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

Tooba Imtiaz[1], Milind Rajadhyaksha[2], Kivanc Kose*,[2], Jennifer Dy*,[1]

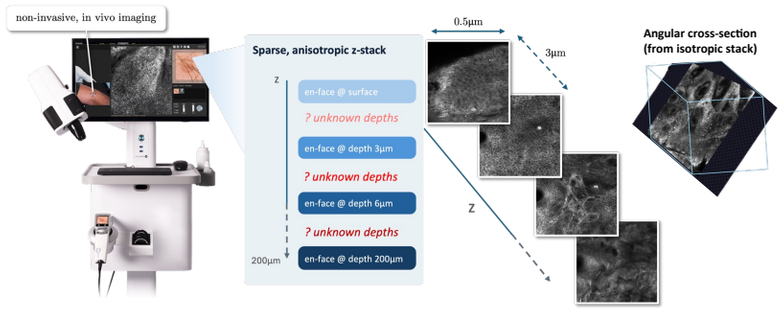
[1] Northeastern University
[2] Memorial Sloan Kettering Cancer Center
* denotes equal contribution

NEURAL INFORMATION PROCESSING SYSTEMS 2026

TL;DR The first feedforward, depth-queryable model that converts a sparse, anisotropic skin z-stack into a dense, isotropic volume — with no per-patient optimization, and sub-second inference

Problem

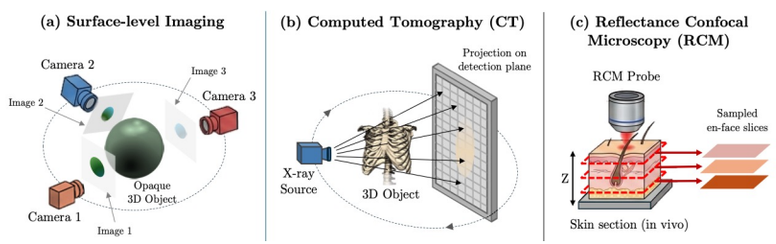
- → RCM provides noninvasive cellular-resolution *optical biopsies* as depth-stacked en-face images
- → However, its 3 µm axial resolution is 6× coarser than 0.5 µm lateral resolution
  - → Anisotropic stacks restrict clinicians to examining this information-rich 3D data through en face slices only



Isotropic stacks would enable clinicians to explore the 3D tissue volume from multiple orientations, including vertical slices that mimic histopathological sections

Comparison of Imaging Paradigms

- → Unlike surface novel view synthesis (geometric parallax), or CT (X-ray attenuation), RCM images internal tissue by axial optical sectioning with no viewpoint change — shallow sections occlude deeper ones.

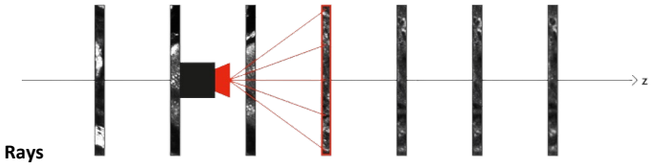


Multiple posed views with geometric parallax

Multiple angle transmission projections; 3D structure from X-ray attenuation

Axial optical sectioning; *en face* imaging of discrete depths

Modeling RCM Acquisition as Virtual Pinhole Camera



ker Rays

Each pixel gets a ray embedding, so transformers built for multi-view geometry can process the stack

Camera Pose

- Identity rotation: $\mathbf{R} = \mathbf{I}_3$
- Pure z-translation: $\mathbf{c}_i = [0, 0, z_i]^T, z_i = \underbrace{i}_{\text{slice index}} \cdot \underbrace{\Delta_z}_{\text{constant axial step-size}}$
- World-to-camera transform: $\mathbf{T}_{\text{w2c}}^{(i)} = \begin{bmatrix} \mathbf{I}_3 & \mathbf{c}_i \\ \mathbf{0}^\top & 1 \end{bmatrix}$ → **Camera pose (c2w):** $\mathbf{T}_{\text{c2w}}^{(i)} = \begin{bmatrix} \mathbf{I}_3 & -\mathbf{c}_i \\ \mathbf{0}^\top & 1 \end{bmatrix}$

Stack Canonicalization

- Mean camera center: $\bar{\mathbf{c}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{c}_i$
- Max. absolute translation: $\mathbf{c}_\infty = \max_{i \in \{1,\ldots,N\}} \|\mathbf{c}_i\|_\infty$

→ Standardized poses: $\mathbf{c}_i \leftarrow (\mathbf{c}_i - \bar{\mathbf{c}})/\mathbf{c}_\infty$

Shared normalized coordinate system preserves relative depth.

Camera intrinsics K

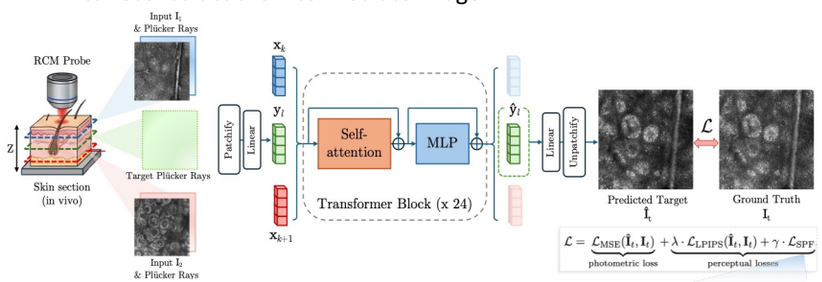
$$\mathbf{K} = \begin{bmatrix} f_x & 0 & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{bmatrix}$$

- Focal length proportional to image resolution: $f_x = f_y = \underbrace{\alpha}_{\text{constant scaling factor}} H$

and $u_0 = v_0 = \frac{H}{2}$

CD-RCM: Architecture

- → Acquire input images, compute input + target Plücker rays
- → Tokenize → feed through transformer → unpatchify target tokens to reconstruct the intermediate image

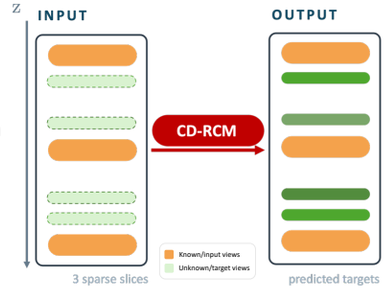


LoRA-adapted Dinov3, trained on dermatology images

- → Train with complementary photometric MSE, LPIPS, and skin-specific perceptual loss using a dermatology-adapted DINOv3 feature space

Training and Inference

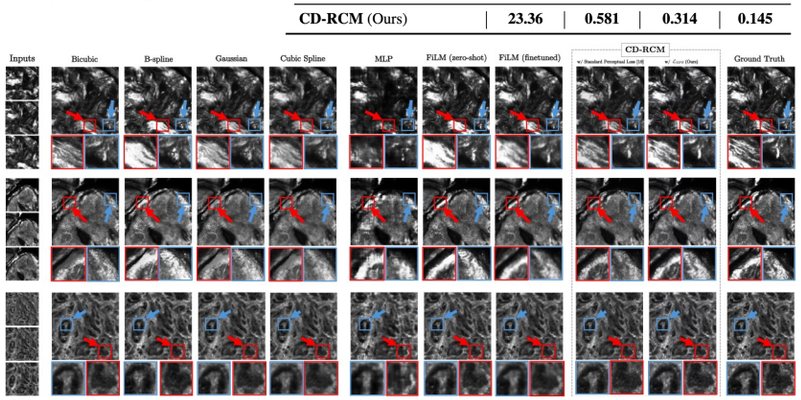
- → **Training:** Sample target slices with varying depth gaps to learn depth-generalizable interpolation
- → **Inference:** Use ***any*** three evenly spaced slices to synthesize arbitrary intermediate depths in a single feedforward pass



Results

Quantitative comparison. CD-RCM outperforms classical interpolation and learned baselines across both photometric (PSNR, SSIM) and perceptual (LPIPS, L_{SPF}) metrics

Method	PSNR ↑	SSIM ↑	LPIPS ↓	$\mathcal{L}_{SPF}$ ↓
Bilinear Interpolation	21.98	0.471	0.377	0.153
B-Spline Interpolation	22.02	0.471	0.375	0.161
Cubic Spline Interpolation	22.09	0.469	0.378	0.163
Gaussian Interpolation	22.87	0.522	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**



Qualitative comparison of NDS methods. CD-RCM best preserves fine cellular structures and depth-dependent morphology, producing predictions closest to GT.

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)

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

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

Acknowledgements: Reading study conducted by Dr. Miguel Cordova, an RCM expert at Dermatology Service of MSKCC. Work supported by grants DoD/USAMRAA HT94252410553, DoD/USAMRAA HT94252410554, NIH/NCI R01CA199673, NIH/NCI U24-CA264369, R01CA293974, and NIH/NCI Grant P30 CA008748.

Paper Website ML Lab @ NEU