and more consistent view-dependent appearance under real imaging conditions, confirming that the learned feature modulation generalizes beyond synthetic rendering.

Overall, these results demonstrate that GNFM effec-

Table 5. Quantitative results on cross-scene training and generalization on the synthetic dataset.

	Dataset	PSNR↑			SSIM↑			LPIPS↓		
		IBRNet	GNT	GNFM (Ours)	IBRNet	GNT	GNFM (Ours)	IBRNet	GNT	GNFM (Ours)
Cross-scenes (Training)	Haemostatic Clamp	31.17	33.12	**35.02**	0.9328	0.9814	**0.9902**	0.0236	0.0197	**0.0140**
	Curved Needle Holder	30.25	31.23	**35.59**	0.9423	0.9806	**0.9925**	0.0354	0.0253	**0.0115**
	DeBakey–Cooley Forcep	28.78	30.01	**33.93**	0.9191	0.9795	**0.9886**	0.0467	0.0288	**0.0165**
	Dissecting Forcep	29.18	30.63	**32.98**	0.9532	0.9811	**0.9828**	0.0397	0.0296	**0.0225**
Generalization (Unseen, w/o finetune)	O-Ring Forcep	25.67	28.15	**33.28**	0.9234	0.9801	**0.9859**	0.0563	0.0372	**0.0193**
	Scalpel	28.98	31.24	**36.64**	0.9513	0.9863	**0.9922**	0.0437	0.0230	**0.0123**
	Surgical Blades	23.03	27.56	**29.06**	0.9287	0.9431	**0.9678**	0.0569	0.0471	**0.0400**
	Umbilical Cord Scissor	25.96	27.81	**30.37**	0.9349	0.9717	**0.9757**	0.0619	0.0415	**0.0301**
Generalization (Unseen, w/ finetune)	O-Ring Forcep	29.06	32.15	**33.57**	0.9577	0.9801	**0.9863**	0.0321	0.0276	**0.0184**
	Scalpel	33.12	36.24	**37.08**	0.9875	0.9907	**0.9926**	0.0421	0.0213	**0.0119**
	Surgical Blades	28.47	29.56	**32.50**	0.9413	0.9657	**0.9814**	0.0456	0.0296	**0.0264**
	Umbilical Cord Scissor	29.54	30.81	**32.17**	0.9325	0.9674	**0.9808**	0.0397	0.0369	**0.0238**

Table 6. Quantitative results on real-world cross-scene training and generalization.

	Scene	PSNR↑		SSIM↑		LPIPS↓	
		GNT	GNFM (Ours)	GNT	GNFM (Ours)	GNT	GNFM (Ours)
Cross-scenes (Training)	Adson Forceps	29.16	**33.17**	0.9283	**0.9537**	0.2043	**0.1621**
	Bayonet Forceps	36.52	**37.76**	0.9537	**0.9800**	0.2542	**0.1895**
	Hemostat	29.47	**32.36**	0.9445	**0.9736**	0.2404	**0.1793**
Generalization (Unseen, w/o finetune)	Scalpel	27.95	**31.95**	0.9318	**0.9342**	0.2655	**0.2224**
	Scissors	30.34	**32.41**	0.9616	**0.9689**	0.1847	**0.1450**
	Thumb Forceps	28.83	**30.56**	0.8752	**0.9115**	0.2340	**0.2238**
Generalization (Unseen, w/ finetune)	Scalpel	27.86	**32.19**	0.9187	**0.9548**	0.2558	**0.1433**
	Scissors	33.65	**35.92**	0.9679	**0.9754**	0.1783	**0.1395**
	Thumb Forceps	28.71	**31.39**	0.8913	**0.9469**	0.2234	**0.2141**

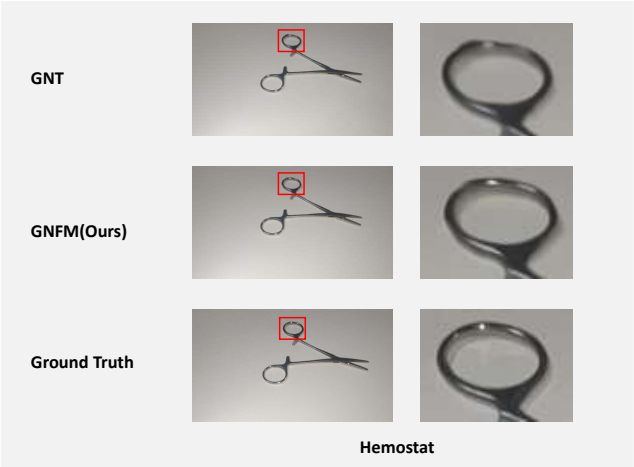

Figure 6. Visualization results of representative instruments under real-world single-scene setting.

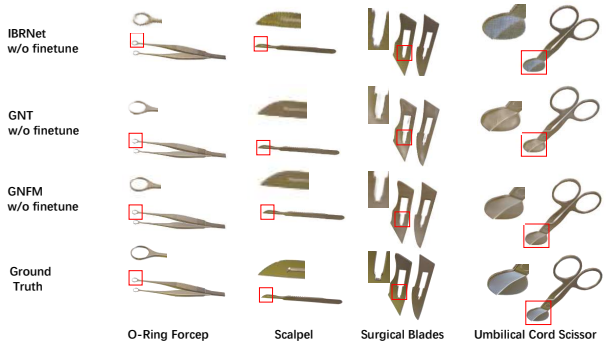

Figure 7. Visualization results of unseen instruments under the synthetic cross-scene setting without finetuning.

tively addresses three key challenges: (i) reconstructing fine-grained surgical instruments, (ii) modeling strong view-dependent specular reflections, and (iii) enabling robust generalization to unseen instruments.

4.6. Ablation studies

To further understand the contribution of each component in GNFM, we conduct a set of ablation studies by gradually introducing our proposed modules on top of the GNT baseline. We report the averaged results across single-scene training, generalization to unseen instruments without finetuning, and generalization to unseen instruments with finetuning. For each setting, PSNR, SSIM, and LPIPS are computed and averaged to avoid excessive numerical clutter. The comparison is summarized in Table 7.

Starting from the GNT baseline, the performance is limited, especially in challenging reflective regions where geometry and appearance interact in a highly view-dependent manner. When we add the Joint Multi-scale Hash Encod-

Table 7. Ablation study of GNFM. We report average PSNR ↑, SSIM ↑, and LPIPS ↓ under single-scene and generalization settings.

Method	Single-scene Avg.	Generalization (Unseen, w/o finetune)	Generalization (Unseen, w/ finetune)
	PSNR↑ / SSIM↑ / LPIPS↓	PSNR↑ / SSIM↑ / LPIPS↓	PSNR↑ / SSIM↑ / LPIPS↓
GNT (Baseline)	33.74 / 0.9854 / 0.0247	28.69 / 0.9703 / 0.0347	32.19 / 0.9759 / 0.0288
+ JMHE	34.43 / 0.9897 / 0.0192	31.17 / 0.9771 / 0.0301	32.14 / 0.9823 / 0.0277
+ V-FiLM	35.01 / 0.9909 / 0.0187	30.23 / 0.9799 / 0.0276	32.59 / 0.9837 / 0.0226
Full (GNFM)	**36.19 / 0.9925 / 0.0140**	**32.33 / 0.9804 / 0.0254**	**33.83 / 0.9853 / 0.0201**

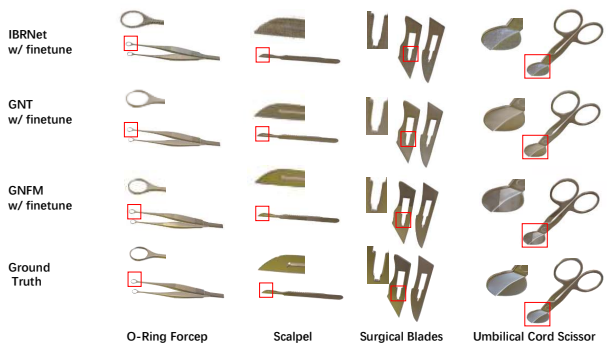

Figure 8. Visualization results of unseen instruments under the synthetic cross-scene setting with finetuning.

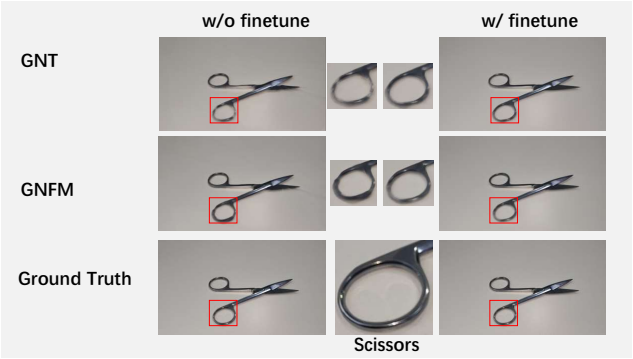

Figure 9. Visualization results of unseen instruments under the real-world cross-scene setting.

ing (JMHE), the model achieves clear improvements in both single-scene and cross-scene evaluations. The reason lies in JMHE's ability to capture fine-grained geometric structures through multi-resolution features, which is particularly beneficial for small surgical instruments with subtle details.

When the View-Conditional Feature-wise Linear Modulation (V-FiLM) module is introduced independently, the network already shows clear improvements in generalization tasks. Unlike JMHE, which focuses on structural representation, V-FiLM enables the model to dynamically adjust feature distributions according to the viewing direction, allowing it to better handle specular highlights and appearance variations across different poses. This effect is most evident in the generalization to unseen instruments without finetuning, where GNFM with only V-FiLM significantly surpasses the baseline, demonstrating that view-aware feature modulation alone substantially enhances robustness even without additional optimization.

Finally, the full GNFM framework, which integrates both JMHE and V-FiLM, achieves the best performance consistently across all evaluation settings. It not only preserves detailed structures in single-scene training but also provides strong adaptability in cross-scene scenarios. With finetuning on unseen instruments, GNFM further consolidates its superiority, indicating that the two components are complementary: JMHE strengthens geometric fidelity, while V-FiLM addresses view-dependent reflectance. Together, they form a unified architecture that effectively solves the dual challenge of small object reconstruction and specular reflection modeling.

5. Conclusion

In robot-assisted surgery, accurate 3D reconstruction of small-scale, highly reflective surgical instruments is essential for ensuring safe operations and intelligent perception. However, this task is extremely challenging due to the tiny size and intricate structures of the instruments, as well as their strong specular reflections, where conventional NeRF methods often fail to recover fine details or model view-dependent appearances. To address these issues, we propose Generalizable NeRF with View-Aware Feature Modulation (GNFM), which employs a Joint Multiresolution Hash Encoder (JMHE) to enhance spatial detail representation and integrates a View-Conditional Feature-wise Linear Modulation (V-FiLM) to capture severe view-dependent appearance variations. Unlike methods relying on material priors, GNFM is entirely data-driven and achieves accurate modeling of reflective instruments without additional assumptions. To further validate its generalization and practicality, we built the Reflective Surgical Instrument Dataset (RSID), a dedicated dataset of real rendered data covering various reflective surgical instruments. Experimental results demonstrate that GNFM significantly outperforms existing methods in small-scale reconstruction, fine detail recovery, and handling of reflective scenes, providing strong support for intelligent perception in robot-assisted surgery.

Acknowledgements

This work was supported in part by the Academy for Advanced Interdisciplinary Studies, Nankai University; in part by the Tianjin Science and Technology Program under the project entitled Research and Application of Key Technologies for Embodied Intelligent Robots in Power Operations; and in part by the National Natural Science Foundation of China under Grants 62293513/62106109/62573243.

References

[1] J. T. Barron, B. Mildenhall, M. Tancik, P. Hedman, R. Martin-Brualla, and P. P. Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5855–5864, 2021. 2, 3

[2] J. M. Blain. *The complete guide to Blender graphics: computer modeling & animation*. AK Peters/CRC Press, 2019. 7

[3] D. Bouget, R. Benenson, M. Omran, L. Riffaud, B. Schiele, and P. Jannin. Detecting surgical tools by modelling local appearance and global shape. *IEEE transactions on medical imaging*, 34(12):2603–2617, 2015. 1

[4] S. Fridovich-Keil, A. Yu, M. Tancik, Q. Chen, B. Recht, and A. Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5501–5510, 2022. 3

[5] Y. Furukawa, C. Hernández, et al. Multi-view stereo: A tutorial. *Foundations and trends® in Computer Graphics and Vision*, 9(1-2):1–148, 2015. 1, 3

[6] S. J. Garbin, M. Kowalski, M. Johnson, J. Shotton, and J. Valentin. Fastnerf: High-fidelity neural rendering at 200fps. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14346–14355, 2021. 2

[7] J. Geng. Structured-light 3d surface imaging: a tutorial. *Advances in optics and photonics*, 3(2):128–160, 2011. 3

[8] X. Gu, Z. Fan, S. Zhu, Z. Dai, F. Tan, and P. Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2495–2504, 2020. 3

[9] L. V. Holland, R. Bliersbach, J. U. Müller, P. Stotko, and R. Klein. Tram-nerf: Tracing mirror and near-perfect specular reflections through neural radiance fields. In *Computer Graphics Forum*, volume 43, page e15163. Wiley Online Library, 2024. 2, 3

[10] X. Huang, C. Wu, X. Xu, B. Wang, S. Zhang, C. Shen, C. Yu, J. Wang, N. Chi, S. Yu, et al. Polarization structured light 3d depth image sensor for scenes with reflective surfaces. *Nature Communications*, 14(1):6855, 2023. 2

[11] B. Kaya, S. Kumar, C. Oliveira, V. Ferrari, and L. Van Gool. Uncertainty-aware deep multi-view photometric stereo. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12601–12611, 2022. 2, 3

[12] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 3

[13] C. Lei, C. Qi, J. Xie, N. Fan, V. Koltun, and Q. Chen. Shape from polarization for complex scenes in the wild. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 12632–12641, 2022. 3

[14] H. Li, W. Yan, D. Liu, L. Qian, Y. Yang, Y. Liu, Z. Zhao, H. Ding, and G. Wang. Evd surgical guidance with retroreflective tool tracking and spatial reconstruction using head-mounted augmented reality device. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 1

[15] L. Ma, V. Agrawal, H. Turki, C. Kim, C. Gao, P. Sander, M. Zollhöfer, and C. Richardt. Specnerf: Gaussian directional encoding for specular reflections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21188–21198, 2024. 2, 3

[16] R. R. Madavan, A. Kaimal, V. Gupta, R. Choudhary, C. Shanmuganathan, K. Mitra, et al. Ganesh: Generalizable nerf for lensless imaging. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 9499–9508. IEEE, 2025. 4

[17] R. Martin-Brualla, N. Radwan, M. S. Sajjadi, J. T. Barron, A. Dosovitskiy, and D. Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7210–7219, 2021. 3

[18] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2, 3

[19] M. M. Morita, D. A. L. Carvajal, I. L. G. Bagur, and G. M. Bilmes. A combined approach of sfm-mvs photogrammetry and reflectance transformation imaging to enhance 3d reconstructions. *Journal of Cultural Heritage*, 68:38–46, 2024. 3

[20] M. Muglikar, L. Bauersfeld, D. P. Moeys, and D. Scaramuzza. Event-based shape from polarization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1547–1556, 2023. 3

[21] T. Müller, A. Evans, C. Schied, and A. Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4):1–15, 2022. 3

[22] Z.-L. Ni, G.-B. Bian, Z.-G. Hou, X.-H. Zhou, X.-L. Xie, and Z. Li. Attention-guided lightweight network for real-time segmentation of robotic surgical instruments. In *2020 IEEE international conference on robotics and automation (ICRA)*, pages 9939–9945. IEEE, 2020. 1

[23] M. Oechsle, S. Peng, and A. Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5589–5599, 2021. 3

[24] K. Peng, R. Islam, J. Quarles, and K. Desai. Tmvnet: Using transformers for multi-view voxel-based 3d reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 222–230, 2022. 1, 3

[25] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. 7

[26] C. Reiser, S. Peng, Y. Liao, and A. Geiger. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14335–14345, 2021. 3

[27] K. Rematas, A. Liu, P. P. Srinivasan, J. T. Barron, A. Tagliasacchi, T. Funkhouser, and V. Ferrari. Urban radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12932–12942, 2022. 3

[28] L. Sestini, B. Rosa, E. De Momi, G. Ferrigno, and N. Padoy. Fun-sis: A fully unsupervised approach for surgical instrument segmentation. *Medical Image Analysis*, 85:102751, 2023. 1

[29] J. Shi, X. Ying, R. Guo, B. Xing, and W. Yue. Normalnerf: Ambiguity-robust normal estimation for highly reflective scenes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 6869–6877, 2025. 3

[30] A. Tewari, J. Thies, B. Mildenhall, P. Srinivasan, E. Tretschk, W. Yifan, C. Lassner, V. Sitzmann, R. Martin-Brualla, S. Lombardi, et al. Advances in neural rendering. In *Computer Graphics Forum*, volume 41, pages 703–735. Wiley Online Library, 2022. 2

[31] J. Thies, M. Zollhöfer, and M. Nießner. Deferred neural rendering: Image synthesis using neural textures. *Acm Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 2

[32] D. Verbin, P. Hedman, B. Mildenhall, T. Zickler, J. T. Barron, and P. P. Srinivasan. Ref-nerf: Structured view-dependent appearance for neural radiance fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5481–5490. IEEE, 2022. 2, 3

[33] D. Verbin, P. P. Srinivasan, P. Hedman, B. Mildenhall, B. Attal, R. Szeliski, and J. T. Barron. Nerf-casting: Improved view-dependent appearance with consistent reflections. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–10, 2024. 3

[34] P. Wang, X. Chen, T. Chen, S. Venugopalan, Z. Wang, et al. Is attention all that nerf needs? *arXiv preprint arXiv:2207.13298*, 2022. 2, 3, 4

[35] P. Wang, L. Liu, Y. Liu, C. Theobalt, T. Komura, and W. Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 3

[36] Q. Wang, Z. Wang, K. Genova, P. P. Srinivasan, H. Zhou, J. T. Barron, R. Martin-Brualla, N. Snavely, and T. Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2021. 2, 3

[37] T. Wang and V. J. Gan. Enhancing 3d reconstruction of textureless indoor scenes with indoreal multi-view stereo (mvs). *Automation in Construction*, 166:105600, 2024. 3

[38] X. Wang, C. Wang, B. Liu, X. Zhou, L. Zhang, J. Zheng, and X. Bai. Multi-view stereo in the deep learning era: A comprehensive review. *Displays*, 70:102102, 2021. 1, 3

[39] Y. Wang, B. Gong, Y. Long, S. H. Fan, and Q. Dou. Efficient endonerf reconstruction and its application for data-driven surgical simulation. *International Journal of Computer Assisted Radiology and Surgery*, 19(5):821–829, 2024. 1

[40] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 2, 8

[41] R. J. Woodham. Photometric method for determining surface orientation from multiple images. *Optical engineering*, 19(1):139–144, 1980. 3

[42] T. Wu, J. Wang, X. Pan, X. XU, C. Theobalt, Z. Liu, and D. Lin. Voxurf: Voxel-based efficient and accurate neural surface reconstruction. In *The Eleventh International Conference on Learning Representations*. 1, 3

[43] Z. Wu, A. Schmidt, R. Moore, H. Zhou, A. Banks, P. Kazanzides, and S. E. Salcudean. Surgpose: a dataset for articulated robotic surgical tool pose estimation and tracking. *arXiv preprint arXiv:2502.11534*, 2025. 1

[44] H. Xu, A. Weld, C. Xu, A. Roddan, J. Cartucho, M. A. Karaoglu, A. Ladikos, Y. Li, Y. Li, D. Shen, et al. Surgripe challenge: Benchmark of surgical robot instrument pose estimation. *Medical Image Analysis*, page 103674, 2025. 1

[45] J. Yang, M. Yoo, J. Cho, and S. Kim. Learning confidence measure with transformer in stereo matching. *Pattern Recognition*, 157:110876, 2025. 3

[46] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)*, pages 767–783, 2018. 1, 3

[47] L. Yariv, J. Gu, Y. Kasten, and Y. Lipman. Volume rendering of neural implicit surfaces. *Advances in neural information processing systems*, 34:4805–4815, 2021. 3

[48] M. Yoshimura, M. M. Marinho, K. Harada, and M. Mitsuishi. Single-shot pose estimation of surgical robot instruments' shafts from monocular endoscopic images. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9960–9966. IEEE, 2020. 1

[49] A. Yu, R. Li, M. Tancik, H. Li, R. Ng, and A. Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5752–5761, 2021. 3

[50] A. Yu, V. Ye, M. Tancik, and A. Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4578–4587, 2021. 2, 3

[51] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 2, 8

[52] R. Zhang, P.-S. Tsai, J. E. Cryer, and M. Shah. Shape-from-shading: a survey. *IEEE transactions on pattern analysis and machine intelligence*, 21(8):690–706, 2002. 3

[53] D. Zhao, D. Lichy, P.-N. Perrin, J.-M. Frahm, and S. Sengupta. Mvpsnet: Fast generalizable multi-view photometric stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12525–12536, 2023. 2