RS-SpecSDF: Reflection-Supervised Surface Reconstruction and Material Estimation for Specular Indoor Scenes

Dongyu Chen
cdy24@mails.tsinghua.edu.cn

Haoxiang Chen
chx20@mails.tsinghua.edu.cn

Qunce Xu
quncexu@tsinghua.edu.cn

Tai-Jiang Mu
taijiang@tsinghua.edu.cn

Tsinghua University
No 30, Shuangqing road, Beijing

Abstract

Neural Radiance Field (NeRF) has achieved impressive 3D reconstruction quality using implicit scene representations. However, planar specular reflections pose significant challenges in the 3D reconstruction task. It is a common practice to decompose the scene into physically real geometries and virtual images produced by the reflections. However, current methods struggle to resolve the ambiguities in the decomposition process, because they mostly rely on mirror masks as external cues. They also fail to acquire accurate surface materials, which is essential for downstream applications of the recovered geometries. In this paper, we present RS-SpecSDF, a novel framework for indoor scene surface reconstruction that can faithfully reconstruct specular reflectors while accurately decomposing the reflection from the scene geometries and recovering the accurate specular fraction and diffuse appearance of the surface without requiring mirror masks. Our key idea is to perform reflection ray-casting and use it as supervision for the decomposition of reflection and surface material. Our method is based on an observation that the virtual image seen by the camera ray should be consistent with the object that the ray hits after reflecting off the specular surface. To leverage this constraint, we propose the Reflection Consistency Loss and Reflection Certainty Loss to regularize the decomposition. Experiments conducted on both our newly-proposed synthetic dataset and real-captured dataset demonstrate that our method achieves high-quality surface reconstruction and accurate material decomposition results without the need of mirror masks.

1. Introduction

3D surface reconstruction, as a basic tool for digital 3D content creation, aims to accurately capture surface shape, texture, and appearance from sensor data of the environment, facilitating detailed digital representation and analysis. The resulting 3D surfaces could be directly employed in various downstream applications including manufacturing, filming, gaming, VR/AR, etc. As NeRF[16] demonstrating impressive performance in 3D reconstruction and novel view synthesis tasks, surface reconstruction with neural implicit representation[27, 33, 11] has made significant progresses in recent years. However, challenges persist in reconstructing indoor scenes with planar specular reflections.

Indoor scenes often feature specular planar reflectors due to the prevalence of glasses, mirrors or artificial objects with glossy surfaces, such as polished wood or marbles. These planar reflectors present significant challenges for two main reasons. Firstly, the reflection inside a planar reflector obeys the multiview consistency that corresponds to the geometry of the reflection virtual image. This often leads reconstruction methods to inaccurately reconstruct the virtual image as real geometries[7, 35], hindering faithful representation of the actual specular surface. Secondly, modeling the appearance of specular reflections solely based on surface position and direction, as typical NeRF-like methods do, proves difficult. Specular reflections from planar surfaces exhibit significant variations in both spatial and directional domains, with high-frequency details mirroring those of the real scene. Thus interpolating such reflections accurately becomes challenging. Moreover, methods targeting non-planar reflectors[22, 40, 13, 4, 9] often approximate specular reflection from distant illumination using environment lighting, which is incompatible with the plane reflection. To address planar reflectors, a common practice is to decompose the reflection from the real scene using alternative spatial representations, such as additional radiance field[7] or multiple radiance fields[35], 3D Gaussians[12, 15], etc.

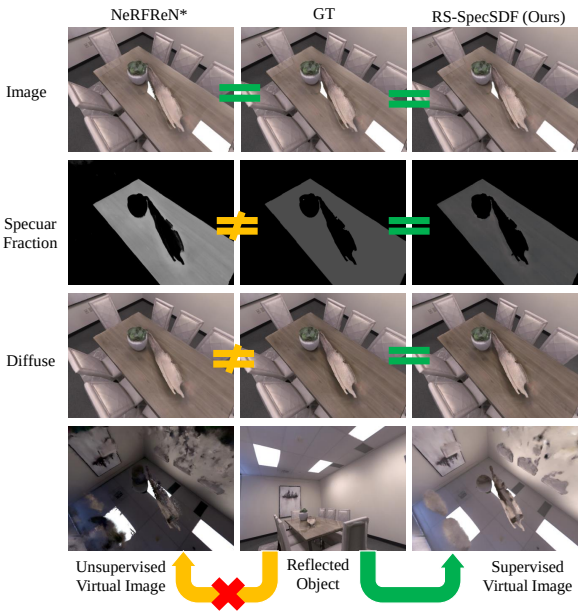

Figure 1. Current methods [7] handling specular planar reflection suffer from light and material ambiguity, leading to inaccurate decomposition of surface material and reflection virtual image even though the rendered image is close to ground truth. Our proposed RS-SpecSDF leverages the environment information (reflected object) as supervision, achieves through reflection ray-casting. Our approach effectively resolves light and material ambiguity.

However, the decomposition of reflection and real scene suffers from significant ambiguities that current methods still cannot resolve. There frequently exists **appearance ambiguity** when distinguishing whether the observed appearance belongs to the virtual image of reflected objects or non-reflective geometries. Thus, it is hard to estimate whether a specular surface exists only using RGB images as the supervision. Most current methods heavily rely on reflection masks[7, 12, 15, 38] to correctly perform decomposition. However, obtaining reflection masks for complex scenes is not straightforward and always requires tedious manual annotations or additional physical equipment during capture process[29]. While some methods[35, 41] are designed without reflection masks, they fail to achieve a complete decomposition between reflection and the real scene. On the other hand, as shown in Figure 1, the decomposition of reflection light and surface material (both the diffuse color and specular fraction) is still challenging. The **light and material ambiguity** exists when the planar surface is texture-less, which is a common scenario in real environments. It leads to inaccurate estimation of surface diffuse color and specular fraction. Current methods either do not consider the diffuse color of the reflector[15, 38, 12], or wrongly model the reflection light and surface material [7, 35].

Our key idea is to use reflection ray-casting as supervision for the decomposition in order to resolve these two ambiguities. The crucial observation behind the idea is that the reflection virtual image should align with the reflected object or the observed surface is unlikely to be specular. It serves as an important clue to distinguish if the surface is specular and to estimate the correct reflection light. Fortunately, in indoor scenarios, corresponding objects appearing in reflections can be found in the reconstructed scene. Therefore, we cast reflection rays at surface points to obtain the depth and radiance of the reflected objects, and utilize these information to supervise the reconstructed virtual image and the surface materials. Specifically, we design two effective constraints: 1) **Reflection Consistency Loss**. The virtual image should have the same depth and radiance as the reflected object, satisfying the consistency requirement of planar reflection, thus can resolve the **light and material ambiguity**; 2) **Reflection Certainty Loss**. The surface specular fraction should highly related to the degree of consistency between the virtual image and reflected objects, which we regard as the certainty of the current virtual image to determine whether it is credible or pseudo. Thus, **appearance ambiguity** could be resolved without reflection masks by constraining the surface specular fraction with the certainty of virtual image.

To evaluate our method, we construct an indoor scene reconstruction dataset based on Replica Dataset[20]. We introduce more specular planar reflectors into the scene and render multi-view images, ground truth diffuse appearance, and a specular fraction map for reconstruction and material estimation tasks. We also validate the effectiveness of our method on a real-captured dataset.

In summary, we present the following contributions:

- We present the RS-SpecSDF framework to reconstruct surfaces for specular scenes without indicating mirror masks.
- We propose reflection ray-casting to supervise the decomposition between reflection and real scene to resolve the ambiguities in planar reflection.
- We build a new indoor scene dataset for 3d reconstruction with complicated specular planar reflectors. Experiments on our proposed dataset can demonstrate the effectiveness of our method.

2. Related Works

2.1. Neural Surface Reconstruction

NeRF [16] has recently emerged as a promising solution to 3d reconstruction task, leveraging an implicit scene representation and volume rendering to synthesize photorealistic images. However, volume density cannot represent high-fidelity surfaces due to the lack of surface constraints. Im-

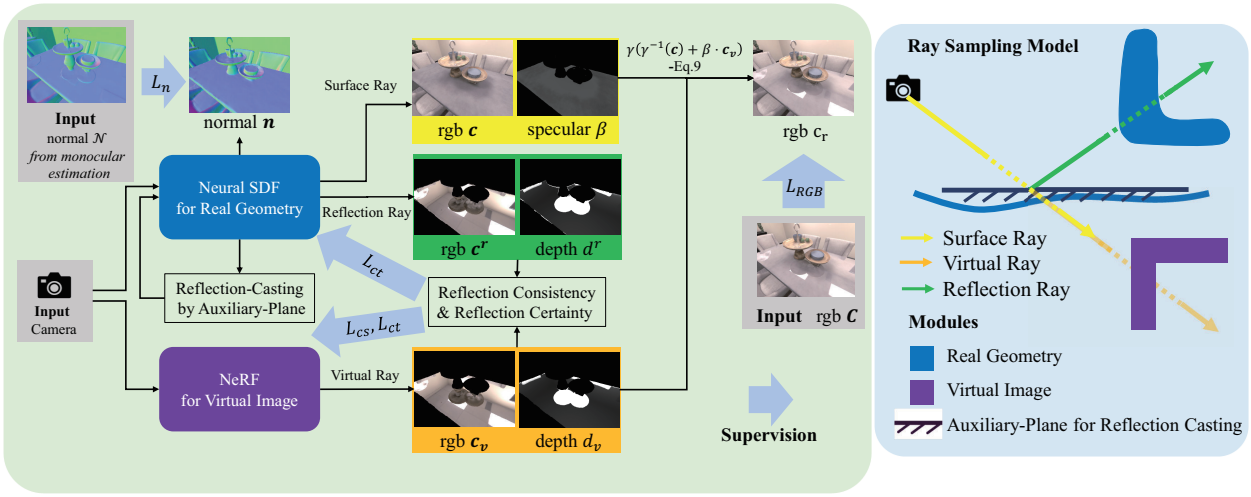

Figure 2. **Overview of RS-SpecSDF.** RS-SpecSDF reconstructs the indoor scenes with specular reflectors by decomposing the real geometry $\mathcal{G}$ and the reflection virtual image $\mathcal{F}$ into a neural SDF and a NeRF. During rendering process, we cast one ray to sample the surface SDF, one ray for the virtual image field, then we compose the final color by Eq. 9 and supervise it with input image. Additionally, we cast a reflection ray on the corresponding Auxiliary-Plane that approximates the accurate position of reflector. And we use the reflection ray to supervise the reconstructed virtual image and surface specular fraction β by our proposed Reflection Consistency Loss L_{cs} and Reflection Certainty Loss L_{ct}. Moverover, we adopt a normal prior correction scheme L_n to supervise the geometry using the predicted normal from monocular estimation.

proved reconstruction of surface geometry can be achieved by employing signed distance field(SDF), NeuS [27] and its concurrent works[33, 18] enable implicit neural SDF optimized by volume rendering. In order to address large texture-less or less observed areas in typical indoor scenes, monocular depth and normal priors[26, 36] are integrated. Specifically NeuRIS[26] has proposed a checking method based on patch matching to adaptively impose normal priors into optimization. Subsequent works follow the success of iNGP[17], representing the SDF with multi-resolution hash grid for faster training[28] and better performance[11]. Recently 3D Gaussian[10] have emerged as a popular choice for 3D representation due to its fast training speed. There are also methods[8, 37, 6, 1] attempting to reconstruct surface with 3D Gaussian splatting and achieving promising results on object surface reconstruction. However, they still cannot outperfrom the state-of-the-art of implicit representations[11] in large scenes[8, 37]. Therefore we continue to use implicit representation for SDF in our work.

2.2. Reflective Object Reconstruction

Reconstructing and rendering highly specular content remains a challenging task. Ref-NeRF and its follow-up works [22, 24, 4] demonstrated that reparameterizing outgoing radiance as a function of the reflected view direction is effective in cases where geometry is estimated accurately. An alternative approach to synthesize specular appearance is inverse rendering[40, 9, 3, 30]: estimating representations of scene materials and lighting which also parameterized as the function of reflected view direction. However this parameterization of reflected view direction is most effective for objects that are mainly illuminated by distant light sources. Progress has been made in solving the near field reflections[13, 32, 39, 25], however they all need geometry learning stage similar to RefNeRF[22] to acquire the near-perfect geometry of objects. SpecNeRF[14] modifies Ref-NeRF's encoding to vary spatially according to a set of optimized 3D Gaussians, which helps the color network represent the near field reflection of reflective objects. However, due to its limited capacity of Gaussians it can only handle the rough surface with weak reflection image[14]. Most recently, NeRF-Casting[23] proposing to trace reflected rays and render feature vectors, also mainly focus on objects which have strong geometry clues provided by the photometric difference between foreground and background. Therefore methods targeting non-planar object still cannot handle the mirror-like planar reflectors that bring great geometry ambiguity. They might model the virtual image as real geometries, thus cannot ensure the geometry correctly initialized for the reflection decomposition.

2.3. Planar Reflection Decomposition

Decomposition of planar specular reflection faces great ambiguity. Most methods[7, 38, 15, 12] requires mirror masks as supervision. NeRFReN[7] use two separated NeRFs for the transmitted and the reflected components, successfully preserving the geometry of transmitted part, but it cannot correctly estimate the reflection fraction due

to the light and material ambiguity. Mirror-NeRF and its concurrent works[38, 15, 12] share the same idea of directly reflecting rays on mirrors to represent the reflection. However, they need mirror masks for every input images to initialize the mirror geometries, and can only handle mirrors without diffuse color. Methods that do not utilize mirror masks either impose specific requirements on capture process to leverage physical cues[29, 31], or are unable to resolve the ambiguities of decomposition[35, 41]. Whelan at al[29] proposed the use of a scanner equipped with an AprilTag[19] to detect mirrors during scene capture. FlashSplat[31] necessitates shooting the scene twice, once with camera flash and once without, employing flash cues to recover transmitted appearance. MS-NeRF[35] models the scene by a group of feature fields in parallel sub-spaces, which allows it to manage multiple mirrors, however, it fails to achieve a clean decomposition of reflections from scene geometry. RefGaussian[41] solely relies on smoothness regularization for reflection decomposition, also resulting in incomplete seperation.

3. RS-SpecSDF

3.1. Overview

As shown in Figure 2, we take the RGB images $C = \{C_1, ..., C_M\}$ and the predicted normal maps $\mathcal{N} = \{\mathcal{N}_1, ..., \mathcal{N}_M\}$ generated by monocular estimation[2] as input, reconstruct the real geometry field $\mathcal{G}$ with its material $\mathcal{M}$ and the virtual image field $\mathcal{F}$ for reflections. We decompose the scene into real geometry and virtual image field (See Sec. 3.3), cast the reflected rays to supervise the decomposition by Reflection Consistency Loss and Reflection Certainty Loss (See Sec. 3.4). To ensure the reflected rays accurate, instead of directly reflect rays on the SDF surface, we use Auxiliary-Planes to approximate the planar specular surface and cast reflected rays from the hitting points (See Sec. 3.5). In order to reconstruct better surfaces at textureless areas, we adopt normal prior as supervision with a correction scheme (See Sec. 3.6). Finally, we introduce our training schedule at Sec. 3.7.

3.2. Preliminaries

Before detailing our method, we will first introduce the necessary preliminaries that underpin our approach.

Neural Radiance Field. Neural Radiance Field (NeRF)[16] represents a scene as a continuous volumetric field, where the density $\sigma \in \mathbb{R}$ and radiance $\mathbf{c} \in \mathbb{R}^3$ at any 3D position $\mathbf{x} \in \mathbb{R}^3$ under viewing direction $\mathbf{d} \in \mathbb{R}^2$ are modeled by a multi-layer perceptron(MLP):

$$f_\theta : (\mathbf{x}, \mathbf{d}) \rightarrow (\sigma, \mathbf{c}) \tag{1}$$

where θ is the learnable parameters. To render a pixel q_k from the camera center $\mathbf{o}$ at viewing direction $\mathbf{d}$, NeRF samples n points with depth $\{t_1, ..., t_n\}$ along the camera ray $R(q_k) = \{\mathbf{x_i} = \mathbf{o} + t_i\mathbf{d}\}_{i=1}^n$, then the rendered color of the ray is computed by:

$$\mathbf{c}(q_k) = \sum_{i=1}^{n} T_i \alpha_i \mathbf{c}_i \tag{2}$$

where $\alpha_i = 1 - \exp(-\sigma_i(t_{i+1} - t_i))$ is the opacity of the i-th point, and $T_i = \prod_{j<i}(1 - \alpha_j)$ is the transmittance after the i-th point. And the total opacity is computed by:

$$\alpha(q_k) = \sum_{i=1}^{n} T_i \alpha_i \tag{3}$$

During training, rendered RGB $\mathbf{c}_r = \mathbf{c}$ will be supervised by input images $\mathbf{C}$,

$$\mathcal{L}_{RGB} = \frac{1}{m} \sum_k ||\mathbf{c}_r(q_k) - \mathbf{C}(q_k)||_1 \tag{4}$$

where m is the sampled ray number for one training iteration.

Neural Surface Reconstruction. We follow NeuS[27] to represent the surface as the zero-level set of a Neural SDF $\{g_\theta(\mathbf{x}) = 0 \,|\mathbf{x} \in \mathbb{R}^3\}$. Then NeRF-like volumetric rendering[16] is applied to render images from neural SDF g_θ. And α_i in Eq. 2 is derived from the SDF value as proposed in NeuS[27]:

$$\alpha_i = \max\left(\frac{\Phi_s(g_\theta(x_i)) - \Phi_s(g_\theta(x_{i+1}))}{\Phi_s(g_\theta(x_i))}, 0\right) \tag{5}$$

where $\Phi_s(x) = \frac{1}{1+e^{-sx}}$ is the Sigmoid function, and s is a trainable parameter to control the density range of a surface, that is, $1/s$ approaches to zero as the training converges.

Total losses to train the neural SDF can be derived as:

$$\mathcal{L}_{sdf} = \mathcal{L}_{RGB} + \lambda_{eik}\mathcal{L}_{eik} + \lambda_{curv}\mathcal{L}_{curv} \tag{6}$$

where $\mathcal{L}_{eik}$ is the eikonal loss[5] enforcing the neural representation to be a valid SDF.

$$\mathcal{L}_{eik} = \frac{1}{nm} \sum_k \sum_{i=1}^{n} (||\nabla^2 g_\theta(x_{k,i})||_2 - 1)^2, \tag{7}$$

where n is the number of sample points of a ray. And $\mathcal{L}_{curv}$ is the curvature loss for smoothness of geometry defined in [11], computed by:

$$\mathcal{L}_{curv} = \frac{1}{nm} \sum_k \sum_{i=1}^{n} |\nabla^2 g_\theta(x_{k,i})|, \tag{8}$$

3.3. Reflection Decomposition

Neural Surface for real geometry To accurately reconstruct the surface of indoor scenes, we use multi-resolution hash grid neural SDF as the representation for real geometry field $\mathcal{G}$ and follow the training strategy proposed in Neuralangelo[11].

In detail, we represent the SDF as $g_\theta : \mathbf{x} \rightarrow g_\theta(\mathbf{x})$. 3D position $\mathbf{x}$ alongside viewing direction $\mathbf{d}$ encode surface radiance as $g_c : (\mathbf{x}, \mathbf{d}) \rightarrow c$, therefore the surface radiance could represent other view-dependent effect which isn't produced by planar reflection. Note that the input RGB images is in low dynamic range(LDR), so surface radiance is stored in standard RGB color space(sRGB): $\mathbf{c} \in [0, 1]^3$. To decompose the reflection from real geometry, a specular fraction value β is learned for each 3D position $\mathbf{x} \in \mathbb{R}^3$ to measure the specular property of objects in real geometry field: $g_\beta : \mathbf{x} \rightarrow \beta$.

NeRF for Virtual Image In order to decompose the reflection from the real geometry, we use a separated NeRF with multi-resolution hash grid encoding strategy [17] to model the virtual field $\mathcal{F}$. Note that composing the virtual field $\mathcal{F}$ with real geometry field $\mathcal{G}$ should be performed with radiance in high dynamic range(HDR) before gamma correction γ, in which the radiance of virtual field should be ranged in $[0, +\infty]^3$ instead of $[0, 1]^3$. Thus we compute radiance of virtual field $\mathcal{F}$ by $\mathbf{c}_v = \mathbf{c}_s \cdot (1 + I)$, where $\mathbf{c}_s \in [0, 1]^3$ is color in sRGB color space and $I \in [0, +\infty]$ is the intensity value to ensure the radiance range in $[0, +\infty]^3$. As the virtual image rarely having view-dependent effect, we do not introduce view information but only position to virtual NeRF: $f_v : \mathbf{x} \rightarrow (\sigma, \mathbf{c}_s, I)$

Composition During the pixel rendering process, we sample two sets of points for the ray $R(q_k)$: one starts from the camera center and queries the neural SDF, and the other starts from the SDF surface and queries the virtual NeRF. For the first set, we perform volume rendering to approximate surface color $\mathbf{c}(q_k)$, specular fraction $\beta(q_k)$, normal $\mathbf{n}(q_k)$ and depth $d(q_k)$ following Eq. 2, where notation (q_k) represents the volume-rendered quantity corresponding to pixel q_k. For the second set, the radiance $\mathbf{c}_v(q_k)$, opacity $\alpha_v(q_k)$ and depth $d_v(q_k)$ from virtual NeRF are computed in the similar way. Particularly, we introduce a random background color to radiance $\mathbf{c}_v(q_k)$ with its opacity $\alpha_v(q_k)$, forcing the opacity to approach 1.

Then we compose radiance from real geometry field $\mathcal{G}$ and virtual field $\mathcal{F}$ by

$$\mathbf{c}_r = \gamma(\gamma^{-1}(\mathbf{c}) + \beta \cdot \mathbf{c}_v) \tag{9}$$

where $\mathbf{c}_r$ is the final render RGB, $\gamma(\cdot)$ is the gamma correction function and the inverse gamma correction $\gamma^{-1}(\cdot)$ is applied on SDF radiance $\mathbf{c}$ due to incompetence between the sRGB range of g_c and HDR space to perform composition. Here the notation (q_k) is omitted.

Our design that decomposes reflection into Neural SDF and NeRF is also based on following considerations: Firstly, we adopt neural SDF to utilize its inherited advantage that the SDF forms the volumetric weights as a single opaque surface when the $1/s$ value approaches zero. Therefore it is unnecessary to add geometric priors to regularize the real scene part as NeRFReN[7]. Secondly, we still employ NeRF as the virtual image field, because the neural SDF needs careful initialization at approximate object position[27]. However, we cannot infer the position of virtual image in advance.

3.4. Reflection Constraints

Based on the reflection physics, reflection image provides a natural cue: the virtual image should exhibit the same appearance with the reflected object, and also symmetrically positioned. To address the appearance ambiguity and the light-material ambiguity during the decomposition of the reflection virtual image and the real geometry, we introduce Reflection Consistency Loss and Reflection Certainty Loss leveraging the reflection cues.

Reflection Consistency Loss. In order to regularize the virtual image to be consistent with reflected object, we propose the reflection consistency loss $\mathcal{L}_{cs}$. It enforces that the depth and radiance of the virtual ray and the reflected ray are equivalent:

$$\mathcal{L}_{cs} = \frac{1}{m} \sum_k \left(|d(q_k) - d^r(q_k)| + ||\mathbf{c}(q_k) - \mathbf{c}^r(q_k)||_1 \right) \cdot \mathbb{1}(\beta(q_k) > \epsilon_s) \cdot \mathbb{1}(\alpha_v(q_k) > \epsilon_v) \tag{10}$$

where superscript 'r' for $\mathbf{c}^r, d^r$ represents the quantity of the reflected object the ray cast on (See Sec3.5) and $\mathbb{1}$ is the indicator function. Since the virtual image is valid only beneath the specular surfaces, the Reflection Consistency Loss should only supervise the rays that passes through specular surface. We set the threshold $\epsilon_s = 0.005$ to determine whether the surface is specular or not. Additionally, as regularizing the radiance and depth of virtual NeRF will encourage the opacity α_v growing, $\mathcal{L}_{cs}$ will not be applied where the virtual image opacity is lower than a threshold $\epsilon_v = 0.5$.

Given the correct radiance of reflection, the specular fraction and diffuse appearance on specular surface can be inferred correctly, thereby addressing the light-material ambiguity. However, when addressing appearance ambiguity at non-specular surfaces, the specular fraction can easily be trapped into the local-minima, attempting to account for the diffuse appearance of non specular surface with a 'pseudo' virtual image. Even when applying Reflection Consistency Loss to regularize the pseudo virtual image, it might still fail to achieve the consistency. This is because the pseudo virtual image under non-specular surface is not well-defined and might have conflict with either the credible virtual im-

age or other pseudo virtual images.

Reflection Certainty Loss. Since $\mathcal{L}_{cs}$ is only able to supervise the specular surface, we propose Reflection Certainty Loss to resolve the appearance ambiguity at non-specular surface. Our approach involves penalizing both the specular fraction β and virtual image opacity α_v where the reconstructed virtual image lacks sufficient consistency with the reflected object. We conceptualize the level of consistency as the 'certainty' of a virtual ray to determine whether the surface is specular.

We define the depth certainty $\mathcal{C}_d$ to represent the NeRF density on a virtual ray that falls within the range of the consistent depth d^r, computed by the accumulated weight of sampled points:

$$\mathcal{C}_d(q_k) = \sum_{i=1}^{n} T_i \alpha_i \cdot \mathbb{1}(|t_i - d^r(q_k)| < \epsilon_d) \tag{11}$$

where $T_i\alpha_i$ is the weights of virtual NeRF, t_i is the depth of sampled points, and $\mathbb{1}$ is an indicator function, with threshold $\epsilon_d = 0.3$.

We observe that the depth consistency of credible virtual image can be naturally satisfied without applying Reflection Consistency Loss, as the reflection image inherently provides geometric information. In contrast, the pseudo virtual image beneath non-specular surfaces fails to generate sufficient density within the range of consistent depth. Leveraging this property, we regularize the non-specular surface by depth certainty before applying $\mathcal{L}_{cs}$.

In order that the pseudo virtual image approaches near-zero opacity, we regularize total opacity α_v of virtual NeRF by depth reflection certainty loss computed as:

$$\mathcal{L}_{dct} = \frac{1}{m} \sum_{k,\beta(q_k)>\epsilon_s} \alpha_v(q_k) \cdot \mathbb{1}(\mathcal{C}_d(q_k) < \epsilon_{dct}) \tag{12}$$

where the threshold $\epsilon_{dct} = 0.5$. Note that the radiance of virtual image $\mathbf{c}_v$ in Eq. 9 has injected random background color, therefore the low opacity will also encourage specular fraction β to approach zero.

However, depth certainty serves as a rough indicator, as the pseudo virtual image may luckily fall within the consistent depth range of reflected objects. To address this limitation, we introduce radiance certainty, defined by the radiance error of the virtual image and reflected object, that is:

$$\mathcal{C}_r(q_k) = ||\mathbf{c}(q_k) - \mathbf{c}^r(q_k)||_1 \tag{13}$$

Since the virtual image requires applying $\mathbf{L}_{cs}$ to ensure the radiance consistency with the reflected object, the radiance certainty loss should be applied concurrently with $\mathbf{L}_{cs}$, which forces the opacity of virtual NeRF to grow. To prevent contradictory regularization on opacity α_v, the radiance certainty loss directly regularizes the surface specular fraction β, computed as:

$$\mathcal{L}_{rct} = \frac{1}{m} \sum_{k,\mathcal{C}_r(q_k)>\epsilon_{rct}} \beta(q_k) \tag{14}$$

where the threshold $\epsilon_{rct} = 0.5$. And the Refection Certainty Loss is composed by these two constrains:

$$\mathcal{L}_{ct} = \lambda_{dct} \cdot \mathcal{L}_{dct} + \mathcal{L}_{rct} \tag{15}$$

where the weighting factor λ_{dct} is set to 0.05.

3.5. Auxiliary-Plane for Reflection Ray-Casting

We observe that the under-constructed SDF surface cannot meet the accuracy requirements for reflection ray-casting. To address this, we introduce K learnable planes $\{P[i]\}_{i=1}^{K}$ to approximate the specular surface and cast reflection rays on these **auxiliary-planes**. Each auxiliary-plane is defined by $P[i] = \{\mathbf{x} \in \mathbb{R}^3 | \mathbf{n}[i] \cdot \mathbf{x} + w[i] = 0\}$, where $\mathbf{n}$ represents the normal vector of the plane. We use the notation i_k to represent the index of corresponding auxiliary-plane of the surface where ray $R(q_k)$ hits. When the camera ray $R(q_k) = \{\mathbf{x} \in \mathbb{R}^3 | \mathbf{x} = \mathbf{o} + t\mathbf{d}\}$ cast reflection on $P[i_k]$, the depth of hitting point can be calculated by:

$$d_p(q_k) = -\frac{\mathbf{n}[i_k] \cdot \mathbf{o} + w[i_k]}{\mathbf{n}[i_k] \cdot \mathbf{d}} \tag{16}$$

We use the hitting point on plane $P[i_k]$ as the reflection point, that is, the reflected ray originates from $\mathbf{o}_r = \mathbf{o} + d_p\mathbf{d}$ and follows the direction $\mathbf{d}_r = \mathbf{d} - 2(\mathbf{d} \cdot \mathbf{n}[i_k])\mathbf{n}[i_k]$. We then sample n_r points along the reflection rays and perform volume rendering to obtain the radiance c^r, depth d^r of reflected objects.

For each sampled rays $R(q_k)$, in order to determine which auxiliary-plane the ray should cast reflection on, we enable the virtual image field to additionally store the normal vector of the corresponding auxiliary-plane $P[i_k]$, denoted as corresponding normal field $f_n : \mathbf{x} \rightarrow \mathbf{n_c}$. This design is based on the fact that all points within the virtual image of a plane reflector should align with the same auxiliary-plane. Consequently the corresponding normal exhibits spatial smoothness, making f_n easy to learn.

To train this corresponding normal field, we regularize the corresponding normal with the normal of SDF surface $\mathbf{n}$.

$$\mathcal{L}_{cn} = \frac{1}{m} \sum_{k,\beta(q_k)>\epsilon_s} ||\mathbf{n_c}(q_k) - \mathbf{n}(q_k)|| \tag{17}$$

Given the corresponding normal $\mathbf{n_c}(q_k)$, we select the auxiliary-plane with the nearest normal to cast reflection, that is $i_k = argmin_i(|\mathbf{n_c} - \mathbf{n}[i]|)$. We also check the distance between the depth $d(q_k)$ of hitting point on SDF and $d_p(q_k)$ of auxiliary-plane $P[i_k]$. If the distance exceed a threshold ϵ_{pd}, which indicates that current hitting point

doesn't correspond to a planar reflector, we directly cast reflection on the SDF surface.

The auxiliary-planes are initialized by pointcloud plane segmentation of the mesh extracted from Neural SDF g_θ, further details can be found in our Supplementary materials. We directly store the plane normal vector $\mathbf{n}$ and position w as learnable parameters, allowing for optimization to ensure accurate reflection. Notably, the auxiliary-planes can be regularized by Reflection Consistency Loss L_{cs}, because the gradient of d^r and $\mathbf{c}^r$ can be backpropagated to the origin $\mathbf{o}_r$ and direction $\mathbf{d}_r$ of reflected rays. Additionally, in order to stabilize the position of auxiliary-planes, we regularize the distance of the SDF surface and the corresponding plane $P[i_k]$, computed by:

$$\mathcal{L}_d = \frac{1}{m} \sum_{k,\beta(q_k)>\epsilon_s} |d(q_k) - d_p(q_k)| \tag{18}$$

3.6. Normal Prior with Correction

In order to guarantee the quality of reconstructed surface for texture-less areas in typical indoor scenes, we follow [26][36] to integrate estimated normal as a prior. We use a pretrained Omnidata model[2] to predict normal maps $\mathcal{N}$ for input RGB images.

Based on the observation that the predicted normal maps $\mathcal{N}$ are easily erroneous inside the plane reflector and always correspond to a blending normal between the virtual image normal and the reflective surface normal, we introduce a checking method for normal reliability and adaptively imposing the prior supervision in the optimization process.

To avoid introducing wrong normal prior on specular surface, we utilize the normal vector of the corresponding auxiliary-plane $P(q_k)$ of the surface as the corrected normal target. Note that the estimated specular fraction cannot accurately delineate the true range of specular surface before the network convergence, due to the detailed geometry above the specular plane remaining under-reconstructed. Therefore, regularizing normal only based on the estimated specular fraction is unreliable, potentially overriding the geometry clue of images at non-specular surface, which might cause the missing of geometries near the reflector as shown in Fig 3. To address this, we propose a normal correction mask M to further assess whether the predicted normal $\mathcal{N}$ falls within the true range of specular surfaces. The scheme originates from our observation that the predicted normal maps on reflective surface always correspond to a blending between the normal of virtual image and specular surface. We perform correction to normal prior $\mathcal{N}$ only if the predicted normal $\mathcal{N}$ predominantly comprises the normal of virtual image $\mathbf{n}_v$ and the normal of the surface $\mathbf{n}$. That is, the projection error of $\mathcal{N}$ should be small enough:

$$E_n = \min_{a>0,b>0} ||\mathcal{N} - (a \cdot \mathbf{n}_v + b \cdot \mathbf{n})||_2 \tag{19}$$

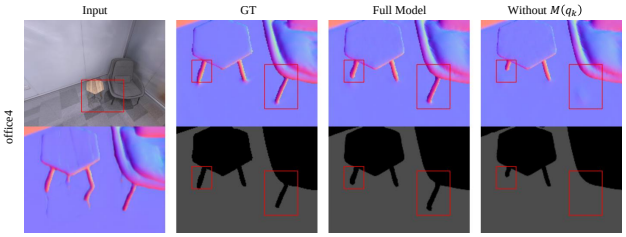

Figure 3. An ablation study showing the effect of our design choices on the checking method for normal prior. Directly perform the correction indicated by the specular fraction β will wrongly correct the normal for tiny non-specular objects above the specular surface (Without $M(q_k)$). With our design of correction mask, we successfully utilize the geometry guidance from normal prior and faithfully reconstruct these geometry details (Full Model).

Therefore we use this as an indicator to determine the normal prior correction mask $M(q_k)$ computed as:

$$M(q_k) = \mathbb{1}[E_n(q_k) < \epsilon_n] \tag{20}$$

where $\mathbb{1}$ is the indicator function. In our experiments, we set $\epsilon_n = 0.1$. Note that we cannot directly obtain the normal of virtual image from our virtual NeRF, so we use the normal of reflected object $\mathbf{n}^r$ to approximate the normal of virtual image $\mathbf{n}_v$ by:

$$\mathbf{n}_v \approx 2(\mathbf{n} \cdot \mathbf{n}^r)\mathbf{n} - \mathbf{n}^r \tag{21}$$

which means to reflect the direction of $\mathbf{n}^r$ by the surface normal $\mathbf{n}$, as the virtual image is symmetric with respect to the reflected object.

Given the normal prior $\mathcal{N}$ and the normal vector $\mathbf{n}[i]$ from assited-plane, along with a correction mask M, the corrected normal prior loss is derived as follows:

$$\begin{aligned}\mathcal{L}_n = \frac{1}{m}[&\sum_{k,\beta(q_k)<\epsilon_s} ||\mathbf{n}(q_k) - \mathcal{N}(q_k)|| + \\ &\lambda_r \cdot \sum_{k,\beta(q_k)>\epsilon_s} ||\mathbf{n}(q_k) - \mathbf{n}[i](q_k)|| \cdot M(q_k)]\end{aligned} \tag{22}$$

where the former part supervises the non-reflective surface by the normal priors while the latter part performs as normal regularization on specular surface with the factor $\lambda_r = 0.5$.

3.7. Optimization

The overall loss used to optimize RS-SpecSDF is the weighted combination of losses from Neuralangelo[11] $\mathcal{L}_{sdf}$ (See Eq. 6), the supervision loss $\mathcal{L}_{ap}$ for auxiliary-planes (Sec. 3.5), the two proposed reflection constrains$\mathcal{L}_{cs}, \mathcal{L}_{ct}$ (Sec. 3.4), and normal priors loss $\mathcal{L}_n$ (Sec. 3.6):

$$\begin{aligned}\mathcal{L}_{total} =&\mathcal{L}_{sdf} + \lambda_{ap}\mathcal{L}_{ap} \\ &+ \lambda_{cs}\mathcal{L}_{cs} + \lambda_{ct}\mathcal{L}_{ct} + \lambda_n\mathcal{L}_n\end{aligned} \tag{23}$$

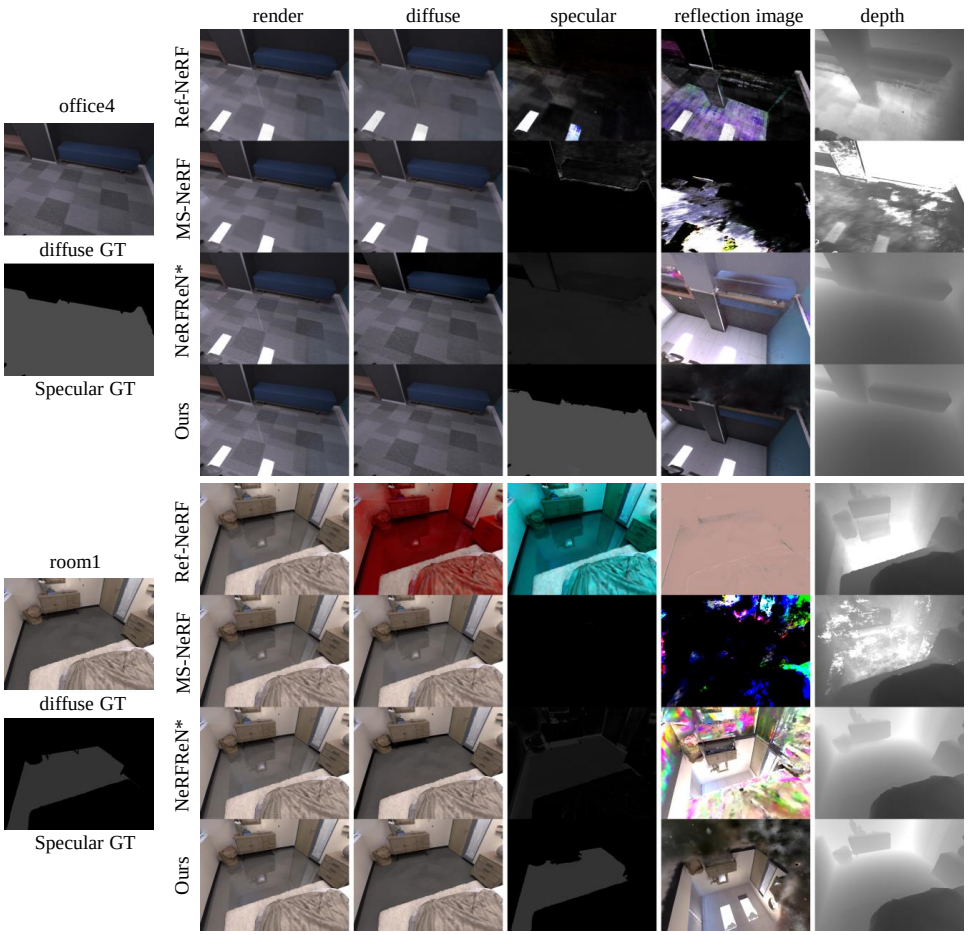

Figure 4. Decomposition components of our approach compared to baselines on our synthetic dataset. All baselines cannot estimate the correct specular fraction map. Our approach produces the decomposition most consistent with the ground truth, due to the effectiveness of reflection constrains.

where the supervisions for auxiliary-planes are integrated as $\mathcal{L}_{ap}$:

$$\mathcal{L}_{ap} = \mathcal{L}_d + \mathcal{L}_{cn} \tag{24}$$

Further more, we propose a universal training strategy to eliminate the necessity of tuning hyper-parameters for each scene. In order to ensure the surface geometry can be initialized and to prevent the reflection virtual NeRF explaining the whole scene, we start the decomposition Eq. 9 after 2k iterations, before that we directly set: $\mathbf{c}_r = \mathbf{c}$.

We extract auxiliary-planes and start casting reflections on them at 50k iterations. The reflection consistency loss $\mathcal{L}_{cs}$ is activated at 100k iterations. And depth certainty loss $\mathcal{L}_{dct}$ and radiance certainty loss $\mathcal{L}_{rct}$, are introduced at 60k and 150k iterations respectively.

4. Experiments

4.1. Experimental settings

Datasets We conduct our experiments on both synthetic dataset and real-captured dataset. According to our survey, the commonly used dataset for plane reflectors contains either forward-facing scenes[7] or mirror reflectors without diffuse component[38, 35]. There lacks an indoor dataset featuring specular planar reflectors with diffuse color. Therefore we conduct experiments on our newly-proposed dataset, based on a synthetic dataset of indoor scenes named Replica Dataset[20]. We add some specular planar reflectors with different reflection fraction ranged from $[0.1, 0.3]$ into the scene, such as tables, walls and floors. The composition of materials and specular reflection was simulated using the equation: $\mathbf{c} = \gamma(\mathbf{c_d} + \beta \cdot \mathbf{c_s})$, where β is the specular fraction, $\mathbf{c_d}$ is the diffuse radiance,

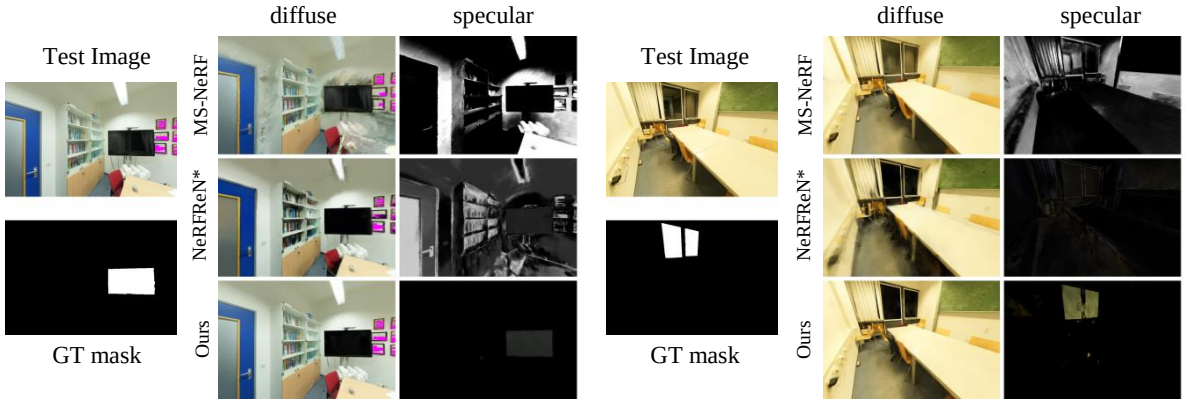

Figure 5. Decomposition comparisons between our approach and baselines on real-captured dataset Scannet++.

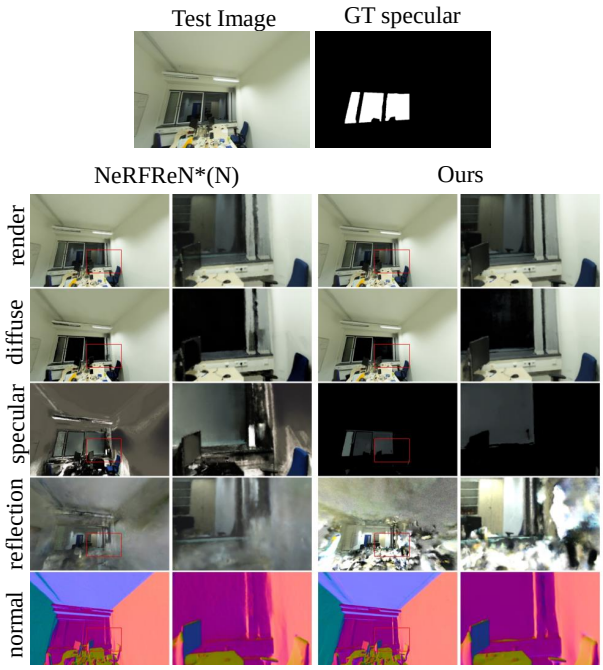

Figure 6. Decomposition components of our approach compared to NeRFReN*(N) on the Scannet++ dataset.

and $\mathbf{c_s}$ is the specular radiance. We selected 5 scenes from our dataset to evaluate the performance of our method. For each scene, we render the composed RGB images around 200 views for training and the ground truth diffuse RGB, normal, depth, and specular fraction around 20 views for evaluation.

Additionally, we conduct experiments on real-captured dataset Scannet++[34]. We select 3 scenes with planar specular reflectors to validate the effectiveness of our method. We use every 16 input images for testing and others for training. However, Scannet++ dataset doesn't provide annotations for planar reflectors, so we manually annotated the specular reflector masks for each test images.

Baselines We compare our method with the available state-of-the-art neural rendering methods dealing with reflections, namely RefNeRF[22], NeRFReN[7], and MS-NeRF[35], serving as our baselines. For fair comparison, we implement a new version called NeRFReN* with SDF representation[11] as transmitted part and hash-grid NeRF[17] as reflected part, sharing the exact same structure as our method. Additionally, we also apply normal priors on NeRFReN*. NeRFReN*(N) directly using normal prior without correction. Besides, we use 2 sub-spaces instead of 8 sub-spaces in MS-NeRF, as each of our specular scenes only has one or two specular surface. And we aim to evaluate its ability to decompose the scene into reflection parts and the real scene. Note that all baselines and our method are trained without mirror masks as supervision.

Parameter Settings We empirically set the weight of each loss as following. The weight factor of normal priors loss λ_n is set to 0.2, and $\lambda_{ap} = 0.1$ for the auxiliary-planes, $\lambda_{cs} = 0.1$ for the Reflection Consistency Loss, $\lambda_{ct} = 1.0$ for the Reflection Certainty Loss. And we use $K = 20$ as the capacity of the auxiliary-planes.

Metrics For novel view synthesis, the metrics include PSNR of rendered image, decomposed diffuse image. To evaluate the accuracy of recovered specular fraction, we render the specular fraction into range $[0, 1]$, and calculate PSNR with the ground truth. Furthermore, we calculate the F-Score of the specular range indicated by the rendered specular fraction map. We evaluate 3D surface reconstruction following the metrics defined in NeuralRecon[21]. Among those metrics, F-score is usually considered as the most suitable metric to evaluate geometry quality. Refer to the supplementary for more details.

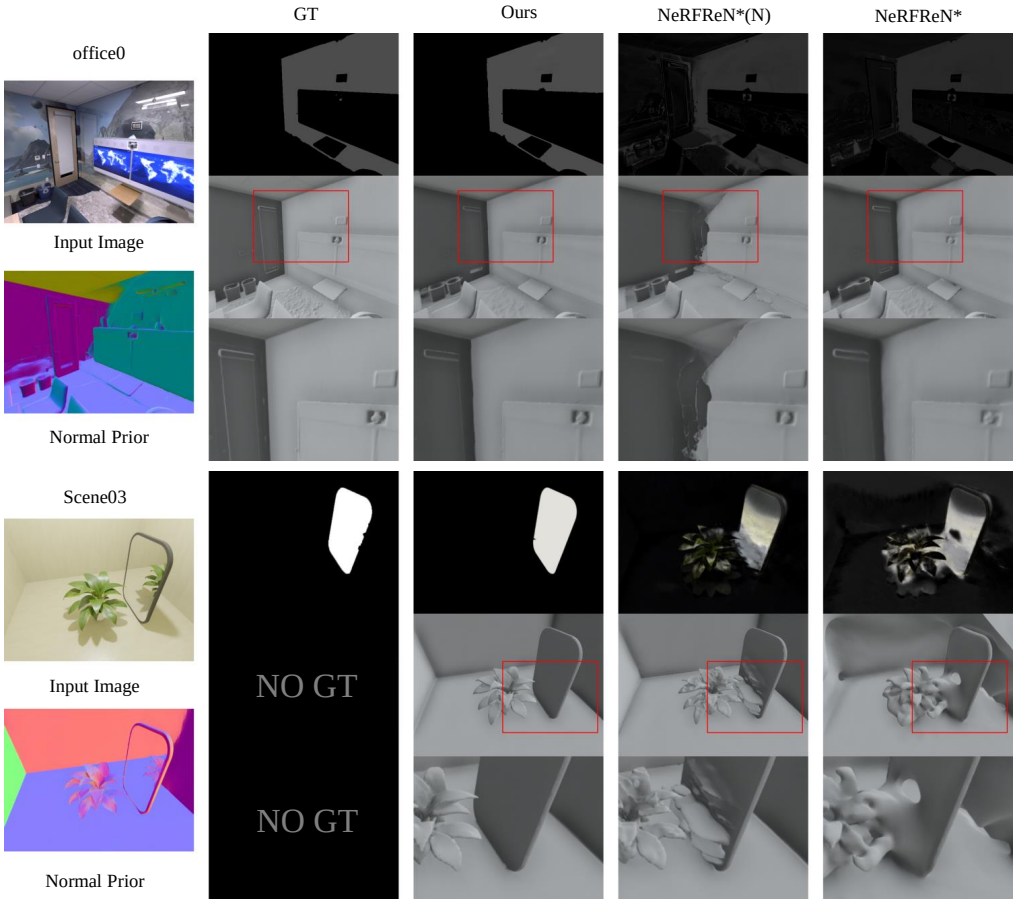

Figure 7. Surface reconstruction results of our approach compared to baselines on our synthetic dataset and a toy scene from MS-NeRF dataset[35]. Our method produces more accurate and higher-fidelity surfaces in all cases. NeRFReN* with or without normal prior all fail at certain cases.

4.2. Comparisons

Novel view synthesis and reflection decomposition Table 1 and Figure 4 show the quantitative and qualitative results compared with the state-of-the-art methods in our synthetic dataset. We achieve the highest PSNR on render

Table 1. **Quantitative comparisons of novel view synthesis results** over 5 scenes of our synthetic dataset.

Method	PSNR(R)↑	PSNR(D)↑	PSNR(S)↑	F-Score(S)↑
RefNeRF	31.934	16.373	6.939	0.275
MSNeRF	32.877	26.346	20.638	0.732
NeRFReN*	36.481	30.570	22.139	0.593
NeRFReN*(N)	36.426	31.451	22.779	0.612
Ours	**37.339**	**37.741**	**36.878**	**0.992**

image, diffuse image and specular fraction map. Additionally, specular range F-Score of our method reaches nearly 1.0, indicating that our method can accurately recover the specular range without guidance of mirror masks. While RefNeRF and MS-NeRF produce seemingly fair novel view

Table 2. **Quantitative comparisons of novel view synthesis results** over 3 scenes of real-captured dataset.

Method	PSNR(R)↑	PSNR(D)↑	F-Score(S)↑
MSNeRF	23.565	22.773	0.508
NeRFReN*	27.498	19.659	0.153
NeRFReN*(N)	27.877	20.983	0.126
Ours	**27.957**	**27.860**	**0.967**

results, they actually fail to decompose the reflection from the real scenes as the low PSNR of diffuse image and specular fraction map shows. RefNeRF models the virtual image as real geometries because it cannot accumulate enough density at the correct surface position, thus it cannot use NeRF's view-dependency to model the reflection. Similarly, MS-NeRF also struggles to decompose the reflection because MS-NeRF relies on the view-inconsistency of reflection virtual image to perform correct decomposition. However, in our dataset, the virtual images are mostly

view-consistent, which conflict with the basic assumption of MS-NeRF. NeRFReN* and NeRFReN*(N) successfully decompose the reflection from the real scene. However, they not only estimate a biased specular fraction at specular surface, but also incorrectly use reflection virtual image to explain some diffuse surfaces. Consequently, they achieve low scores on both specular PSNR and specular F-score. In contrast, our solution can correctly decompose the reflection and faithfully recover the specular fraction of the surface and achieve better results in all cases.

Qualitative and quantitative results on the Scannet++ datasets are provided in Figure 5 and Table 2. Note that the Scannet++ dataset doesn't provide the ground truth for diffuse appearance, so we compare the rendered image outside the mirror mask as the diffuse component. Furthermore, we can only annotate the specular fraction as a binary mask without its value, therefore we do not evaluate on the specular PSNR metrics. We can see that other baselines fail to recover the correct range of specular reflectors. In contrast, our method shows great effectiveness in decomposing the reflection from the diffuse appearance. We also visualize and compare the decomposition components produced by our method and NeRFReN*(N) in Figure 6. Our method recover the correct range of specular surface, thus won't explain the diffuse appearance by pseudo virtual image, resulting in better novel view synthesis results on both render image and diffuse image.

3D Reconstruction. Table 3 and Figure 7 present the quantitative and qualitative results of the surface reconstruction compared with baselines on the synthetic scenes. Note that since RefNeRF and MS-NeRF are not designed for surface reconstruction and all fail to decompose the virtual image from real geometries, we haven't evaluated their results for 3D reconstruction. Our approach fully utilizes the normal prior based on our normal correction scheme and correct estimation of the range of specular surface. Consequently, the correction to normal priors doesn't influence the geometry at non-specular surfaces. NeRFReN* without normal priors struggles to reconstruct accurate surface at low-texture areas both on specular surface and non-specular surface. NeRFReN*(N) directly supervised with normals will generate wrong geometry due to the wrongly predicted normal at the specular surface as shown in the example of *office0* in Figure 7. Above results demonstrate that only our method can utilize the normal prior and perform appropriate correction, thus achieving faithful surface reconstruction with robustness.

4.3. Ablation Study

In this section we carefully ablate the key components of our method to justify our design choices.

Reflection Constrains. Our main component is the reflection constrains. Thus we provide ablation studies on the reflection consistency loss and reflection certainty loss to demonstrate their effectiveness as shown in Table 4 and Figure 8. Model (a) only uses $\mathcal{L}_{cs}$ and without $\mathcal{L}_{ct}$, so only virtual image at specular surface is constrained correctly. Wrong specular fraction remains uncontrained at non-specular surface. Model (b) uses $\mathcal{L}_{ct}$ without $\mathcal{L}_{cs}$. Note that the radiance certainty calculation requires the virtual image having consistent appearance with the reflected object. Therefore, it becomes unavailable during the absence of $\mathcal{L}_{cs}$. We only adopt depth certainty loss into our certainty loss. It achieves slightly higher F-Score on specular range than Model(a) but still cannot estimate the accurate specular fraction for specular surface. Moreover, the depth certainty loss $\mathcal{L}_{dct}$ is a rougher indicator compared with $\mathcal{L}_{rct}$. Thus it cannot produce the same accurate specular range as our full model (Model (c)).

Table 3. **Quantitative comparisons of reconstruction results** over 5 scenes of our synthetic dataset.

Method	Accu.↓	Comp.↓	Prec.↑	Recall.↑	F-Score↑
NeRFReN*	0.504	0.663	92.261	89.159	90.678
NeRFReN*(N)	0.421	0.449	95.931	94.623	95.272
Ours	**0.363**	**0.374**	**98.039**	**97.424**	**97.729**

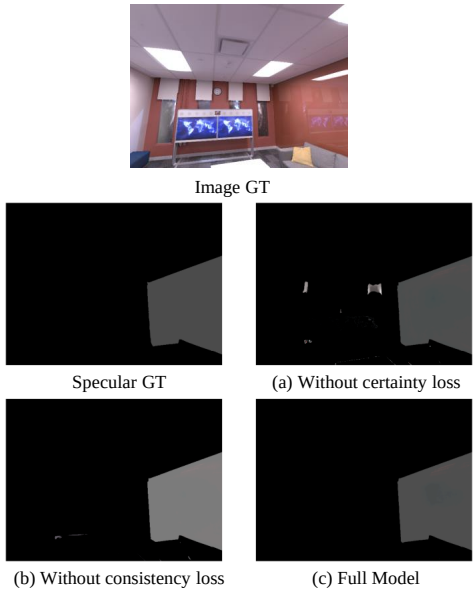

Figure 8. An ablation study showing the effect of the reflection constrains.

Table 4. **Novel view synthesis results for ablation studies of the reflection constrains** of our method over 5 scenes of our synthetic dataset.

Method	PSNR(R)↑	PSNR(D)↑	PSNR(S)↑	F-Score(S)↑
(a) without certainty	36.808	35.358	31.663	0.980
(b) without consistency	36.949	36.121	28.572	0.983
(c) Full Model	**37.339**	**37.741**	**36.878**	**0.992**

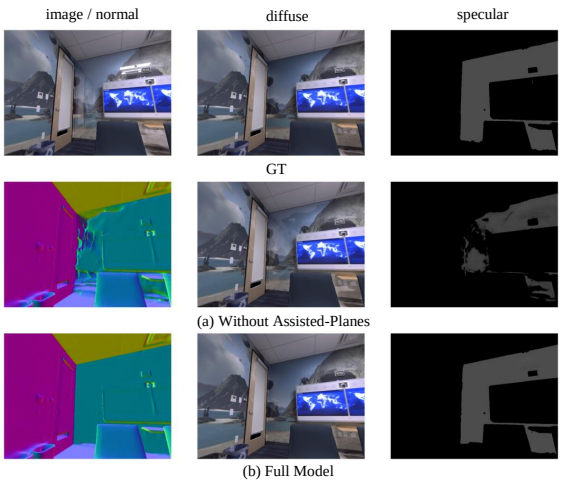

Figure 9. An ablation study showing the effect of casting reflection rays on Auxiliary-Planes.

Normal Prior Correction. Figure 3 shows the qualitative results removing the normal correction mask M computed by normal prior projection. Some thin structures above the specular surface might remain under-reconstructed, causing the ray that should hit the non-specular thin structure to reach the specular surface. Without the normal correction mask, normal correction will be applied to pull this ray's normal to the plane, hindering the growth of the detailed geometry. Our correction mask can recognize that the predicted normal does not belong to the specular surface. With better utilization of the geometry guidance provided by normal prior, our normal prior correction scheme effectively reconstructs detailed geometry near the specular surface.

Auxiliary-Planes. As shown in Figure 9, Model(a) directly casting reflection rays on SDF surface cannot achieve correct decomposition results. At early stage of training, SDF surface still remains incomplete compared with true reflector. Therefore, casting reflection rays on SDF surface, might produce biased reflection rays, causing the Reflection Constrains wrongly constrain the decomposition.

5. Conclusions

We have presented RS-SpecSDF, a framework to reconstruct surfaces for specular scenes without indicating mirror masks and decompose the unbiased reflection from the real scene. Our method utilizes reflection ray-casting as a supervision. Specifically, we have proposed Reflection Consistency Loss and Reflection Certainty Loss to regularize the decomposition of reflection virtual image and surface material. Therefore we can acquire the accurate range of specular surface and the unbiased specular fraction for material estimation task. Based on the accurate range of specular surface, we have proposed a correction method for predicted normal from monocular estimation which is possibly wrong at specular surface. Our method can fully leverage the normal prior as supervision and faithfully reconstruct the surfaces.

References

[1] D. Chen, H. Li, W. Ye, Y. Wang, W. Xie, S. Zhai, N. Wang, H. Liu, H. Bao, and G. Zhang. Pgsr: Planar-based gaussian splatting for efficient and high-fidelity surface reconstruction. *arXiv preprint arXiv:2406.06521*, 2024. 3

[2] A. Eftekhar, A. Sax, J. Malik, and A. Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10786–10796, 2021. 4, 7

[3] Y. Fan, I. Skorokhodov, O. Voynov, S. Ignatyev, E. Burnaev, P. Wonka, and Y. Wang. Factored-neus: Reconstructing surfaces, illumination, and materials of possibly glossy objects. *arXiv preprint arXiv:2305.17929*, 2023. 3

[4] W. Ge, T. Hu, H. Zhao, S. Liu, and Y.-C. Chen. Ref-neus: Ambiguity-reduced neural implicit surface learning for multi-view reconstruction with reflection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4251–4260, 2023. 1, 3

[5] A. Gropp, L. Yariv, N. Haim, M. Atzmon, and Y. Lipman. Implicit geometric regularization for learning shapes. *arXiv preprint arXiv:2002.10099*, 2020. 4

[6] A. Guédon and V. Lepetit. Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering. *arXiv preprint arXiv:2311.12775*, 2023. 3

[7] Y.-C. Guo, D. Kang, L. Bao, Y. He, and S.-H. Zhang. Nerfren: Neural radiance fields with reflections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18409–18418, 2022. 1, 2, 3, 5, 8, 9

[8] B. Huang, Z. Yu, A. Chen, A. Geiger, and S. Gao. 2d gaussian splatting for geometrically accurate radiance fields. *arXiv preprint arXiv:2403.17888*, 2024. 3

[9] Y. Jiang, J. Tu, Y. Liu, X. Gao, X. Long, W. Wang, and Y. Ma. Gaussianshader: 3d gaussian splatting with shading functions for reflective surfaces. *arXiv preprint arXiv:2311.17977*, 2023. 1, 3

[10] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):1–14, 2023. 3

[11] Z. Li, T. Müller, A. Evans, R. H. Taylor, M. Unberath, M.-Y. Liu, and C.-H. Lin. Neuralangelo: High-fidelity neural surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8456–8465, 2023. 1, 3, 4, 5, 7, 9

[12] J. Liu, X. Tang, F. Cheng, R. Yang, Z. Li, J. Liu, Y. Huang, J. Lin, S. Liu, X. Wu, et al. Mirrorgaussian: Reflecting 3d gaussians for reconstructing mirror reflections. *arXiv preprint arXiv:2405.11921*, 2024. 1, 2, 3, 4

[13] Y. Liu, P. Wang, C. Lin, X. Long, J. Wang, L. Liu, T. Komura, and W. Wang. Nero: Neural geometry and brdf reconstruction of reflective objects from multiview images. *ACM Transactions on Graphics (TOG)*, 42(4):1–22, 2023. 1, 3

[14] L. Ma, V. Agrawal, H. Turki, C. Kim, C. Gao, P. Sander, M. Zollhöfer, and C. Richardt. Specnerf: Gaussian directional encoding for specular reflections. *arXiv preprint arXiv:2312.13102*, 2023. 3

[15] J. Meng, H. Li, Y. Wu, Q. Gao, S. Yang, J. Zhang, and S. Ma. Mirror-3dgs: Incorporating mirror reflections into 3d gaussian splatting. *arXiv preprint arXiv:2404.01168*, 2024. 1, 2, 3, 4

[16] 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. 1, 2, 4

[17] 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, 5, 9

[18] 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

[19] E. Olson. Apriltag: A robust and flexible visual fiducial system. In *2011 IEEE international conference on robotics and automation*, pages 3400–3407. IEEE, 2011. 4

[20] J. Straub, T. Whelan, L. Ma, Y. Chen, E. Wijmans, S. Green, J. J. Engel, R. Mur-Artal, C. Ren, S. Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 2, 8

[21] J. Sun, Y. Xie, L. Chen, X. Zhou, and H. Bao. Neuralrecon: Real-time coherent 3d reconstruction from monocular video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15598–15607, 2021. 9

[22] 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. 1, 3, 9

[23] 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. *arXiv preprint arXiv:2405.14871*, 2024. 3

[24] F. Wang, M.-J. Rakotosaona, M. Niemeyer, R. Szeliski, M. Pollefeys, and F. Tombari. Unisdf: Unifying neural representations for high-fidelity 3d reconstruction of complex scenes with reflections. *arXiv preprint arXiv:2312.13285*, 2023. 3

[25] H. Wang, W. Hu, L. Zhu, and R. W. Lau. Inverse rendering of glossy objects via the neural plenoptic function and radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19999–20008, 2024. 3

[26] J. Wang, P. Wang, X. Long, C. Theobalt, T. Komura, L. Liu, and W. Wang. Neuris: Neural reconstruction of indoor scenes using normal priors. In *European Conference on Computer Vision*, pages 139–155. Springer, 2022. 3, 7

[27] 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. 1, 3, 4, 5

[28] Y. Wang, Q. Han, M. Habermann, K. Daniilidis, C. Theobalt, and L. Liu. Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3295–3306, 2023. 3

[29] T. Whelan, M. Goesele, S. J. Lovegrove, J. Straub, S. Green, R. Szeliski, S. Butterfield, S. Verma, R. A. Newcombe, M. Goesele, et al. Reconstructing scenes with mirror and glass surfaces. *ACM Trans. Graph.*, 37(4):102, 2018. 2, 4

[30] T. Wu, J.-M. Sun, Y.-K. Lai, and L. Gao. De-nerf: Decoupled neural radiance fields for view-consistent appearance editing and high-frequency environmental relighting. In *ACM SIGGRAPH 2023 conference proceedings*, pages 1–11, 2023. 3

[31] M. Xie, H. Cai, S. Shah, Y. Xu, B. Y. Feng, J.-B. Huang, and C. A. Metzler. Flash-splat: 3d reflection removal with flash cues and gaussian splats. 4

[32] Y. Yao, J. Zhang, J. Liu, Y. Qu, T. Fang, D. McKinnon, Y. Tsin, and L. Quan. Neilf: Neural incident light field for physically-based material estimation. In *European Conference on Computer Vision*, pages 700–716. Springer, 2022. 3

[33] 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. 1, 3

[34] C. Yeshwanth, Y.-C. Liu, M. Nießner, and A. Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023. 9

[35] Z.-X. Yin, J. Qiu, M.-M. Cheng, and B. Ren. Multi-space neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12407–12416, 2023. 1, 2, 4, 8, 9, 10

[36] Z. Yu, S. Peng, M. Niemeyer, T. Sattler, and A. Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *Advances in neural information processing systems*, 35:25018–25032, 2022. 3, 7

[37] Z. Yu, T. Sattler, and A. Geiger. Gaussian opacity fields: Efficient and compact surface reconstruction in unbounded scenes. *arXiv preprint arXiv:2404.10772*, 2024. 3

[38] J. Zeng, C. Bao, R. Chen, Z. Dong, G. Zhang, H. Bao, and Z. Cui. Mirror-nerf: Learning neural radiance fields for mirrors with whitted-style ray tracing. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4606–4615, 2023. 2, 3, 4, 8

[39] J. Zhang, Y. Yao, S. Li, J. Liu, T. Fang, D. McKinnon, Y. Tsin, and L. Quan. Neilf++: Inter-reflectable light fields for geometry and material estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3601–3610, 2023. 3

[40] K. Zhang, F. Luan, Q. Wang, K. Bala, and N. Snavely. Physg: Inverse rendering with spherical gaussians for physics-based material editing and relighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5453–5462, 2021. 1, 3

[41] R. Zhang, T. Luo, W. Yang, B. Fei, J. Xu, Q. Zhou, K. Liu, and Y. He. Refgaussian: Disentangling reflections from 3d gaussian splatting for realistic rendering. *arXiv preprint arXiv:2406.05852*, 2024. 2, 4