CD-RCM: Generalizable Continuous-Depth Novel View Synthesis for Reflectance Confocal Microscopy

Tooba Imtiaz[1], **Milind Rajadhyaksha**[2], **Kivanc Kose**[2,*], **Jennifer Dy**[1,*]
[1]Northeastern University, [2]Memorial Sloan Kettering Cancer Center
{imtiaz.t,j.dy}@northeastern.edu, {rajadhym,kosek}@mskcc.org
[*] denotes equal contribution

Abstract

Reflectance confocal microscopy (RCM) provides noninvasive, cellular-resolution "optical biopsies" of human skin *in vivo* by acquiring en-face images at successive depths, forming a sparse z-stack. Due to optical limitations, these stacks are anisotropic 3D volumes with lateral resolution ($0.5\mu m$) $\sim$6 times higher compared to axial resolution, which is defined by the optical sectioning ($3\mu m$), limiting the interpretation of tissue. Our goal is to provide continuous-depth visualization by interpolating intermediate sections and making the 3D volume isotropic. Such a representation permits arbitrary-direction sectioning, including histopathology-like cross-sectional examination, without requiring per-patient optimization. To that end, we introduce the first RCM-specific novel-view synthesis (NVS) approach, CD-RCM, a feedforward model that predicts realistic, unseen depths from sparsely sampled RCM stacks. Classical neural rendering methods focus on reconstruction from surface-level multi-view observations. In contrast to surface-level camera views, RCM can acquire optically sectioned en-face images of tissue beyond the surface up to $200\mu m$. However, during visualization of the RCM stacks, observations of the shallower sections (towards the surface) obscure the deeper ones. This unique axial imaging geometry and layer-dependent anatomical organization motivated our development of a tailored architectural and training framework that explicitly accounts for RCM's depth-resolved, occlusive imaging physics. Experiments demonstrate that CD-RCM achieves high-fidelity novel-view synthesis with sub-second inference time.

1 Introduction

Early detection and precise staging of skin disease rely on observing microstructural changes across depth, often beginning at anatomically meaningful boundaries such as the dermal–epidermal junction (DEJ). Optical sectioning microscopy techniques like Reflectance Confocal Microscopy (RCM) [Gonzalez et al., 2008] enable noninvasive *in vivo* imaging of skin with near-histologic resolution and are widely used for diagnosing tumors, inflammatory conditions, and guiding surgical margin assessment [Edwards et al., 2017, Huzaira et al., 2001, Wojarska et al., 2025, Richarz et al., 2022].

Clinicians acquire RCM data as *z-stacks* — sequences of en-face optical sections sampled at discrete depths. Although lateral resolution is high ($\sim$0.5 μm), axial resolution is governed by optical sectioning ($\sim$3 μm), resulting in strongly anisotropic volumes that limit depth-wise interpretation, particularly near diagnostically critical transitions such as the DEJ. Prior studies note the tradeoff between slice spacing and acquisition time, and the sensitivity of downstream algorithms to sampling density [Bozkurt et al., 2017, Ghanta et al., 2016]. Conventional interpolation methods (*e.g.,* spline-based) lack anatomical priors and fail to preserve depth-dependent skin morphology.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

In this work, we introduce *Continuous-Depth Novel-View Synthesis for RCM* (CD-RCM), a feedforward model for continuous-depth novel-view synthesis from sparse RCM stacks. Our approach leverages the structured axial acquisition process of optical sectioning microscopy and models depth variation as a queryable dimension. Unlike prior methods, CD-RCM predicts anatomically consistent intermediate slices in a single forward pass, without per-stack optimization, and generalizes across patients and imaging sites. Rather than performing optical super-resolution, CD-RCM exploits the substantial overlap between adjacent sections, acquired every $1.5\,\mu$m relative to an axial section thickness of approximately 3–5 μm, to densify sampling along depth. It can estimate intermediate optical sections consistent with structures already resolved by RCM, but cannot recover sub-resolution structures or spatial frequencies absent from the acquired measurements. The resulting volume is more densely and isotropically sampled, while its intrinsic optical resolution remains unchanged.

To preserve fine cellular structure, we further propose a *skin-specific perceptual loss*, based on a domain-adapted feature extractor trained on RCM data. This enables the model to capture microtextural patterns that are not well represented by perceptual losses trained on natural images.

In summary, we make the following contributions:

1. We introduce the first formulation of arbitrary-depth novel-view synthesis for RCM, explicitly addressing the anisotropic and see-through nature of RCM stacks.
2. We develop CD-RCM, a feedforward, depth-queryable model that incorporates the axial imaging physics of RCM and generalizes across skin sites without per-stack optimization.
3. Our experiments demonstrate that CD-RCM achieves high-fidelity novel-depth synthesis in a single inference pass, outperforming conventional interpolation baselines. In a blinded expert evaluation, CD-RCM received ratings comparable to acquired slices and was selected as the reconstruction most similar to ground truth in majority cases.

2 Related Work

Our work draws on two areas of prior research: algorithmic analysis of RCM stacks, and neural methods for 3D reconstruction and novel-view synthesis. We review each below, highlighting how existing methods address related problems but leave a gap at their intersection — namely, feedforward synthesis of arbitrary intermediate depths from sparsely sampled confocal stacks.

2.1 RCM stacks and downstream analysis

RCM enables non-invasive, *in vivo* optical sectioning of skin with near-cellular lateral resolution. Unlike conventional histopathology, which provides vertically sectioned, stained tissue spanning multiple layers, RCM produces grayscale *en face* images based on intrinsic reflectance contrast. The en face geometry and anisotropic resolution make interpretation challenging, as clinicians must mentally reconstruct cross-sectional structure from sparsely sampled depth stacks.

Prior computational work has focused on algorithmic analysis of discretely acquired stacks — including DEJ localization [Kose et al., 2020, Ghanta et al., 2016], strata delineation, and layer thickness estimation [Bozkurt et al., 2017, 2021] — treating stacks as fixed collections of optical sections rather than modifying or densifying the underlying volume.

From a visualization standpoint, RCM's anisotropic resolution yields inadequate reconstructions from any angle other than en face. Perpendicular sections analogous to histopathology can be constructed from stacks but suffer from sparse axial sampling; naïve interpolation introduces blocking artifacts and unrealistic structures. Dense 3D visualization [Song et al., 2024] could enhance structural understanding of tissue organization, but alternative volumetric modalities such as light-sheet microscopy [Liu et al., 2023] and LC-OCT [Latriglia et al., 2023] are either impractical for routine in vivo imaging or offer lower lateral resolution than RCM.

Crucially, each RCM optical section integrates signal from adjacent depths due to the finite axial point-spread function, yet this overlapping information has not been leveraged for dense axial reconstruction. We formulate continuous-depth reconstruction as a feedforward novel-view synthesis problem that exploits this property and the structured axial acquisition geometry to densify sparse stacks without needing per-stack optimization.

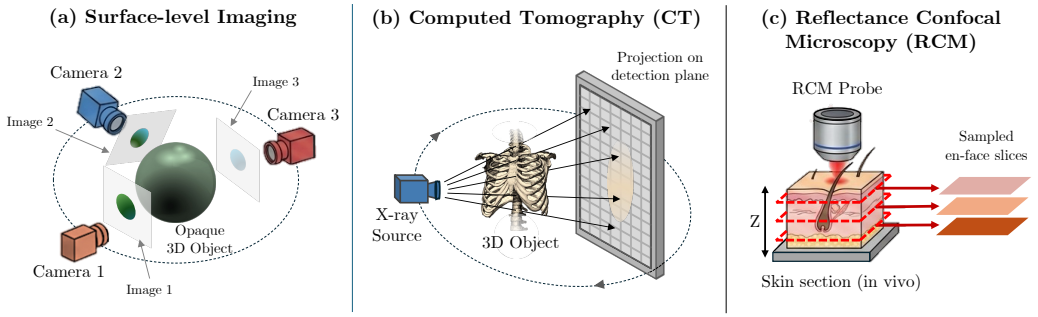

Figure 1: **Comparison of imaging paradigms.** (a) Surface-level imaging reconstructs opaque objects from multiple posed viewpoints with geometric parallax. (b) CT acquires transmission projections from multiple angles and reconstructs 3D structure from X-ray attenuation. (c) RCM performs axial optical sectioning, capturing discrete en-face slices of internal tissue at successive depths.

2.2 Neural rendering, 3D reconstruction and view synthesis

3D reconstruction and novel-view synthesis have progressed from implicit representations [Mescheder et al., 2019, Chibane et al., 2020] to NeRF [Mildenhall et al., 2020] and 3D Gaussian Splatting [Kerbl et al., 2023], enabling increasingly high-fidelity reconstruction with reduced optimization time. However, two limitations persist. Most methods require per-scene optimization, and while recent generalizable frameworks [Charatan et al., 2024, Zhang et al., 2024, Jin et al., 2025, Imtiaz et al., 2025] enable feedforward inference, they are primarily designed for surface-level reconstruction of opaque objects under known camera poses (Figure 1a), offering limited support for internal geometry.

Tomographic reconstruction offers a parallel perspective, as recovering 3D structure from sparse measurements is an ill-posed inverse problem (Figure 1b). Recent radiative Gaussian splatting approaches [Cai et al., 2024, Zha et al., 2024, Li et al., 2025] tailor rendering operators to X-ray physics for sparse-view CT, but like surface-based NVS methods, they require per-case optimization and do not generalize across subjects or hardware settings without retraining. Video frame interpolation [Reda et al., 2022] has also been adapted for intermediate-plane synthesis in 3D tomography [Gambini et al., 2024].

As illustrated in Figure 1, RCM occupies a distinct regime that differs fundamentally from both surface-level and CT imaging modalities. Unlike standard NVS, RCM captures internal tissue via axial optical sectioning with minimal viewpoint variation; unlike CT, its contrast is governed by depth-dependent tissue optics rather than geometric parallax. These properties motivate a specialized, generalizable approach for depth-conditioned synthesis of en face sections from sparse axial stacks.

Recent neural-field approaches have also addressed inverse problems in computational microscopy. Cao et al. [2024] jointly recover scene dynamics and structure from multi-shot measurements, while Kang et al. [2026] estimate optical aberrations and sample structure in two-photon microscopy. Unlike these per-sample, physics-based reconstruction methods, CD-RCM operates on already reconstructed optical sections and learns an amortized cross-stack model for continuous-depth synthesis.

3 Method

CD-RCM builds upon transformer-based novel-view synthesis architectures [Charatan et al., 2024, Zhang et al., 2024, Jin et al., 2025], adapting them to model internal skin structures captured by reflectance confocal microscopy. We describe our components below and highlight key differences from conventional surface-level NVS frameworks.

3.1 Overview

An RCM *stack* consists of N en-face grayscale images $\mathbf{I}_i \in \mathbb{R}^{H\times W}$, $i = 1 \dots N$, each comprising a 3 μm thick optical section of tissue. By moving the focusing depth through the tissue, the microscope translates the focal plane along the axial (z) direction at $>$1.5 μm steps with negligible lateral (x, y) motion, imaging up to 150–200 μm deep. Stacks typically contain 50–65 slices.

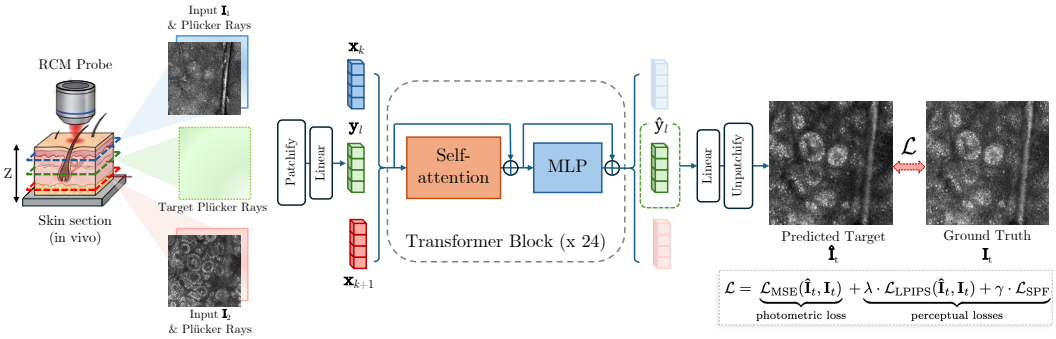

Figure 2: **Overview of CD-RCM.** Sparse input RCM slices and their Plücker ray embeddings are mapped to a unified token space and processed by a decoder-only transformer. Target ray tokens condition the synthesis; updated target tokens are decoded to produce $\hat{I}_t$. Training uses a weighted combination of photometric, LPIPS, and skin-specific perceptual feature loss $\mathcal{L}_{\text{SPF}}$.

RCM does not operate as a conventional camera-based imaging system. Instead, it raster-scans a focused beam using galvanometric mirrors and collects back-scattered light through a confocal pinhole, which enforces optical sectioning and provides depth selectivity. Since each slice corresponds to a fixed lateral field of view at a known axial position, we approximate the imaging process as a virtual pinhole camera undergoing pure z-axis translation (Section 3.2). This abstraction enables us to leverage Plücker ray embeddings and transformer architectures designed for multi-view geometry.

Given sparse inputs $\mathbf{I}_i$, CD-RCM synthesizes target slices $\mathbf{I}_t$ at arbitrary unseen depths (Figure 2) within the range of the collected stack. As illustrated in Figure 2, we map input slices and their corresponding Plücker embeddings into a unified token space, process them using a decoder-only transformer, and decode only the updated target tokens to generate $\hat{\mathbf{I}}_t$. We train the model using a combination of photometric reconstruction loss and perceptual supervision, including our proposed skin-specific perceptual feature loss.

Below, we describe the virtual camera parameters (Section 3.2), skin-specific perceptual feature extractor (Section 3.4), CD-RCM architecture (Section 3.3), and overall training objective (Section 3.5).

3.2 Virtual Camera Parameters

RCM forms images via a scanning laser and confocal pinhole that optically section tissue at controlled focal depths. Although no physical camera parameters are available, the acquisition process provides strong geometric priors, *i.e.,* consecutive slices share the same lateral field of view and differ primarily in axial position. Note that, unlike surface-level datasets, where pose estimation requires feature matching across viewpoints, RCM stacks exhibit near-perfect lateral alignment and systematic axial translation. Inter-slice correspondence arises from imaging different axial sections of the same underlying three-dimensional tissue structures (e.g., cells, vessels, and collagen), rather than from observing a static object under varying viewpoints. Consequently, the depth variation reflects progressive optical sectioning rather than geometric parallax. We exploit this by modeling acquisition as a virtual pinhole camera with fixed orientation translating along z.

Let slice index $i \in \{1, \ldots, N\}$ denote the depth position within a stack. We approximate the camera motion as a pure translation, *i.e.,* $\mathbf{c}_i = [0, 0, z_i]^T$, $z_i = i \cdot \Delta_z$. Here Δ_z is a constant axial step size. In practice, Δ_z corresponds to the physical spacing between optical sections (approximately 3μm), mapped to normalized model coordinates.

Assuming fixed camera orientation, the world-to-camera transform is

$$\mathbf{T}^{(i)}_{\text{w2c}} = \begin{bmatrix} \mathbf{I}_3 & \mathbf{c}_i \\ \mathbf{0}^\top & 1 \end{bmatrix} \tag{1}$$

where $\mathbf{I}_3$ denotes identity rotation. Neural rendering formulations typically adopt the camera-to-world convention. Since the transform is purely translational, its inverse admits a closed-form expression:

$$\mathbf{T}^{(i)}_{\text{c2w}} = \begin{bmatrix} \mathbf{I}_3 & -\mathbf{c}_i \\ \mathbf{0}^\top & 1 \end{bmatrix} \tag{2}$$

This matrix defines the camera pose associated with slice I_i.

Stack Canonicalization. Although poses are synthetically constructed, per-stack normalization improves numerical stability and learning consistency across stacks. Given a set of translations $\{\mathbf{c}_i\}$, we compute the mean camera center $\bar{\mathbf{c}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{c}_i$ and the maximum absolute translation $\mathbf{c}_\infty = \max_{i\in\{1,\dots,N\}} \|\mathbf{c}_i\|_\infty$ to standardize all poses via $\mathbf{c}_i \leftarrow (\mathbf{c}_i - \bar{\mathbf{c}})/\mathbf{c}_\infty$. This embeds each stack within a shared normalized coordinate system while preserving relative depth relations.

Camera Intrinsic Parameters Camera intrinsics K are similarly approximated using a fixed pinhole model:

$$\mathbf{K} = \begin{bmatrix} f_x & 0 & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3}$$

where focal lengths are set proportional to image resolution, *i.e.,* $f_x = f_y = \alpha H$, with α a constant scaling factor, and $u_0 = v_0 = \frac{H}{2}$. Altogether, this approximation provides a stable geometric reference without requiring physical calibration.

3.3 CD-RCM Architecture

Inspired by the decoder-only transformer design of LVSM [Jin et al., 2025], we train a feedforward depth-queryable model that renders intermediate slices between acquired slices. Unlike LVSM, which operates on surface views with geometric parallax, CD-RCM models purely axial variation and depth-conditioned reflectance.

Tokenizing Input Views. For each input view $\mathbf{I}_i$, we compute pixel-wise Plücker ray embeddings [Plücker, 1865] $\mathbf{P}_i \in \mathbb{R}^{H\times W\times 6}$ from the virtual camera parameters, and tokenize both into non-overlapping $p \times p$ patches following ViT [Dosovitskiy et al., 2021] to obtain $\mathbf{I}_{ij}$ and $\mathbf{P}_{ij}$, respectively. We channel-wise concatenate each pair, and map to d-dimensional input patch tokens $\mathbf{x}_{ij}$ via a linear layer:

$$\{\mathbf{x}_{ij}\}_{j=1}^{HW/p^2} = \text{Linear}_{\text{input}}(\text{Concat}(\mathbf{I}_{ij}, \mathbf{P}_{ij})) \in \mathbb{R}^d. \tag{4}$$

Lastly, we flatten these to obtain a 1D input token sequence $\mathbf{x}_K$ of length $K = NHW/p^2$.

Tokenizing Target Views. For each target depth $t = 1, \dots, T$, we similarly compute Plücker embeddings $\mathbf{P}_t$, patchify into $p \times p$ patches to obtain $\mathbf{P}_{tj}$, and map these to d-dimensional target tokens $\mathbf{y}_{tj}$ by a linear layer:

$$\{\mathbf{y}_{tj}\}_{j=1}^{HW/p^2} = \text{Linear}_{\text{target}}(\mathbf{P}_{tj}) \in \mathbb{R}^d. \tag{5}$$

We flatten these to form a 1D target token sequence $\mathbf{y}_L$ of length $L = THW/p^2$.

Transformer-based Decoder. We concatenate $[\mathbf{x}_K, \mathbf{y}_L]$ and process the sequence with a decoder-only transformer comprising 24 self-attention blocks (with QK-normalization [Henry et al., 2020]) and MLPs. Here, the term *decoder* refers to a GPT-style architecture [Radford et al., 2019] without causal masking, *i.e.,* attention is bidirectional across all tokens. After transformation, we retain only the updated target tokens and apply a linear layer followed by sigmoid activation to recover pixel intensities:

$$\hat{\mathbf{I}}_\ell = \text{Sigmoid}(\text{Linear}_{\text{out}}(\hat{\mathbf{y}}_\ell)) \in \mathbb{R}^{p^2} \tag{6}$$

Finally, we unpatchify (reshape) the intensity patches to recover the 2D synthesized target slices $\hat{\mathbf{I}}_t$.

3.4 Skin-Specific Perceptual Feature Extractor

Standard perceptual losses [Johnson et al., 2016, Zhang et al., 2018] emphasize natural-image semantics over the micro-textural patterns that dominate RCM imagery. To provide supervision sensitive to skin-specific structures (cellular morphology, speckle statistics, layer-dependent contrast), we train a domain-adapted perceptual backbone on a large auxiliary dataset of RCM mosaic crops.

We initialize a ViT [Dosovitskiy et al., 2021] from self-supervised DINOv3 [Siméoni et al., 2025] and adapt it to the skin domain via LoRA [Hu et al., 2022] applied to all attention projections:

$$\phi_s(\cdot) = \phi_{\text{DINOv3}}(\cdot; \theta_{\text{base}} + \Delta\theta_{\text{LoRA}}), \tag{7}$$

where only $\Delta\theta_{\text{LoRA}}$ is optimized. A teacher network ϕ_t is maintained via EMA *i.e.,* $\theta_t \leftarrow m\theta_t + (1-m)\theta_s$, with momentum coefficient m. We follow the standard DINO multi-crop training paradigm [Siméoni et al., 2025], where we generate 2 global (224×224) and 8 local (96×96) crops per image. The teacher processes global crops, while the student processes all crops, promoting local–global consistency and robustness. For training, we use cross-entropy distillation between student and teacher distributions, computing losses from both CLS and mean-pooled patch tokens. Once trained, ϕ_s is frozen and used as a skin-specific feature extractor.

3.5 Training Objective

We train CD-RCM using a weighted combination of complementary photometric and perceptual losses:

$$\mathcal{L} = \underbrace{\mathcal{L}_{\text{MSE}}(\hat{\mathbf{I}}_t, \mathbf{I}_t)}_{\text{photometric loss}} + \underbrace{\lambda \cdot \mathcal{L}_{\text{LPIPS}}(\hat{\mathbf{I}}_t, \mathbf{I}_t) + \gamma \cdot \mathcal{L}_{\text{SPF}}}_{\text{perceptual loss}}, \tag{8}$$

where λ and γ are scalar weights. $\mathcal{L}_{\text{MSE}}$ computes pixel-wise mean squared error between $\hat{\mathbf{I}}_t$ and $\mathbf{I}_t$, enforcing structural accuracy. $\mathcal{L}_{\text{LPIPS}}$ [Zhang et al., 2018] captures higher-order perceptual structure. Following [Imtiaz et al., 2025], we use LPIPS over the VGG-based loss [Johnson et al., 2016] of prior works [Zhang et al., 2024, Jin et al., 2025]. Our *skin-specific perceptual feature* loss,

$$\mathcal{L}_{\text{SPF}} = \frac{1}{T}\|\phi_s(\hat{\mathbf{I}}_t) - \phi_s(\mathbf{I}_t)\|_1, \tag{9}$$

complements $\mathcal{L}_{\text{LPIPS}}$ by enforcing consistency in a feature space adapted to dermatologic imagery, encouraging preservation of diagnostically relevant morphological details.

3.6 Training and Inference

During training, we sample sequences of M consecutive slices from each stack, use the first, middle, and last as conditioning inputs $\mathbf{I}_i$, and randomly select four intermediate slices as targets $\mathbf{I}_t$. This randomized sampling across varying depth separations prevents overfitting to fixed spacing and enables generalization to arbitrary depths. At inference, CD-RCM takes any three evenly spaced slices and predicts target views at arbitrary intermediate depths in a single feedforward pass — no per-stack optimization, iterative refinement, or perceptual feature computation is required. Effectively, this allows synthesis at continuous axial positions, which, in practice, corresponds to generating non-integer-index slices between discretely acquired sections.

4 Experiments

We describe our experimental setup (Section 4.1) and discuss the results obtained (Section 4.2). We also ablate key design choices in Section 4.3.

4.1 Experimental Setup

Dataset. We conduct our experiments using a dataset comprising 216 RCM stacks acquired *in vivo*. Each stack corresponds to a single skin site and contains between 50 and 65 en-face optical sections sampled along the axial (z) direction. Imaging was performed on healthy skin of volunteers without any special preparation, with images obtained from the arm and trunk of healthy volunteers aged 20–50 years. It is important to note that each stack originates from a distinct skin site, and even when acquired from the same subject, RCM images do not exhibit visual similarity across sites due to strong spatial and structural variability. As a result, stacks can be treated as independent samples. While subject-level metadata is unavailable, our train/test split is performed at the stack level and is effectively disjoint. We perform a post hoc audit for potential overlap between training and evaluation stacks reported in Appendix B.

Data were collected using a commercially available reflectance confocal microscope (VivaScope 1500, Caliber Imaging and Diagnostics, Rochester, NY), a system widely used in clinical imaging and dermatologic studies [Rajadhyaksha et al., 2017, Atak et al., 2023]. The microscope employs a stabilized imaging window that adheres to the region of interest, ensuring a fixed lateral field of view during stack acquisition.

Each slice is a single-channel intensity image of resolution 1000×1000 pixels, corresponding to a lateral resolution of approximately $0.5\,\mu m$. The optical sectioning thickness (axial resolution) of the imaging system is approximately $3\,\mu m$, while consecutive slices are acquired at $1.5\,\mu m$ intervals. This results in axial oversampling relative to the optical sectioning limit, yet the stacks remain strongly anisotropic due to the disparity between lateral and axial resolution.

Preprocessing. We register each stack using SIFT-based [Lowe, 2004] affine alignment implemented in Fiji [Schindelin et al., 2012] to reduce motion-induced inter-slice misalignment. After registration, we extract 960×960 center crops to remove boundary artifacts from the imaging window and resize to 512×512 via bilinear interpolation. The final five slices of each stack are excluded due to low signal content. We use a train-test split of approximately 70/30, yielding 156 training and 60 evaluation stacks.

Implementation Details. We train CD-RCM on four NVIDIA A6000 GPUs using a coarse-to-fine strategy [Zhang et al., 2024, Jin et al., 2025]. We first train CD-RCM from scratch for 20k steps at 256×256 resolution (batch size 8 per GPU), followed by 10k finetuning steps at 512×512 (batch size 4 per GPU). We set $M = 9$ for all reported results. At inference, $10\times$ densification of a full stack takes 0.8 seconds on a single A6000. Additional hyperparameters and architectural details are provided in the Appendix.

4.2 Results

We compare CD-RCM against four conventional interpolation methods: bilinear, B-spline, cubic spline, and Gaussian interpolation, and two learned baseline families: a NeRF-like coordinate MLP and FILM [Reda et al., 2022]. We evaluate photometric fidelity using PSNR and SSIM and perceptual similarity using LPIPS and $\mathcal{L}_{SPF}$. Baseline and evaluation details are provided in the Appendix.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	$\mathcal{L}_{\text{SPF}}$ ↓
Bilinear Interpolation	21.98	0.471	0.377	<u>0.153</u>
B-Spline Interpolation	22.02	0.471	<u>0.375</u>	0.161
Cubic Spline Interpolation	22.09	0.469	0.378	0.163
Gaussian Interpolation	<u>22.87</u>	<u>0.522</u>	0.477	0.164
MLP (NeRF)	19.10	0.445	0.520	0.480
FiLM zero-shot (pretrained)	20.24	0.380	0.440	0.156
FiLM fine-tuned (5 epochs)	22.83	0.510	0.496	0.181
CD-RCM (Ours)	**23.36**	**0.581**	**0.314**	**0.145**

Table 1: **Quantitative comparison.** CD-RCM outperforms classical interpolation and learned baselines across photometric (PSNR, SSIM) and perceptual (LPIPS, $\mathcal{L}_{SPF}$) metrics.

As shown in Table 1 and Figure 3, CD-RCM achieves the best performance across all quantitative metrics while producing reconstructions that most closely preserve the ground-truth tissue structure. Clinicians generally prefer reconstructions that preserve high-frequency details, even with mild noise, over overly smoothed images that obscure diagnostically relevant structures. Among classical baselines, Gaussian interpolation introduces noticeable blur. Although this smoothing can improve PSNR, which favors reduced variance, it degrades perceptual fidelity and obscures diagnostically relevant detail. The per-stack NeRF-like coordinate MLP performs poorly despite accurately fitting the observed slices, indicating that photometric optimization without a cross-stack anatomical prior is insufficient for reconstructing unseen tissue structure.

FiLM [Reda et al., 2022] exhibits a similar tradeoff between photometric and perceptual fidelity. Zero-shot FILM produces relatively sharp predictions but introduces inter-slice discontinuities and morphologically implausible structures. Fine-tuning FiLM on RCM substantially improves PSNR and SSIM by producing smoother predictions, but suppresses fine morphological detail and worsens both

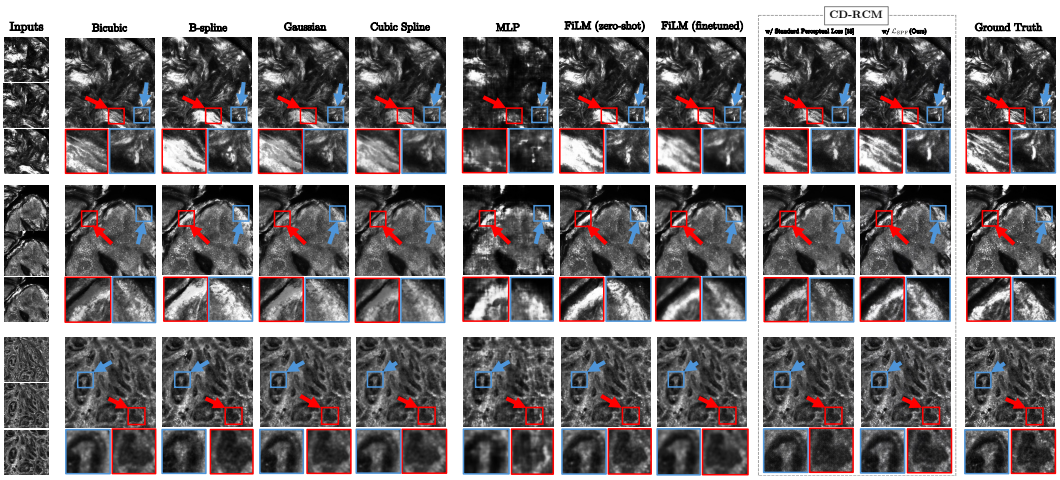

Figure 3: **Qualitative comparison of novel-depth synthesis methods.** We compare CD-RCM with classical interpolation and learned baselines. CD-RCM better preserves fine cellular structures and depth-dependent morphology, producing predictions closest to the ground truth. Insets highlight representative structural differences and reconstruction artifacts.

perceptual metrics. In contrast, CD-RCM preserves layered tissue structure and fine cellular features while maintaining continuity across depth, achieving the strongest photometric and perceptual fidelity.

Cross-sectional views of densified stacks Figure 4 shows the cross-sectional and arbitrary-plane visualization of CD-RCM–densified stacks. Using CD-RCM, we densify RCM stacks so that axial sampling matches the actual sample dimensions, enabling isotropic 3D visualization. We show sagittal and coronal cross-sections (left) and arbitrary oblique cuts at specified azimuth, elevation, and offset angles (right) for two example stacks. Compared to the sparsely sampled input stacks, CD-RCM produces markedly smoother structural transitions along the z-axis, reducing the staircase artifacts visible in the original cross-sections and revealing continuous tissue morphology that may facilitate identification of clinically relevant landmarks.

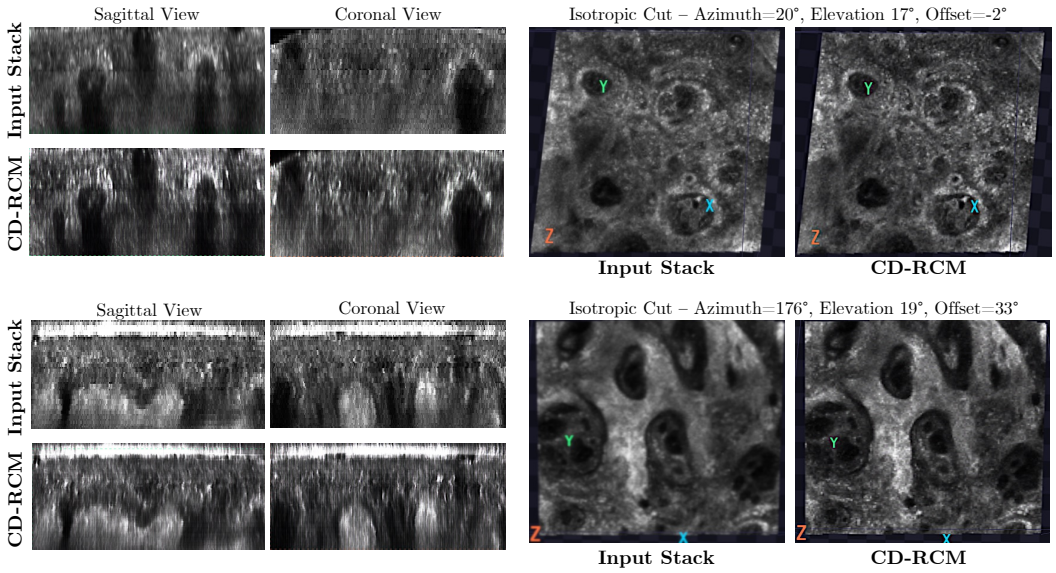

Figure 4: **Cross-sectional and arbitrary-plane visualization of densified stacks.** We densify the input stacks using CD-RCM to match the axial and lateral resolution of the actual sample, enabling isotropic 3D visualization. Sagittal, coronal, and oblique cross-sections show markedly smoother structural transitions compared to the sparsely sampled input stacks.

Training objective	PSNR ↑	SSIM ↑	LPIPS ↓	$\mathcal{L}_{\text{SPF}}$ ↓
($\mathcal{L}_{\text{MSE}}$) + $\mathcal{L}_{\text{LPIPS}}$	21.60	0.480	0.343	0.273
+ Perceptual Loss [Johnson et al., 2016]	22.56	0.547	0.386	0.258
+ $\mathcal{L}_{\text{LPIPS}}$ + Perceptual Loss [Johnson et al., 2016]	<u>22.65</u>	<u>0.567</u>	<u>0.329</u>	<u>0.216</u>
+ $\mathcal{L}_{\text{LPIPS}}$ + $\mathcal{L}_{\text{SPF}}$ (multi-feature)	22.39	0.522	0.338	0.246
+ $\mathcal{L}_{\text{LPIPS}}$ + $\mathcal{L}_{\text{SPF}}$(eq. (9))	**23.35**	**0.597**	**0.288**	**0.172**

Table 2: **Ablation of training objectives.** We compare different perceptual loss combinations for training CD-RCM. Best and second-best results are **bolded** and <u>underlined</u>, respectively. Our proposed $\mathcal{L}_{\text{SPF}}$ (eq. (9)) yields the best performance across all metrics.

4.3 Ablation Study

We ablate key design choices of CD-RCM, mainly the training objective and the design of our skin-specific perceptual loss $\mathcal{L}_{\text{SPF}}$. All models are trained at 256×256 resolution for 20k steps.

Training Objective. Table 2 reports performance across perceptual loss configurations, with $\mathcal{L}_{\text{MSE}}$ included in all experiments. LPIPS alone yields the weakest results. The VGG-based perceptual loss of Johnson et al. [2016] performs better, and combining it with LPIPS further boosts performance. However, replacing the VGG-based loss with our proposed $\mathcal{L}_{\text{SPF}}$ (eq. (9)) yields the best overall performance, simultaneously improving all four metrics. Figure 3 (columns 5 and 6) qualitatively confirms that $\mathcal{L}_{\text{SPF}}$ produces higher-fidelity reconstructions that more closely match the ground truth.

Design of $\mathcal{L}_{\text{SPF}}$. We also ablate the design of $\mathcal{L}_{\text{SPF}}$ by comparing our final-layer MAE formulation (eq. (9)) against a multi-layer feature variant analogous to the standard perceptual loss [Johnson et al., 2016]. As shown in the last two rows of Table 2, the single-layer design outperforms the multi-layer alternative across all metrics.

CD-RCM demonstrates that feedforward novel-depth synthesis can efficiently densify sparse RCM stacks without per-stack optimization, producing anatomically plausible intermediate slices in a single forward pass. More broadly, our results highlight the potential of adapting neural rendering methods to domain-specific imaging physics in computational microscopy. A promising direction is to accelerate stack acquisition by sparsely sampling tissue sections and reconstructing intermediate slices with CD-RCM. This could speed up clinical imaging and reduce motion artifacts by shortening the time patients must remain still (see Limitations on registration).

4.4 Expert Evaluation

Image-similarity metrics do not fully capture morphological fidelity, particularly when smoothing improves PSNR while suppressing fine tissue structure. We therefore conducted a blinded evaluation with an RCM expert. Images were anonymized and presented in randomized order. The study comprised two complementary tasks, namely:

- **Image quality.** The reader independently evaluated 25 en face slices: five acquired slices and five each from CD-RCM, fine-tuned FILM, Gaussian, and B-spline interpolation. For each slice, the reader rated 1–3 its suitability for evaluating tissue morphology, presence of morphologically implausible structures or artifacts, and confidence in interpretation.
- **Similarity to ground truth.** For 25 additional cases, the reader was shown two acquired slices surrounding a withheld ground-truth slice and reconstructions from CD-RCM, fine-tuned FILM, Gaussian, and B-Spline interpolation in randomized order. The reader selected the reconstruction most similar to the ground truth. To evaluate generalization beyond the training spacing, inputs were separated by 6 slices ($9\,\mu$m), $1.5\times$ the spacing used during training, with the target midway between them.

As shown in Table 3, CD-RCM received ratings comparable to acquired slices for suitability (2.40 vs. 2.40), implausibility (1.60 vs. 1.80), and confidence (2.60 vs. 2.40). Gaussian interpolation and fine-tuned FILM received lower suitability and confidence ratings, consistent with the loss of fine

Method	**Suitability↑**	**Implausibility↓**	**Confidence↑**
Acquired (ground truth)	2.40 (0.89)	1.80 (0.84)	2.40 (0.89)
B-Spline Interpolation	2.60 (0.55)	1.80 (0.84)	2.80 (0.45)
Gaussian Interpolation	1.60 (0.55)	2.00 (0.71)	2.00 (0.71)
FILM fine-tuned	1.80 (0.45)	2.00 (0.71)	2.00 (0.71)
CD-RCM (Ours)	2.40 (0.55)	1.60 (0.55)	2.60 (0.55)

Table 3: **Expert evaluation of image quality.** For 5 images per method, we report the blinded ratings of individual slices on a 1-3 scale. CD-RCM was rated comparably to ground-truth slices.

Method	Times chosen	Rate
B-Spline Interpolation	3	12.0%
Gaussian Interpolation	0	0.0%
FILM fine-tuned	0	0.0%
CD-RCM (Ours)	22	**88.0%**

Table 4: **Expert evaluation of similarity to ground truth.** For 25 cases, we report the frequency with which each reconstruction was judged most similar to the withheld ground truth. CD-RCM was rated most similar in the majority of cases.

morphological detail from smoothing observed in Section 4.2. Given the single reader and small sample size, we treat these ratings as indicative rather than as a formal non-inferiority test.

In the ground-truth comparison reported in Table 4, CD-RCM was selected in 22/25 cases, compared with 3/25 for linear interpolation and none for FILM or Gaussian interpolation. Notably, FILM achieved the highest PSNR in 9/25 cases, whereas CD-RCM achieved the lowest LPIPS in all 25, further illustrating that expert judgments aligned better with perceptual than pixel-wise similarity.

4.5 Limitations

Our pipeline uses SIFT-based registration to reduce inter-slice motion, but residual misalignment persists in some stacks and can degrade reconstruction quality. Moreover, curating a larger and more diverse dataset, spanning a wider range of skin types, and filtering out slices with low signal content may further improve robustness and performance.

4.6 Future Work

This work represents a first-of-its-kind, proof-of-concept study for continuous-depth reconstruction from sparse RCM stacks. Our goal is to motivate further development toward clinical applications, including larger-scale data collection and comprehensive human studies.

Future work will integrate CD-RCM into interactive clinical visualization tools to evaluate whether densified stacks improve downstream analyses, such as DEJ localization and strata delineation. We also aim to extend CD-RCM to RCM mosaics acquired at consecutive depths to enable dense visualization over larger tissue regions. Finally, adapting our framework to other optical sectioning modalities, such as line-field confocal OCT (LC-OCT), could broaden its impact beyond RCM.

5 Conclusion

We introduced CD-RCM, the first feedforward framework for continuous-depth novel-view synthesis tailored to reflectance confocal microscopy. By modeling RCM acquisition as axial translation and using a decoder-only transformer, CD-RCM synthesizes anatomically plausible en-face slices at arbitrary depths without per-stack optimization and with sub-second inference. We further introduced a skin-specific perceptual loss adapted to the micro-textural patterns of RCM. Quantitatively, CD-RCM outperforms interpolation baselines across photometric and perceptual metrics, with ablations supporting the proposed design choices. A blinded expert evaluation further supports the morphological fidelity of the synthesized slices, with CD-RCM rated comparably to acquired RCM images and most similar to withheld ground truth. By enabling axial densification and arbitrary-orientation virtual sectioning, CD-RCM takes a step toward making continuous 3D visualization of *in vivo* skin a practical clinical tool.

Acknowledgments

The reading study was conducted by Dr. Miguel Cordova, an RCM imaging expert with approximately 20 years of experience acquiring and interpreting reflectance confocal microscopy images at the Dermatology Service of Memorial Sloan Kettering Cancer Center (MSKCC).

This work is supported by grants DoD/USAMRAA HT94252410553 and DoD/USAMRAA HT94252410554, NIH/NCI R01CA199673, NIH/NCI U24-CA264369, R01CA293974, and NIH/NCI Cancer Center Support Grant P30 CA008748.

References

Mehmet Fatih Atak, Banu Farabi, Cristian Navarrete-Dechent, Gennady Rubinstein, Milind Rajadhyaksha, and Manu Jain. Confocal microscopy for diagnosis and management of cutaneous malignancies: clinical impacts and innovation. *Diagnostics*, 13(5):854, 2023.

A Bozkurt, K Kose, C Alessi-Fox, JG Dy, DH Brooks, and M Rajadhyaksha. Unsupervised delineation of stratum corneum using reflectance confocal microscopy and spectral clustering. *Skin Research and Technology*, 23(2):176–185, 2017.

Alican Bozkurt, Kivanc Kose, Jaume Coll-Font, Christi Alessi-Fox, Dana H Brooks, Jennifer G Dy, and Milind Rajadhyaksha. Skin strata delineation in reflectance confocal microscopy images using recurrent convolutional networks with attention. *Scientific reports*, 11(1):12576, 2021.

Yuanhao Cai, Yixun Liang, Jiahao Wang, Angtian Wang, Yulun Zhang, Xiaokang Yang, Zongwei Zhou, and Alan Yuille. Radiative gaussian splatting for efficient x-ray novel view synthesis. In *European Conference on Computer Vision*, pages 283–299. Springer, 2024.

Ruiming Cao, Nikita S Divekar, James K Nunez, Srigokul Upadhyayula, and Laura Waller. Neural space–time model for dynamic multi-shot imaging. *Nature Methods*, 21(12):2336–2341, 2024.

David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *CVPR*, pages 19457–19467, 2024.

Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174*, 2016.

Julian Chibane, Gerard Pons-Moll, et al. Neural unsigned distance fields for implicit function learning. *Advances in Neural Information Processing Systems*, 33:21638–21652, 2020.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. volume 35, pages 16344–16359, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

SJ Edwards, G Osei-Assibey, R Patalay, V Wakefield, and C Karner. Diagnostic accuracy of reflectance confocal microscopy using vivascope for detecting and monitoring skin lesions: a systematic review. *Clinical and Experimental Dermatology*, 42(3):266–275, 2017.

Esmaeil Enjilela, Ting-Yim Lee, Gerald Wisenberg, Patrick Teefy, Rodrigo Bagur, Ali Islam, Jiang Hsieh, and Aaron So. Cubic-spline interpolation for sparse-view ct image reconstruction with filtered backprojection in dynamic myocardial perfusion imaging. *Tomography*, 5(3):300, 2019.

Laura Gambini, Cian Gabbett, Luke Doolan, Lewys Jones, Jonathan N Coleman, Paddy Gilligan, and Stefano Sanvito. Video frame interpolation neural network for 3d tomography across different length scales. *Nature Communications*, 15(1):7962, 2024.

Sindhu Ghanta, Michael I Jordan, Kivanc Kose, Dana H Brooks, Milind Rajadhyaksha, and Jennifer G Dy. A marked poisson process driven latent shape model for 3d segmentation of reflectance confocal microscopy image stacks of human skin. *IEEE Transactions on Image Processing*, 26(1): 172–184, 2016.

Salvador Gonzalez, M. Gill, and Allan C. Halpern, editors. *Reflectance Confocal Microscopy of Cutaneous Tumors: An Atlas with Clinical, Dermoscopic and Histological Correlations*. CRC Press, 1 edition, 2008. doi: 10.3109/9780203091562.

Alex Henry, Prudhvi Raj Dachapally, Shubham Shantaram Pawar, and Yuxuan Chen. Query-key normalization for transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4246–4253, 2020.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

Misbah Huzaira, Francisca Rius, Milind Rajadhyaksha, R. Rox Anderson, and Salvador González. Topographic variations in normal skin, as viewed by in vivo reflectance confocal microscopy. *Journal of Investigative Dermatology*, 116(6):846–852, 2001. ISSN 0022-202X. doi: https://doi.org/10.1046/j.0022-202x.2001.01337.x. URL https://www.sciencedirect.com/science/article/pii/S0022202X15412552.

Tooba Imtiaz, Lucy Chai, Kathryn Heal, Xuan Luo, Jungyeon Park, Jennifer Dy, and John Flynn. Lvt: Large-scale scene reconstruction via local view transformers. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pages 1–12, 2025.

Haian Jin, Hanwen Jiang, Hao Tan, Kai Zhang, Sai Bi, Tianyuan Zhang, Fujun Luan, Noah Snavely, and Zexiang Xu. Lvsm: A large view synthesis model with minimal 3d inductive bias. In *ICLR*, 2025.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

Iksung Kang, Hyeonggeon Kim, Ryan Natan, Qinrong Zhang, Stella X Yu, and Na Ji. Adaptive optical correction for in vivo two-photon fluorescence microscopy with neural fields. *Nature Methods*, pages 1–10, 2026.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM TOG*, 42(4):139–1, 2023.

Kivanc Kose, Alican Bozkurt, Christi A Fox, Melissa Gill, Dana Brooks, Jennifer Dy, and Milind Rajadhyaksha. Detection of the dej and segmentation of its morphological patterns in rcm images of melanocytic skin lesions. In *Microscopy Histopathology and Analytics*, pages MW2A–1. Optica Publishing Group, 2020.

Flora Latriglia, Jonas Ogien, Clara Tavernier, Sébastien Fischman, Mariano Suppa, Jean-Luc Perrot, and Arnaud Dubois. Line-field confocal optical coherence tomography (lc-oct) for skin imaging in dermatology. *Life*, 13(12), 2023. ISSN 2075-1729. doi: 10.3390/life13122268. URL https://www.mdpi.com/2075-1729/13/12/2268.

Thomas Martin Lehmann, Claudia Gonner, and Klaus Spitzer. Survey: Interpolation methods in medical image processing. *IEEE transactions on medical imaging*, 18(11):1049–1075, 2002.

Yingtai Li, Xueming Fu, Han Li, Shang Zhao, Ruiyang Jin, and S Kevin Zhou. 3dgr-ct: Sparse-view ct reconstruction with a 3d gaussian representation. *Medical Image Analysis*, 103:103585, 2025.

Jonathan T.C. Liu, Adam K. Glaser, Chetan Poudel, and Joshua C. Vaughan. Nondestructive 3d pathology with light-sheet fluorescence microscopy for translational research and clinical assays. *Annual Review of Analytical Chemistry*, 16(Volume 16, 2023):231–252, 2023. ISSN 1936-1335. doi: https://doi.org/10.1146/annurev-anchem-091222-092734. URL https://www.annualreviews.org/content/journals/10.1146/annurev-anchem-091222-092734.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4460–4470, 2019.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4): 1–15, 2022.

Julius Plücker. Xvii. on a new geometry of space. *Philosophical Transactions of the Royal Society of London*, (155):725–791, 1865.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Milind Rajadhyaksha, Ashfaq Marghoob, Anthony Rossi, Allan C. Halpern, and Kishwer S. Nehal. Reflectance confocal microscopy of skin in vivo: From bench to bedside. *Lasers in Surgery and Medicine*, 49(1):7–19, 2017. doi: https://doi.org/10.1002/lsm.22600. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/lsm.22600`.

Fitsum Reda, Janne Kontkanen, Eric Tabellion, Deqing Sun, Caroline Pantofaru, and Brian Curless. Film: Frame interpolation for large motion. In *European Conference on Computer Vision*, pages 250–266. Springer, 2022.

N. A. Richarz, A. Boada, A. Jaka, J. Bassas, C. Ferrándiz, J. M. Carrascosa, and O. Yélamos. Challenges for new adopters in pre-surgical margin assessment by handheld reflectance confocal microscope of basal cell carcinoma: A prospective single-center study. *Dermatology Practical & Conceptual*, 12(4):e2022162, Oct 2022. doi: 10.5826/dpc.1204a162.

Johannes Schindelin, Ignacio Arganda-Carreras, Erwin Frise, Verena Kaynig, Mark Longair, Tobias Pietzsch, Stephan Preibisch, Curtis Rueden, Stephan Saalfeld, Benjamin Schmid, et al. Fiji: an open-source platform for biological-image analysis. *Nature methods*, 9(7):676–682, 2012.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Andrew H. Song, Mane Williams, Drew F.K. Williamson, Sarah S.L. Chow, Guillaume Jaume, Gan Gao, Andrew Zhang, Bowen Chen, Alexander S. Baras, Robert Serafin, Richard Colling, Michelle R. Downes, Xavier Farré, Peter Humphrey, Clare Verrill, Lawrence D. True, Anil V. Parwani, Jonathan T.C. Liu, and Faisal Mahmood. Analysis of 3d pathology samples using weakly supervised ai. *Cell*, 187(10):2502–2520.e17, 2024. ISSN 0092-8674. doi: https://doi.org/10.1016/j.cell.2024.03.035. URL `https://www.sciencedirect.com/science/article/pii/S0092867424003519`.

Monika Wojarska, Klaudia Kokot, Paulina Bernecka, Natalia Domańska, Agata Libik, Dana Bunevich, Dominika Nowakowska, Magdalena Dzido, Wiktoria Borzyszkowska, Wojciech Kazimierczak, and Jerzy Jankau. In vivo confocal microscopy in the surgical treatment of keratinocyte carcinomas: A systematic review. *Journal of Clinical Medicine*, 14(16), 2025. ISSN 2077-0383. doi: 10.3390/jcm14165779. URL `https://www.mdpi.com/2077-0383/14/16/5779`.

Ruyi Zha, Tao Jun Lin, Yuanhao Cai, Jiwen Cao, Yanhao Zhang, and Hongdong Li. R^2-gaussian: Rectifying radiative gaussian splatting for tomographic reconstruction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. Gs-lrm: Large reconstruction model for 3d gaussian splatting. In *ECCV*, pages 1–19. Springer, 2024.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims in the abstract and introduction accurately reflect the contributions and scope of the paper. We present a feedforward model for continuous-depth reconstruction from sparse RCM stacks and demonstrate improvements over classical interpolation baselines. We carefully frame the work as a proof-of-concept on a limited dataset and avoid making claims beyond the evaluated setting.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are detailed in Section 4.5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The complete model architecture is detailed in Section 3. The details of the dataset collection and experimental setup are discussed in Section 4.1 and Section A.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We are unable to release the data at this point. However, we plan on releasing the model code and trained checkpoints post-acceptance. The paper, to the best of our knowledge, sufficiently describes the details needed to reproduce our architecture and experiments.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The data collection/splits, as well as implementation details, are discussed in Section 4.1. The details of hyperparameters are provided in Section A.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The paper reports quantitative metrics averaged over the test set but does not include error bars or statistical significance analysis. We focus on consistent trends across multiple metrics and qualitative comparisons. This work serves as a proof-of-concept pilot on a smaller medical dataset, to motivate large-scale data collection and enable more thorough analysis in future studies on RCM data and similar medical imaging modalities.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The details about compute requirements for training and inference are provided under Implementation Details of **??**. The inference time is also discussed under **??**. The training times for both skin-specific perceptual feature extraction and CD-RCM are reported in Section A.1 and Section A.2, respectively.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The collected dataset is fully anonymized with no metadata on subject identity. The patient data is not published, further eliminating any privacy concerns.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: A detailed discussion on the broader impacts of this work is available in Appendix Section F.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The work does not involve releasing a dataset or model with a high risk of misuse, such as generative models for media synthesis or large-scale scraped data. CD-RCM is developed for biomedical imaging and operates on specialized reflectance confocal microscopy (RCM) data, which is not broadly accessible or applicable outside this domain. The model is intended for research and visualization purposes rather than deployment in open-ended generative settings. As such, no specific safeguards for high-risk misuse are required.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Our code builds on top of LVSM [Jin et al., 2025] (BY-NC-SA), which is explicitly credited in the main paper.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The paper introduces a new model and training pipeline for continuous-depth reconstruction. We provide detailed descriptions of the model architecture, training procedure, and evaluation protocol in the main paper and appendix. These details are sufficient for reproducibility. Any released code or supplementary materials will include documentation to facilitate understanding and reuse.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: This work does not involve crowdsourcing experiments or interaction with human subjects. The dataset used in this study was collected in prior work by our collaborators; all data were fully anonymized and obtained with appropriate consent for research use. Therefore, no participant instructions, screenshots, or compensation details apply to this paper.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This work does not require dedicated, direct interaction with human subjects. The dataset used in this study was collected in prior work by our collaborators, with appropriate consent and ethical approvals obtained at the time of data collection. All data used in this paper are fully anonymized and were provided for research use, and no additional risks to participants were introduced by this work.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The paper does not use large language models (LLMs) as part of the core methodology. Any use of LLMs was limited to writing or editing assistance and does not affect the scientific content, methodology, or results.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.

Appendix

A Further Implementation Details

A.1 Skin-specific perceptual feature extractor

For training the skin-specific perceptual feature extractor ϕ_s (Sec. 3.5), we use LoRA with rank 16 and optimize using AdamW Loshchilov and Hutter [2017] with a constant learning rate of 1×10^{-4}. The teacher network is updated using EMA with momentum $m = 0.9995$. The projection MLP h has a bottleneck dimension of 256 and an output dimension of 4096. The teacher and student temperatures in Eqs. 9 and 10 are set to $\tau_t = 0.04$ and $\tau_s = 0.1$, respectively. We train the model for 4 epochs using one NVIDIA A6000 GPU, which takes 3 days.

A.2 CD-RCM

For training CD-RCM (Sec. 3.6), we tokenize images and Plücker embeddings using a patch size $p = 8$, and set the embedding dimension (Eqs. 6 and 7) $d = 768$ of the linear projectors. For optimization, we use AdamW Loshchilov and Hutter [2017] with $\beta_1 = 0.9$, $\beta_2 = 0.95$, and a weight decay of 0.05 applied to all parameters except LayerNorm layers. We employ a cosine learning rate schedule with linear warmup, using 2000 warmup steps and a peak learning rate of 4×10^{-4} for both training resolutions. The loss weights in Eq. 12 are set to $\lambda = 0.5$ and $\gamma = 0.05$. The full model has 170.8M trainable parameters. We train CD-RCM using four NVIDIA A6000 GPUs. The first stage of training (*i.e.,* at 256 resolution) takes 2 days, and finetuning at 512 resolution takes another 2 days.

Following Jin et al. [2025], we use gradient skipping for gradient norm > 5.0 and gradient-clipping at 1.0 to stabilize training. Lastly, to boost training and inference efficiency, we use FlashAttention-v2 Dao et al. [2022] within the attention blocks, gradient checkpointing Chen et al. [2016], and mixed-precision training with BF16 data type.

B Dataset Overlap Audit

As stated in Section 4, subject identifiers were not retained with the RCM stacks. Therefore, although the train/test split was performed at the stack level, we cannot guarantee that different skin sites from the same volunteer do not appear across both splits. We therefore conducted a post hoc similarity analysis to assess whether duplicate, adjacent, or overlapping acquisitions of the same skin site were present across the training and evaluation sets.

We embedded all training and evaluation slices using the skin-specific feature extractor from Section 3.5 and computed nearest-neighbor cosine similarity for each of the 60 evaluation stacks against the training set. One evaluation stack had very high similarity (0.98) to a training stack, and visual inspection showed matching ring and cellular structures under a lateral shift, consistent with an overlapping acquisition site. Seven additional stacks had similarities of 0.75-0.85, but no corresponding anatomical structures were observed, suggesting feature-space similarity rather than true overlap.

To assess the effect of these potentially overlapping cases, we repeated the aggregate evaluation after excluding these 8 identified stacks. Performance on the remaining 52 evaluation stacks was essentially unchanged, and the excluded stacks did not show anomalously high reconstruction performance relative to the full evaluation set. Thus, while the absence of subject identifiers prevents us from establishing a strictly subject-disjoint split, this analysis indicates that detectable same-site or near-duplicate overlap does not account for the reported performance. "'latex

C Additional Ablation

Positional encoding. CD-RCM conditions each token on the Plücker ray map $\mathbf{P}_i$ (eqs. (4) and (5)). Because the virtual cameras differ only by axial translation, $\mathbf{P}_i$ is fully determined by pixel position (u, v) and canonicalized depth z_i. To test whether this representation is important, we replace $\mathbf{P}_i$ with a learned multiresolution hash encoding [Müller et al., 2022] of (u, v, z_i) while keeping the remaining architecture and training setup fixed.

Positional encoding	PSNR↑	SSIM↑	LPIPS↓	$\mathcal{L}_{\text{SPF}}$ ↓
Plücker ray map (default)	22.82	0.574	**0.281**	**0.156**
Multiresolution hash [Müller et al., 2022]	**23.10**	**0.605**	0.298	0.187

Table 5: **Ablation of positional encoding.** Both models are evaluated after 10k training steps on 300 windows from the 60 evaluation stacks; only the encoding of (u, v, z) differs.

We normalize (u, v, z_i) to $[0, 1]^3$ and use a 16-level hash grid with resolutions increasing geometrically from 16 to 256 and two features per level, yielding a 32-channel positional map. The hash-grid parameters are learned jointly with the transformer.

Both variants are trained on 256×256 inputs using the full objective ($\mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{LPIPS}} + \mathcal{L}_{\text{SPF}}$) and evaluated after 10k steps of the same 20k-step schedule. As shown in Table 5, the two encodings perform comparably overall: the hash encoding slightly improves PSNR and SSIM, while Plücker rays achieve better perceptual scores. This suggests that CD-RCM does not depend critically on Plücker coordinates, and that alternative continuous depth encodings can provide similar performance.

D Baselines Methods

We adopt four well-established classical interpolation techniques widely adopted for processing medical imaging data as the baselines in our experiments Enjilela et al. [2019], Lehmann et al. [2002]. For completeness, we also evaluate two learning-based baselines: a per-stack coordinate MLP, representative of NeRF-style per-scene optimization [Mildenhall et al., 2020], and FILM [Reda et al., 2022], a state-of-the-art video frame interpolation network, which we evaluate both zero-shot and after fine-tuning on our training stacks.

We additionally evaluated a diffusion-based model for intermediate slice prediction, but it failed to reconstruct high-frequency structures. This is consistent with the known data requirements of diffusion models and their sensitivity to noisy inputs, which are characteristic of RCM imagery.

Below, we explain each baseline:

D.1 Bilinear Interpolation

Bilinear interpolation estimates each intermediate slice as a per-pixel linear blend of the two acquired slices that bracket the target depth. Since the slices of a registered stack are laterally aligned, no in-plane resampling is required, and the method reduces to linear interpolation along the axial direction. Given the bracketing slices $I(z_a, x, y)$ and $I(z_b, x, y)$ with $z_a < z < z_b$, the interpolated intensity at depth z is

$$\hat{I}(z, x, y) = (1 - \tau)\, I(z_a, x, y) + \tau\, I(z_b, x, y), \qquad \tau = \frac{z - z_a}{z_b - z_a}. \tag{10}$$

D.2 B-Spline Interpolation.

B-spline interpolation models the intensity along the axial direction as a smooth spline function and evaluates it at intermediate depths. For our case, it treats the three input slices as samples of a smooth intensity function along the axial direction and estimates missing slices by applying third-order spline interpolation independently at each pixel location. For each pixel location (x, y), let the observed slices be $I(z_i, x, y)$ for $i \in \{1, \ldots, N\}$. The interpolated intensity at depth z is computed as

$$\hat{I}(z, x, y) = \sum_k c_k(x, y)\, \beta_3(z - k), \tag{11}$$

where $\beta_3(\cdot)$ denotes the cubic B-spline basis function and $c_k(x, y)$ are spline coefficients derived from the observed samples. In practice, we implement this interpolation using `map_coordinates` of `scipy.ndimage` library with spline order 3, which evaluates the cubic B-spline along the depth axis independently for each pixel.

D.3 Cubic Spline Interpolation.

Cubic spline interpolation fits a natural cubic spline along the depth axis for each pixel independently, using the three observed slices as control points. Given observed samples $(z_i, I(z_i, x, y))$, a piecewise cubic polynomial is constructed, such that

$$S(z) = \begin{cases} a_1 z^3 + b_1 z^2 + c_1 z + d_1, & z_1 \le z < z_2 \\ \vdots & \\ a_{N-1} z^3 + b_{N-1} z^2 + c_{N-1} z + d_{N-1}, & z_{N-1} \le z \le z_N \end{cases} \tag{12}$$

such that

$$S(z_i) = I(z_i, x, y), \quad S'(z_i^-) = S'(z_i^+), \quad S''(z_i^-) = S''(z_i^+), \tag{13}$$

with natural boundary conditions

$$S''(z_1) = S''(z_N) = 0. \tag{14}$$

Intermediate slices are obtained by evaluating $S(z)$ at the desired depths z.

We implement this method using `CubicSpline` function of `scipy.interpolate` with boundary condition set to `natural`, *i.e.,*, the second derivative at curve ends are zero (eq. (14)).

D.4 Gaussian Interpolation.

Gaussian interpolation first generates intermediate slices using cubic B-spline interpolation and then applies spatial Gaussian smoothing to each synthesized slice. Let $\tilde{I}(z, x, y)$ denote the B-spline interpolated slice. The final result is obtained by convolving each slice with a Gaussian kernel:

$$\hat{I}(z, x, y) = (\tilde{I} * G_\sigma)(x, y) = \sum_{u,v} \tilde{I}(z, x-u, y-v)\, G_\sigma(u, v), \tag{15}$$

where

$$G_\sigma(u, v) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{u^2 + v^2}{2\sigma^2}\right). \tag{16}$$

We use $\sigma = 1$ in our implementation. This smoothing produces visually smooth transitions across depths but suppresses high-frequency cellular structures.

D.5 Coordinate MLP (NeRF-like).

As a per-scene optimization baseline, we fit a coordinate MLP $f_\theta : (x, y, z) \mapsto I$ independently to each evaluation stack, in the spirit of NeRF [Mildenhall et al., 2020]. Since our virtual camera undergoes a pure axial translation without parallax (Section 3.2), each observed pixel directly samples the reflectance at (x, y, z) rather than integrating it along a ray through an occluding volume; we therefore omit volume rendering and regress the intensity directly. All coordinates are normalized to $[-1, 1]$, with z normalized over the depth range of the input slices, and are mapped through the positional encoding of NeRF with $L = 10$ frequency bands,

$$\gamma(p) = \big(p, \sin(2^0 \pi p), \cos(2^0 \pi p), \ldots, \sin(2^{L-1} \pi p), \cos(2^{L-1} \pi p)\big), \tag{17}$$

applied to each coordinate separately. The MLP has four hidden layers of 256 units with ReLU activations, followed by a sigmoid output. For each stack, we optimize it with an MSE loss on the pixels of the three input slices only, using Adam with a learning rate of 2×10^{-3} for 1,500 iterations of 8,192 randomly sampled pixels, and then query it at the target depths. We use one window of three inputs and four targets per evaluation stack, sampled as for CD-RCM. The MLP thus has no information beyond the three inputs of the stack it is fit to, which isolates the contribution of the cross-stack prior learned by CD-RCM.

D.6 FILM.

FILM [Reda et al., 2022] synthesizes a frame at an arbitrary time $t \in (0, 1)$ between two input frames using a shared multi-scale feature extractor, bidirectional flow estimation, and a fusion decoder. Since FILM operates on pairs of RGB frames, we replicate each grayscale slice to three channels, pad it to a multiple of 64 pixels, and condition the model on the two acquired slices that bracket the target depth, setting t to the relative depth of the target between them. We tile each evaluation stack into non-overlapping windows of five consecutive slices, use the first and last slices as inputs, and predict the three intermediate slices at $t \in \{0.25, 0.5, 0.75\}$, yielding 643 windows over the 60 evaluation stacks.

Zero-shot. We apply the pretrained checkpoint directly, without any training on RCM data.

Fine-tuned. We fine-tune all parameters of FILM on our 156 training stacks, tiled into five-slice windows in the same way (1,581 windows, *i.e.,* 4,743 training samples per epoch, as each intermediate slice is a separate sample). Instead of FILM's original combination of L_1, perceptual, and style losses, we use $L_1 + 0.5 \cdot \mathcal{L}_{\text{LPIPS}}$, matching the photometric–perceptual objective of CD-RCM. We optimize with AdamW [Loshchilov and Hutter, 2017] at a learning rate of 2×10^{-5}, using 100 linear warmup steps followed by cosine decay, and a batch size of 9 for 5 epochs.

E Evaluation Metrics

We evaluate reconstruction quality using both photometric and perceptual metrics. Specifically, we report peak signal-to-noise ratio (PSNR), structural similarity (SSIM), and the learned perceptual image patch similarity (LPIPS). We also use our proposed skin-specific perceptual feature loss $\mathcal{L}_{\text{SPF}}$ (Eq. 13) as a skin domain-aware perceptual similarity metric.

E.1 PSNR

Peak signal-to-noise ratio (PSNR) measures pixel-wise reconstruction fidelity between a predicted image $\hat{I}$ and the ground truth image I. It is defined as:

$$\text{PSNR} = 10 \log_{10} \left(\frac{\text{MAX}^2}{\text{MSE}} \right), \tag{18}$$

where MAX denotes the maximum possible pixel intensity (*i.e.,* 255 in our case), and MSE is the mean squared error:

$$\text{MSE} = \frac{1}{HW} \sum_{x,y} (I(x, y) - \hat{I}(x, y))^2. \tag{19}$$

Higher PSNR values indicate better reconstruction fidelity.

E.2 SSIM

Structural Similarity Index (SSIM) evaluates perceptual image quality by comparing luminance, contrast, and structural information between two images. It is defined as:

$$\text{SSIM}(I, \hat{I}) = \frac{(2\mu_I \mu_{\hat{I}} + C_1)(2\sigma_{I\hat{I}} + C_2)}{(\mu_I^2 + \mu_{\hat{I}}^2 + C_1)(\sigma_I^2 + \sigma_{\hat{I}}^2 + C_2)}, \tag{20}$$

where μ, σ^2, and $\sigma_{I\hat{I}}$ denote the mean, variance, and covariance of the images, respectively, and C_1, C_2 are constants used for numerical stability. SSIM ranges from -1 to 1, with higher values indicating greater structural similarity.

E.3 LPIPS

Learned Perceptual Image Patch Similarity (LPIPS) Zhang et al. [2018] measures perceptual similarity in the feature space of a pretrained convolutional network (we use the VGG backbone in our implementation, as it is shown in the original paper to be the strongest variant). Given feature activations $\phi_l(\cdot)$ at layer l, LPIPS is computed as

$$\text{LPIPS}(I, \hat{I}) = \sum_l w_l \, \|\phi_l(I) - \phi_l(\hat{I})\|_2^2, \tag{21}$$

where w_l are learned channel-wise weights. Unlike pixel-based metrics, LPIPS better captures perceptual differences in texture and structure. Lower LPIPS values indicate greater perceptual similarity.

F Broader Impacts

This work has potential positive societal impact in biomedical imaging and clinical workflows. By enabling continuous-depth reconstruction from sparse RCM stacks, CD-RCM may improve visualization, reduce acquisition burden, and support more consistent analysis of skin microstructure. In the long term, such tools could assist clinicians in diagnosis and monitoring, particularly in settings where dense sampling is time-consuming or impractical.

Potential negative impacts may arise from over-reliance on synthesized data. As CD-RCM generates intermediate slices that are not directly observed, there is a risk that inaccuracies or hallucinated structures could be misinterpreted as real tissue features, potentially affecting clinical decisions if used without appropriate validation. To mitigate these risks, CD-RCM should be used with clear communication that synthesized slices are model predictions.

Additionally, the current model is a proof-of-concept study and is trained on a limited dataset of healthy skin from a single imaging setup. To broaden generalization across populations, devices, or pathological conditions, more exhaustive, diverse data should be collected for training and evaluating the model, incorporating uncertainty estimation, and validating the method in downstream clinical tasks before deployment in real-world settings. The same is discussed in Section 4.6.