DEPTH ESTIMATORS ARE IMPLICIT NEURAL FIELDS FOR 3D SCENE GEOMETRY INPAINTING AND RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

The 3D geometry of real-world scene data is often incomplete. Mainstream methods use depth estimators to inpaint missing structure. However, their prediction results can be inconsistent with observed geometry, or unreliable on out-of-distribution data. To solve these problems, we propose Neural Depth Field (NDF). Our key insight is that a depth estimator can also be a scene-level implicit field. As an estimator, it adapts to the target domain by learning observed depth data. While as an implicit field, it fits the existing geometry to maintain consistency. Under this view, NDF addresses both problems through a single test-time optimization. Experiments evaluate completion accuracy and geometric consistency across indoor scans, driving sequences, and height maps. On ScanNet Clean50, NDF reaches 97.66% accuracy at $\delta_{1.05}$ on non-prompt pixels, 0.99 percentage points above our CAPA implementation. On Birmingham, it reduces heldout height MAE by 41.8% relative to the frozen estimator. These results support fitting pretrained depth estimators as scene-conditioned fields for geometry completion.

1 INTRODUCTION

3D scene data (Dai et al., 2017a; Savva et al., 2019; Hu et al., 2022) is crucial for visual understanding Hong et al. (2023); Majumdar et al. (2024), spatial modeling Bar et al. (2025); Zheng et al. (2024), and embodied intelligence Krantz et al. (2020); Jian et al. (2025). These downstream tasks require complete and high-fidelity geometry observations, while real-world sensory data is often structurally incomplete. For instance, indoor scans suffer from missing surfaces in occluded regions (Newcombe et al., 2011; Dai et al., 2017b); drone scans exhibit geometric gaps from constrained flight perspectives(Hu et al., 2022; Yin et al., 2024); and satellite imagery contains data voids caused by cloud, shadows, or matching failures (Mari et al., 2022; Lee et al., 2026).

A prevalent approach to recovering continuous geometry is depth estimators. These models are trained on large-scale geometric data and can recover complete depth maps from partial observations(Yang et al., 2024b; Lin et al., 2025a; Yu et al., 2026). However, two main difficulties remain. First, the completed region must remain consistent with the observed part. Second, certain data (e.g., satellite imagery) may be out-of-distribution for common depth estimators.

To solve these problems, we propose Neural Depth Field (NDF), which builds upon pretrained depth estimators. Our key insight is that a depth estimator can be treated as both a one-shot predictor and a scene-level implicit depth field. As a predictor, it learns domain-specific priors for depth inpainting by optimizing on existing depth observations. While as an implicit depth field, it fits the observed region to maintain geometric consistency. Under this perspective, NDF solves the two aforementioned challenges through a single unified test-time optimization objective.

We evaluate NDF on indoor scans, driving sequences, and large-scale height maps. On ScanNet Clean50, it achieves 97.66% pixel-heldout accuracy at $\delta_{1.05}$, exceeding our CAPA implementation by 0.99 percentage points. On DrivingStereo, DDAD, and Waymo, the corresponding gains over CAPA are 0.80, 1.27, and 0.66 points. For Birmingham height completion, NDF reduces heldout MAE by 41.8% relative to the frozen estimator. A separate ScanNet spatial-heldout evaluation tests

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

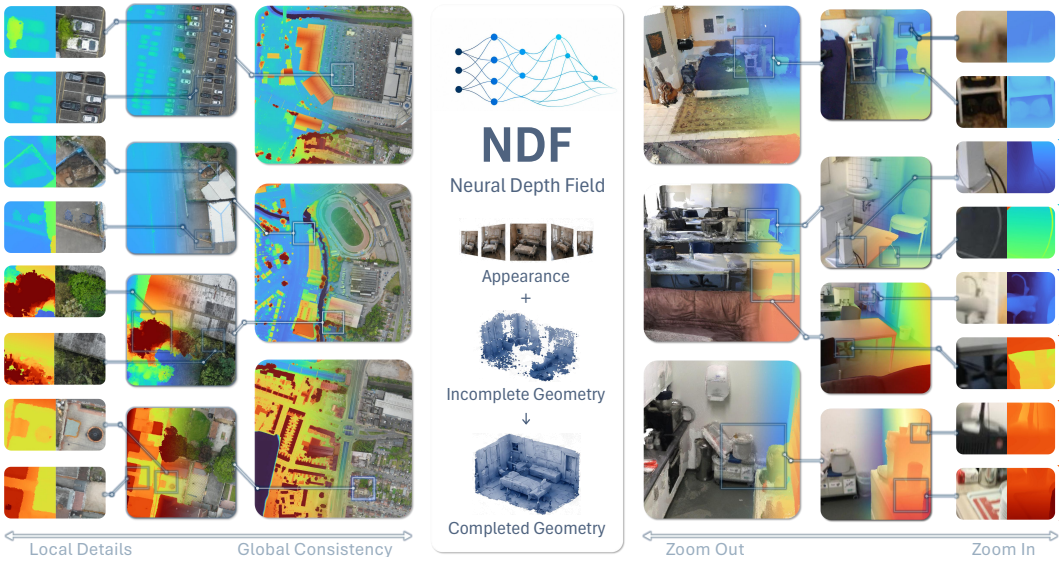


Figure 1: **Neural Depth Field (NDF)** is a test-time optimization approach designed for scene geometry inpainting task. Given the scene's appearance (e.g., multi-view RGB images) and its partial geometry (e.g., incomplete depth maps) as input, NDF inpaints missing structures with high precision and global consistency. The results preserve fine-grained details such as chimneys in satellite imagery (left) and wires in indoor scans (right). Meanwhile, the inpainted regions blend seamlessly with the observed geometry, and maintain coherence across overlapping inference windows.

prediction beyond adaptation-visible surfaces. Qualitative comparisons show finer structures and reduced cross-window inconsistencies, complementing the task-specific metrics.

In summary, our contributions are:

1. We are the first to systematically formulate depth estimators as implicit neural fields for unified 3D scene inpainting and reconstruction.

2. We demonstrate improved completion accuracy and geometric consistency across indoor scans, driving sequences, and height maps, with separate evaluations of unsupervised pixels and adaptation-unseen surfaces.

3. We provide a scalable pipeline for high-fidelity scene data inpainting and generation, together with processed scene datasets to support future research.

2 Related Work

2.1 Depth Estimation and Completion

Depth estimation recovers scene geometry from visual observation. Some methods rely only on RGB information, and predict depth directly from a monocular view (Eigen et al., 2014; Yang et al., 2024b; Hu et al., 2024; Lin et al., 2025a), or multiple view stereo (Yao et al., 2018; Wang et al., 2024; 2025). They learn visual-to-geometric priors from large-scale data (Li & Snavely, 2018; Roberts et al., 2021; Cabon et al., 2020; Yang et al., 2024a). However, RGB-only prediction does not directly determine metric scale. This can result in predicted depth scales that differ substantially from observed depth.

Other methods use prompted depth (Ma & Karaman, 2018; Park et al., 2020; Hu et al., 2021; Lin et al., 2025b; Yu et al., 2026) to condition the depth generation process. Yet the conditioning signal is often sparse, and provides weak constraints from existing depth, where completed geometry can still drift from the true metric scale.

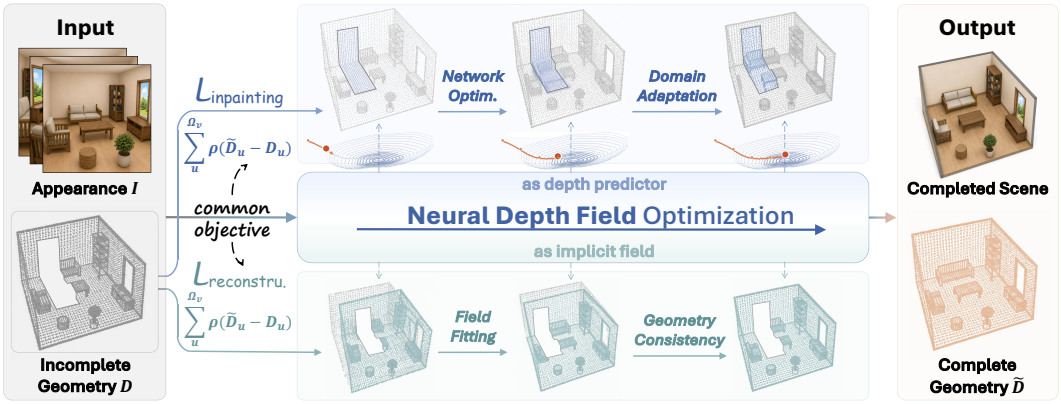


Figure 2: **Nerual Depth Field Optimization.** NDF takes as input scene appearance I and incomplete geometry D with observed set Ω_v. At each optimization step, NDF renders scene geometry, and predicts depth $\tilde{D}$ conditioned on the appearance I. As an **implicit neural field**, the rendered geometry $\tilde{D}$ should fit the observed part $D|_{\Omega_v}$. And as a **depth estimator**, the prediction $\tilde{D}$ should be supervised by the ground truth depth $D|_{\Omega_v}$ to adapt the prior distribution. Two objectives can be both described by a pixel-wise discrepancy loss. By optimizing this shared loss, NDF samples inpainting from a domain-specific prior, while maintaining consistency with the observed geometry.

NDF places observed depth directly in the optimization objective, so the pretrained geometric prior is retained while the field is forced to align with the metric scale of the current scene.

2.2 Scene Reconstruction with Neural Fields

Scene reconstruction recovers 3D structure from low-dimensional observations. Some methods directly optimize geometry from original measurements, such as multi-view images (Mildenhall et al., 2020; Kerbl et al., 2023; Mari et al., 2022; Aira et al., 2025; Xu et al., 2025b) or partial scans (Newcombe et al., 2011; Whelan et al., 2015; Dai et al., 2017b; Schöps et al., 2019). These methods have limited extrapolation ability for missing regions, where no direct measurement is available.

Recent works further introduce depth estimators as additional supervision for neural-field optimization (Deng et al., 2022; Wang et al., 2023; Yu et al., 2022; Chung et al., 2024; Xu et al., 2025a). However, this supervision is often indirect. The depth prediction and the optimized scene field remain separated and may become inconsistent.

NDF treats the depth estimator itself as the implicit geometry field. In this way, the depth prediction and scene field are inherently consistent.

2.3 3D Scene Inpainting

3D scene inpainting completes missing scene regions from incomplete observations. When inpainting geometry, most methods rely on pretrained priors for structural extrapolation (Weber et al., 2024; Liu et al., 2024; Huang et al., 2025; Dahaghin et al., 2025; Wang et al., 2026). However, inpainting targets may differ from the data distribution where the prior is learned. This gap makes direct prior extrapolation unreliable, where generated geometry can be mismatched to the current scene.

NDF fits the depth estimator on observed geometry from the target scene. This turns inpainting from direct pretrained prior extrapolation into scene-specific test-time adaptation.

3 Method

3.1 Problem Formulation

We define a scene as an underlying world S. An RGB-D image is an observation $\mathcal{O}$ sampled from this scene. Given an observation parameter c that specifies the sampling configuration, such as

position, scale, and viewing angle, the observation operator produces:

$$\mathcal{O}_c = (I_c, D_c, M_c) = \mathcal{R}(S, c). \tag{1}$$

Each observation is defined on a pixel domain $\Omega_c = \{1, \dots, H\} \times \{1, \dots, W\}$, where $I_c : \Omega_c \to \mathbb{R}^3$ is the RGB image, $D_c : \Omega_c \to \mathbb{R}$ is the incomplete scene depth, and $M_c : \Omega_c \to \{0, 1\}$ is the depth-validity mask. When c is fixed, we write $\mathcal{O} = (I, D, M)$ on Ω. Let $u \in \Omega$ denote a pixel.

$$\Omega_v = \{u \in \Omega : M(u) = 1\}, \qquad \Omega_h = \Omega \setminus \Omega_v. \tag{2}$$

Only $D|_{\Omega_v}$ is observed. For an observation, the goal of depth inpainting is to find a depth estimator f^*, whose predicted depth $\tilde{D}$ best fits the complete depth observation $\hat{D}$. Without loss of generality, we assume that f is a monocular depth estimator to simplify the derivation:

$$\tilde{D} = f(I, D, M), \qquad f^* = \arg\min_f |\hat{D} - \tilde{D}|. \tag{3}$$

3.2 Neural Depth Field

Let f_θ be a mono-depth estimator with parameters θ. For the input image I, it gives a one-shot depth prediction $f_\theta(I)$. Since I is fixed for a given observation, θ defines a scene-level implicit depth field $F_\theta : \Omega \to \mathbb{R}$, where

$$F_\theta = f_\theta(I) \implies F_\theta(u) = [f_\theta(I)](u). \tag{4}$$

Therefore, the mono-depth model has two roles. As a predictor, f_θ maps RGB to depth. As a field, F_θ represents the depth surface of the current observation. Our objective is to minimize the discrepancy between the predicted and complete depth:

$$\theta^* = \arg\min_\theta \mathbb{E}_{c\sim\mu}[\ell_c(\theta)] = \arg\min_\theta \mathbb{E}_{c\sim\mu}\left[\int_{\Omega_c} \rho\Big(F_{\theta,c}(u), \hat{D}_c(u)\Big)\, d\nu_c(u)\right], \tag{5}$$

where θ^* is the optimization target, μ is the distribution of c, ℓ_c is the loss term, $\rho : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ is a pointwise loss, and ν_c is the normalized counting measure on Ω_c. Omitting c for clarity, the loss can be decomposed into an observed-depth fitting term and a hidden-depth term:

$$\ell(\theta) = \int_\Omega M(u)\rho(F_\theta(u), D(u))\, d\nu(u) + \int_\Omega (1 - M(u))\rho\Big(F_\theta(u), \hat{D}(u)\Big)\, d\nu(u). \tag{6}$$

Since $D|_{\Omega_v} = \hat{D}|_{\Omega_v}$, the observed depth is a partial observation of the complete depth generated by the same scene. We therefore formulate the estimation of $\hat{D}$ on Ω_h as a prior-conditioned maximum a posteriori (MAP) inference problem, using the partial observation together with any available prior Π. When F_θ is used as the neural parameterization of this estimate, the practical NDF objective becomes

$$\ell_{\text{NDF}}(\theta) = \int_\Omega M(u)\rho(F_\theta(u), D(u))\, d\nu(u) + \alpha\, \Pi(F_\theta)\,, \tag{7}$$

where α is a balancing coefficient and $\Pi(F_\theta)$ denotes a prior-induced constraint, which may be a pretrained depth prior, a smoothness regularizer, or any other task-specific objective. Through the NDF parameterization, this objective uses the observed depth for reconstruction in Ω_v and the prior to infer the missing depth in Ω_h.

3.3 Proxy Geometry

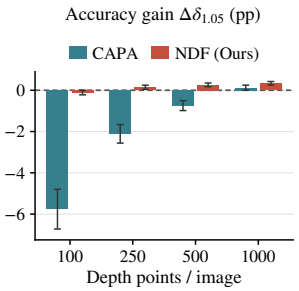
4 Experiments

The experiments examine whether fitting scene observations improves geometry prediction and provides useful constraints for reconstruction. Observation access and evaluation support are defined separately for each task; pixel-heldout completion, spatially heldout prediction and cross-view consistency are not interchangeable tests.

Table 1: Clean50 completion and ablations with 500 depth points per image. Accuracy is in %; bold marks the best score.

	Pixel-heldout			Observed		
Method	$\delta_{1.05}$ ↑	MAE (m)↓	RMSE (m)↓	$\delta_{1.05}$ ↑	MAE (m)↓	RMSE (m)↓
Frozen	97.41	0.02258	0.10770	98.66	0.01358	0.08111
CAPA	96.67	0.02452	0.10152	99.91	0.00416	0.01641
TestPromptDC	96.85	0.02416	0.10433	**99.99**	**0.00138**	**0.00515**
NDF (Ours)	**97.66**	**0.01980**	**0.08533**	99.64	0.00808	0.03412
w/o decoder prior	94.54	0.03163	0.08891	97.40	0.02175	0.05610
w/o RGB content	93.83	0.03842	0.13655	99.23	0.01214	0.04074
w/o depth prompt	49.92	0.13589	0.22915	50.78	0.13362	0.22645

Figure 3: Accuracy gain over Frozen versus depth density (paired 95% CIs).



4.1 Indoor Scene Geometry

A pretrained depth estimator can be fitted to scene observations as a conditional depth field. The useful question is not simply whether its training error decreases, but whether the fitted field predicts unsupervised geometry and supplies useful constraints to a separate reconstruction. We distinguish pixel completion from geometry unseen during adaptation because their observations and evaluation regions differ; depth-guided reconstruction is reported in Appendix A.5.

4.1.1 Does Sparse Fitting Improve Unsupervised Pixels?

Protocol. We use 50 ScanNet (Dai et al., 2017a) scans (Clean50), with 3,897 stride-20 images at 640×480. In the main comparison, every method receives the same 500 sensor-depth samples per image. We report these supplied points (*Observed*) separately from the remaining valid pixels within 10 m (*Pixel-heldout*), then macro-average per-scan metrics. Observed measures fit to the supplied depths. Adaptation and evaluation share RGB views, so pixel-heldout completion does not test unseen views or surfaces.

Methods. All methods use InfiniDepth-DepthSensor (Yu et al., 2026). NDF fits its 1.45M-parameter implicit decoder, shared across each scan. Our ports of CAPA (Ke et al., 2026) and TestPromptDC (Jeong et al., 2025) instead fit scene-shared VPT tokens and per-image RGB prompts. Calibration and fitting budgets differ; implementation details and additional baseline compatibility tests appear in Appendix A.

Result. NDF improves pixel-heldout $\delta_{1.05}$ over Frozen, CAPA and TestPromptDC by 0.25, 0.99 and 0.82 percentage points, respectively, winning on 47, 49 and 49 of the 50 scans. It also reduces MAE by 12.3% relative to Frozen. TestPromptDC fits observed points most closely, but this does not translate into better completion (Table 1). Figure 4 complements these averages: NDF preserves slender faucet stems and the continuity of a thin pole, while the prompt-adapted predictions merge or fragment these details.

Component ablations. Randomizing the decoder, replacing RGB with a constant image, or removing spatial depth prompting lowers $\delta_{1.05}$ by 3.12, 3.84 and 47.74 points, respectively, with a decrease in every scan. This supports the value of decoder pretraining and input conditioning under the fixed fitting budget. Removing the prompt retains sparse loss and scale recovery; input-intervention limits are detailed in Appendix A.2.

Observation density. With 100–1,000 nested observations per image and a common mask excluding all 1,000 points, NDF exceeds CAPA in mean $\delta_{1.05}$ at every density (Figure 3; Appendix A.3). Its gain over Frozen grows from 0.13 to 0.33 points over 250–1,000 observations, but at 100 points it is slightly lower (95.61% versus 95.72%).

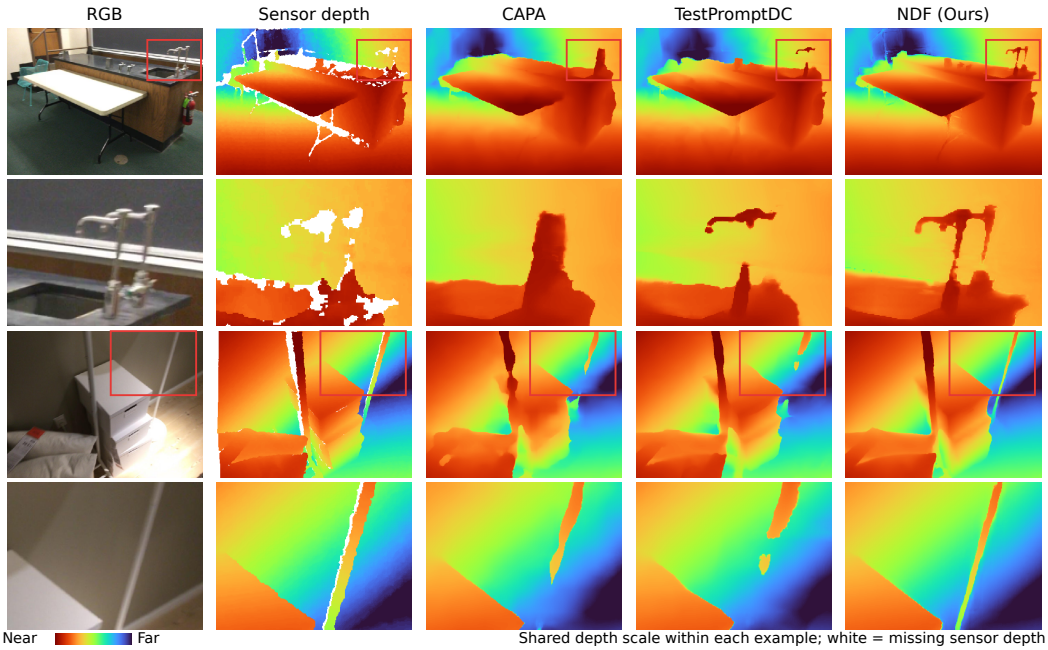


Figure 4: Qualitative comparison of depth completion on ScanNet. Full views and close-ups highlight faucets (top) and a slender pole (bottom). CAPA and TestPromptDC blur or fragment these thin structures, while NDF (ours) preserves finer boundaries and more continuous shapes. Sensor depth is shown for reference; white indicates missing measurements. Depth maps share the same color scale within each example.

4.1.2 NDF Improves Prediction of Adaptation-unseen Geometry

Separate spatial protocol. We evaluate VGGT (Wang et al., 2025) on 31 eligible ScanNet scans under fixed spatial-heldout splits, masks and method configurations. Training, guard and heldout views are disjoint. The *Heldout* region excludes train-visible mesh triangles and a 10 cm dilation on 5 cm triangle-centroid voxels, followed by connected-region selection and mask erosion. We also evaluate *Overlap*: surfaces visible in both training and heldout views, excluding sparse-depth-supervised voxels and their 10 cm neighborhood. Only Heldout tests prediction beyond adaptation-visible surfaces.

NDF adapts a shared depth head for each scene. Heldout RGB is read only for final inference, never for adaptation. Each method's scene-level affine calibration is fitted on training sparse depth and frozen. Thus the task predicts geometry unseen *during adaptation*, not geometry without a visual observation at inference.

Table 2: Geometry prediction on 31 ScanNet scans, evaluated in heldout views. Heldout excludes training-visible surfaces; Overlap is training-visible but excludes sparse-depth-supervised neighborhoods. All rows use VGGT, training-only calibrated predictions and the same evaluator. Entries are scan-macro means; bold marks the best mean within each region, not statistical significance. Acc./Comp. are directional surface distances; CD-L1 is their mean. Coverage, raw results and paired intervals appear in Appendix A.4.

		2D error ↓			3D distance (m) ↓			3D F-score (%) ↑		
Region	Method	MAE (m)	RMSE (m)	AbsRel	Acc.	Comp.	CD-L1	2 cm	5 cm	10 cm
Heldout	Frozen	0.3814	0.4398	0.1937	0.2745	0.1531	0.2138	9.05	18.73	33.46
	CAPA	0.3352	0.3835	0.1696	0.2499	0.1508	0.2003	9.20	22.35	39.21
	NDF	**0.2454**	**0.2966**	**0.1207**	**0.1828**	**0.0798**	**0.1313**	**13.04**	**29.90**	**50.00**
Overlap	Frozen	0.3653	0.4375	0.1951	0.2571	0.1982	0.2276	6.69	16.51	33.66
	CAPA	0.3292	0.3876	0.1745	0.2451	0.1956	0.2203	10.27	20.77	36.93
	NDF	**0.2258**	**0.2840**	**0.1159**	**0.1789**	**0.1071**	**0.1430**	**11.45**	**27.44**	**48.18**

Result and limits. The Heldout results in Table 2 support improved prediction beyond adaptation-visible surfaces under this fixed protocol. NDF reduces MAE by 0.1360 m relative to Frozen (paired scene-bootstrap 95% CI: 0.0745–0.2018 m) and by 0.0898 m relative to CAPA (0.0233–0.1567 m), improving 21/31 and 20/31 scans, respectively. Its CD-L1 gain over CAPA is 0.0691 m (0.0257–0.1157 m; 21/31 wins). The F@5cm mean increases by 7.55 percentage points, but its interval includes zero (-3.15–17.45). On Overlap, NDF also reduces MAE and CD-L1 relative to CAPA, by 0.1034 m and 0.0773 m.

The methods retain several implementation differences. CAPA uses layer-shared VPT, joint context of up to four views, and different resize antialiasing and calibration sampling; NDF uses single-view input.

4.2 Street-Scene Depth Completion

4.2.1 Setting

Task. Street scenes combine long depth ranges, sparse measurements, and changing viewpoints. We test whether fitting a conditional depth field to a driving sequence improves depth completion in views excluded from adaptation.

Dataset. We use DrivingStereo (Yang et al., 2019), DDAD (Guizilini et al., 2020), and Waymo (Sun et al., 2020), with 31, 40, and 40 locked fields, respectively. A field is a continuous clip or an official sequence. Each has fixed, disjoint train, guard, and heldout views, totaling 1,521 heldout images. All methods receive the same RGB images and sparse depths. Heldout RGB and depth prompts are used only for inference, with prompt pixels excluded from scoring.

Metrics. We report accuracy $\delta_{1.05}$ and $\delta_{1.25}$ in percent, together with MAE, RMSE, and AbsRel, as in the indoor evaluation. Metrics pool valid non-prompt pixels within each field and then average fields equally. The fixed GT range is $(0.001, 200]$ m.

4.2.2 Implementation

All three methods use InfiniDepth-DepthSensor (Yu et al., 2026). Frozen performs no adaptation; CAPA (Ke et al., 2026) fits sequence-shared VPT prompts through our InfiniDepth implementation; NDF fits the shared 1.45M-parameter implicit depth decoder while freezing the backbone. Adaptation reads only training views, and parameters reset only at field boundaries. The RGB–depth interface preserves native resolution and converts metric depths to model units and back. We report native metric predictions, without additional affine alignment. Optimization settings, calibration differences, and budgets are specified in Appendix B.

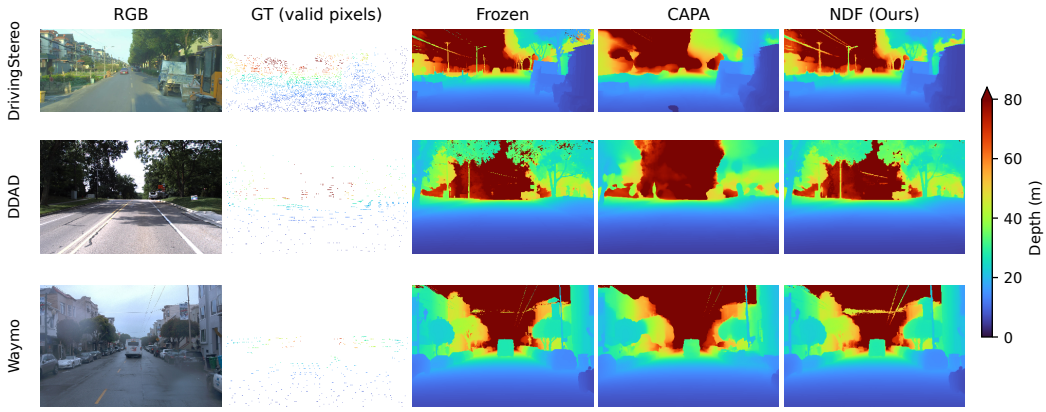


Figure 5: Qualitative street-depth completion on heldout views from DrivingStereo (top), DDAD (middle), and Waymo (bottom). NDF (ours) retains roadside structure and fine boundaries that are smoothed in the CAPA outputs. White in the GT column denotes missing reference depth. All depth panels use one color scale, saturated above 80 m for display only; evaluation retains the full 200 m GT range. These examples illustrate structural differences; selection details appear in Appendix B.

4.2.3 Quantitative Result

Table 3: Street completion with InfiniDepth-DepthSensor. Field-macro metrics exclude heldout-view depth prompts. Bold marks the best score per dataset and metric.

	DrivingStereo (31 fields)			DDAD (40 fields)			Waymo (40 fields)		
Metric	Frozen	CAPA	NDF (Ours)	Frozen	CAPA	NDF (Ours)	Frozen	CAPA	NDF (Ours)
$\delta_{1.05}$ (%) ↑	93.61	93.93	**94.73**	**87.03**	84.12	85.39	93.07	92.51	**93.17**
$\delta_{1.25}$ (%) ↑	98.14	98.83	**99.19**	95.65	94.35	**95.76**	97.52	97.47	**98.31**
MAE (m) ↓	1.8924	0.6901	**0.6534**	3.0037	2.8032	**2.2046**	0.9219	0.8040	**0.6790**
RMSE (m) ↓	22.5408	**3.1458**	4.0303	21.5934	11.6895	**8.4964**	5.3464	3.3825	**2.8526**
AbsRel ↓	0.0578	0.0226	**0.0202**	0.0636	0.0654	**0.0453**	0.0308	0.0251	**0.0209**

Table 3 shows that NDF improves both accuracy thresholds, MAE, and AbsRel over CAPA on all three datasets. The $\delta_{1.05}$ gains are 0.80, 1.27, and 0.66 percentage points on DrivingStereo, DDAD, and Waymo, with wins in 26/31, 34/40, and 33/40 fields. MAE decreases by 5.3%, 21.4%, and 15.5%, respectively. The trade-offs remain visible: Frozen has higher $\delta_{1.05}$ on DDAD, and CAPA has lower RMSE on DrivingStereo. Thus sparse fitting improves completion overall without removing every large error.

4.2.4 Qualitative Result

Figure 5 compares roadside, tree-lined road, and urban-street views. In DrivingStereo, NDF preserves utility poles and overhead wires that CAPA merges into the background, with sharper building and vegetation boundaries. Frozen also preserves many of these details; the gains from fitting are quantified in Table 3.

4.3 Satellite Height Map Inpainting

4.3.1 Setting

We complete missing heights in Birmingham's 13673×14004 canvas at 0.1 m resolution by fitting one field across RGB-height windows. We reserve $1{,}734{,}799$ valid heights for evaluation, disjoint from training and prompting. Missing RGB is completed first (Appendix C). AbsRel, MAE (m), and accuracy measure *Heldout* completion, *Observed* training fit, and *Overlap* disagreement between windows in invalid regions. Relative-depth results use scale-shift-invariant (SSI) metrics.

4.3.2 IMPLEMENTATION

We fit InfiniDepth using height-derived depth, MAE, and overlap consistency. The metric variant additionally uses sparse prompts and transferred-mask augmentation. Depth conversion, optimization, and ablations are detailed in Appendix D.

4.3.3 QUANTITATIVE RESULT

Table 4: Birmingham metric-height completion and ablations. Bold/underline mark best/second-best results; MAE is in meters.

	Heldout			Overlap			Observed		
Variant	AbsRel	MAE	$\delta_{1.025}$	AbsRel	MAE	$\delta_{1.025}$	AbsRel	MAE	$\delta_{1.025}$
Dummy Baseline	0.1756	2.0900	0.4113	0.1396	1.6752	0.4062	0.1750	2.1791	0.4087
InfiniDepth	0.0432	0.4584	0.7917	0.0199	0.2397	0.8506	0.0201	0.2701	0.8557
InfiniDepth NDF (Ours)	**0.0244**	**0.2668**	**0.8644**	**0.0083**	**0.0880**	<u>0.9393</u>	**0.0084**	**0.1204**	<u>0.9396</u>
w/o augmentation	<u>0.0270</u>	<u>0.2910</u>	<u>0.8589</u>	<u>0.0093</u>	<u>0.0974</u>	**0.9424**	<u>0.0085</u>	<u>0.120</u>	**0.9425**
w/o pretrained prior	0.0314	0.3657	0.8089	0.0129	0.1354	0.8947	0.0130	0.1905	0.8975
w/o depth estimator	0.1031	1.2384	0.5863	0.0333	0.4099	0.7518	0.0505	0.6975	0.7207

Table 5: Birmingham relative-depth completion with SSI evaluation. Bold marks the best result.

	Heldout SSI			Overlap SSI			Observed SSI		
Variant	AbsRel	MAE	$\delta_{1.1}$	AbsRel	MAE	$\delta_{1.1}$	AbsRel	MAE	$\delta_{1.1}$
Dummy Baseline	0.1778	2.3150	0.5520	0.1305	1.5301	0.5156	0.1746	2.4384	0.4811
InfiniDepth	0.1851	2.5441	0.5613	0.1496	1.7138	0.4875	0.1522	2.1700	0.5428
InfiniDepth NDF (Ours)	**0.1206**	**1.6527**	**0.6910**	**0.1077**	**1.2371**	**0.5999**	**0.1101**	**1.5434**	**0.6928**

NDF lowers heldout MAE for metric and relative depth (Tables 4–5), and metric Overlap/Observed MAE by 63.3%/55.4%. Removing augmentation, decoder pretraining, or RGB conditioning increases heldout error.

4.3.4 QUALITATIVE RESULT

Figures 6–7 in Appendix D.1 show individual and stitched windows. NDF preserves sharper structures than reconstruction alone and reduces the cross-window inconsistencies of the pretrained predictor.

5 CONCLUSION

We presented Neural Depth Field, which treats a pretrained depth estimator as a conditional field fitted to scene observations. A shared set of scene-specific parameters combines the pretrained geometric prior with measured depths. Experiments on indoor scans, driving sequences, and height-map completion show improved prediction beyond the supplied depth observations. A separate spatial-heldout evaluation supports prediction of geometry unseen during adaptation. Component ablations show the value of decoder pretraining and input conditioning under a fixed fitting budget. Gains remain sensitive to observation density and do not always transfer to downstream reconstruction. Future work will address sparse-supervision reliability and more effective depth-guided reconstruction.

AI USE STATEMENT

In this work, we used OpenAI's ChatGPT for two purposes. First, GPT-5.6-Sol was used to generate initial annotations for a dataset. All AI-generated annotations were subsequently reviewed item by item by human annotators and corrected where necessary. Second, we use the ChatGPT to polish the language of the manuscript, including its grammar, clarity, and readability. It was not used to generate or alter the research ideas, hypotheses, methodology, experimental design, results, interpretations, scientific claims, or conclusions of this work. Other tasks listed as requiring disclosure under the ICLR policy were either not performed with generative AI or were not applicable to this

work. We reviewed all AI-assisted work and take full responsibility for the final content of the paper and the accompanying dataset.

REFERENCES

Ferran Aira, Miguel Pineda, Gabriele Facciolo, Thibaud Ehret, and Tamy Boubekeur. Gaussian splatting for efficient satellite image photogrammetry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *Advances in Neural Information Processing Systems*, 2025.

Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*, 2020.

Jaeyoung Chung, Jeongtaek Oh, and Kyoung Mu Lee. Depth-regularized optimization for 3d gaussian splatting in few-shot images. *arXiv preprint arXiv:2311.13398*, 2024.

Mahtab Dahaghin, Milind G. Padalkar, Matteo Toso, and Alessio Del Bue. Splatfill: 3d scene inpainting via depth-guided gaussian splatting. *arXiv preprint arXiv:2509.07809*, 2025.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017a.

Angela Dai, Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Christian Theobalt. Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration. *ACM Transactions on Graphics*, 2017b.

Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in Neural Information Processing Systems*, 2014.

Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3D packing for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3D-LLM: Injecting the 3d world into large language models. In *Advances in Neural Information Processing Systems*, 2023.

Mu Hu, Shuling Wang, Bin Li, Shiyu Ning, Li Fan, and Xiaojin Gong. PENet: Towards precise and efficient image guided depth completion. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 2021.

Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *arXiv preprint arXiv:2404.15506*, 2024.

Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Niki Trigoni, and Andrew Markham. SensatUrban: Learning semantics from urban-scale photogrammetric point clouds. *International Journal of Computer Vision*, 130:316–343, 2022.

Sheng-Yu Huang, Zi-Ting Chou, and Yu-Chiang Frank Wang. 3DGIC: 3d gaussian inpainting with depth-guided cross-view consistency. *arXiv preprint arXiv:2502.11801*, 2025.

Chanhwi Jeong, Inhwan Bae, Jin-Hwi Park, and Hae-Gon Jeon. Test-time prompt tuning for zero-shot depth completion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9443–9454, 2025.

Yingzhao Jian, Zhongan Wang, Yi Yang, and Hehe Fan. Endowing GPT-4 with a humanoid body: Building the bridge between off-the-shelf VLMs and the physical world. *arXiv preprint arXiv:2511.00041*, 2025.

Bingxin Ke, Qunjie Zhou, Jiahui Huang, Xuanchi Ren, Tianchang Shen, Konrad Schindler, Laura Leal-Taixé, and Shengyu Huang. Depth completion as parameter-efficient test-time adaptation. *arXiv preprint arXiv:2602.14751*, 2026.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 2023.

Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Beyond the navgraph: Vision-and-language navigation in continuous environments. In *European Conference on Computer Vision*, 2020.

Jie-Ying Lee, Yi-Ruei Liu, Shr-Ruei Tsai, Wei-Cheng Chang, Chung-Ho Wu, Jiewen Chan, Zhenjun Zhao, Chieh Hubert Lin, and Yu-Lun Liu. Skyfall-gs: Synthesizing immersive 3d urban scenes from satellite imagery. *arXiv preprint arXiv:2510.15869*, 2026.

Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025a.

Haotong Lin, Sida Peng, Jingxiao Chen, Songyou Peng, Jiaming Sun, Minghuan Liu, Hujun Bao, Jiashi Feng, Xiaowei Zhou, and Bingyi Kang. Prompting depth anything for 4k resolution accurate metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025b.

Zhiheng Liu, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Jie Xiao, Kai Zhu, Nan Xue, Yu Liu, Yujun Shen, and Yang Cao. Infusion: Inpainting 3d gaussians via learning depth completion from diffusion prior. *arXiv preprint arXiv:2404.11613*, 2024.

Fangchang Ma and Sertac Karaman. Sparse-to-dense: Depth prediction from sparse depth samples and a single image. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 2018.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, Karmesh Yadav, Qiyang Li, Ben Newman, Mohit Sharma, Vincent Berges, Shiqi Zhang, Pulkit Agrawal, Yonatan Bisk, Dhruv Batra, Mrinal Kalakrishnan, Franziska Meier, Chris Paxton, Sasha Sax, and Aravind Rajeswaran. OpenEQA: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Roger Mari, Gabriele Facciolo, and Thibaud Ehret. Sat-nerf: Learning multi-view satellite photogrammetry with transient objects and shadow modeling using rpc cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2022.

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 *Proceedings of the European Conference on Computer Vision*, 2020.

Richard A. Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J. Davison, Pushmeet Kohli, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *IEEE International Symposium on Mixed and Augmented Reality*, 2011.

Hyoungseob Park, Anjali Gupta, and Alex Wong. Test-time adaptation for depth completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20519–20529, 2024.

Jinsun Park, Kyungdon Joo, Zhe Hu, Chi-Kuei Liu, and In So Kweon. Non-local spatial propagation network for depth completion. In *Proceedings of the European Conference on Computer Vision*, 2020.

Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

Thomas Schöps, Torsten Sattler, and Marc Pollefeys. BAD SLAM: Bundle adjusted direct rgb-d slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Minseok Seo, Wonjun Lee, Jaehyuk Jang, and Changick Kim. Efficient test-time optimization for depth completion via low-rank decoder adaptation. *arXiv preprint arXiv:2603.01765*, 2026.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo Open Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant CNNs. In *Proceedings of the International Conference on 3D Vision*, 2017.

Massimiliano Viola, Kevin Qu, Nando Metzger, Bingxin Ke, Alexander Becker, Konrad Schindler, and Anton Obukhov. Marigold-DC: Zero-shot monocular depth completion with guided diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Guangcong Wang, Zhaoxi Chen, Chen Change Loy, and Ziwei Liu. Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Christian Rupprecht, David Novotny, and Andrea Vedaldi. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Shaoxiang Wang, Shihong Zhang, Christen Millerdurai, Rüdiger Westermann, Didier Stricker, and Alain Pagani. Inpaint360gs: Efficient object-aware 3d inpainting via gaussian splatting for 360 degree scenes. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 117–127, 2026.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jérôme Revaud. DUSt3R: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Ethan Weber, Aleksander Holynski, Varun Jampani, Noah Snavely, and Angjoo Kanazawa. Nerfiller: Completing scenes via generative 3d inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Thomas Whelan, Stefan Leutenegger, Renato F. Salas-Moreno, Ben Glocker, and Andrew J. Davison. Elasticfusion: Dense slam without a pose graph. *Robotics: Science and Systems*, 2015.

Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. Depthsplat: Connecting gaussian splatting and depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025a.

Zhiwen Xu, Yifan Zhang, Chen Li, and Rui Wang. Skysplat: Generalizable 3d gaussian splatting from sparse multi-temporal satellite images. *arXiv preprint arXiv:2508.09479*, 2025b.

Guorun Yang, Xiao Song, Chaoqin Huang, Zhidong Deng, Jianping Shi, and Bolei Zhou. DrivingStereo: A large-scale dataset for stereo matching in autonomous driving scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024a.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *Advances in Neural Information Processing Systems*, 2024b.

Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision*, 2018.

Yihang Yin, Yusheng Xu, Jonathan Li, Wenkai Li, Jie Shen, Min Liu, Cheng Wang, and Bisheng Yang. CUS3D: A fine-grained 3d urban-scene dataset for semantic segmentation using unmanned aerial vehicle photogrammetry. *Remote Sensing*, 16(6):1079, 2024.

Hao Yu, Haotong Lin, Jiawei Wang, Jiaxin Li, Yida Wang, Xueyang Zhang, Yue Wang, Xiaowei Zhou, Ruizhen Hu, and Sida Peng. Infinidepth: Arbitrary-resolution and fine-grained depth estimation with neural implicit fields. *arXiv preprint arXiv:2601.03252*, 2026.

Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. MonoSDF: Exploring monocular geometric cues for neural implicit surface reconstruction. In *Advances in Neural Information Processing Systems*, 2022.

Wenzhao Zheng, Weiliang Chen, Yuanhui Huang, Borui Zhang, Yueqi Duan, and Jiwen Lu. OccWorld: Learning a 3d occupancy world model for autonomous driving. In *European Conference on Computer Vision*, 2024.

A Indoor Protocols and Additional Results

A.1 Clean50 Completion: Implementation and Budget

All rows use the same 50 scans, 3,897 stride-20 frames and fixed 500-point prompts. Valid sensor depth is in $(10^{-6}, 10]$ m. Excluding prompt pixels leaves 1,059,940,922 valid non-prompt pixels. Each field has equal weight after its pixel statistics are accumulated. RGB from every evaluated view is available during adaptation; this is not novel-view evaluation.

Pixel-heldout and observed support. Pixel-heldout excludes the supplied depth samples; Observed evaluates the original 500 supplied values per image, totaling 1,948,500 points. It measures agreement with available input or supervision, not generalization. Both use the same final metric predictions, with no additional fitting, scale alignment or prediction replacement. MAE, RMSE and AbsRel are pooled within each scan and then scan-macro averaged; δ_τ is the percentage satisfying $\max(\tilde{D}/D, D/\tilde{D}) < \tau$. This completion protocol has neither the legacy spatial-heldout mask nor a separate overlap split.

NDF parameter scope. Head updates eight tensors in InfiniDepth's implicit-head output MLP (1,451,329 parameters). FullFT updates the RGB feature path, conditioning modules and depth decoder (358,148,001 parameters), excluding frozen teachers and unrelated outputs. Decoder and feature-path learning rates are 3×10^{-5} and 3×10^{-6}. Both use AdamW, weight decay 10^{-4}, gradient clipping at 1, and inverse-depth Smooth-L1 with $\beta = 0.05$ in sparse-median-normalized coordinates, without a scheduler. Complete shuffled cycles give exactly E10 or E100 exposures per frame. Combinations freeze the fitted CAPA prompt before NDF fitting.

CAPA port. The implementation allocates $24 \times 1 \times 64 \times 1024$ prompt values but uses the first slice at every transformer block: 65,536 effective shared parameters, not 1,572,864 independently trained values. Tokens are inserted after CLS and before the four storage tokens and patch sequence, then removed after each block. Xavier initialization uses seed 47, not an official InfiniDepth VPT checkpoint. AdamW uses learning rate 2×10^{-4}, weight decay 0.01 and gradient clipping at 1. A 150-step linear schedule from factor 1 to 0.1 stops after 100 steps per scan. Each step accumulates single-view losses on $\lceil 0.1N \rceil$ sampled frames.

CAPA minimizes metric L1 after detached affine alignment to the 500 sparse points and applies per-image sparse affine calibration at inference. Native Frozen/NDF instead recover metric depth through their sparse inverse-depth median. This compares specified pipelines, not VPT against decoder tuning in isolation. The RGB-only row uses the sparse points only for detached inverse-depth RANSAC calibration, not as a network prompt.

TestPromptDC port. TestPromptDC is ported from its official UniDepthV2 implementation to the frozen InfiniDepth DepthSensor checkpoint. It fits an additive RGB prompt separately for each image for 50 AdamW updates, with learning rate 0.002, sparse affine alignment and L1-plus-RMSE supervision. Final inference follows the last update, retaining sparse calibration and the inherited $[0.001, 10]$ m clamp. All 3,897 frames complete 50 updates (194,850 total). This is a per-image InfiniDepth port, not an official native-model result or an equal-budget comparison to NDF.

The baseline Observed evaluation verifies all 11,691 final prediction hashes and preserves per-frame sparse-input hashes and per-field statistics. Arithmetic and invalid-support tests pass. No baseline model is rerun and no new calibration is fitted; the archived calibration differences remain. Baseline pixel-heldout values are reused from the original per-field metrics rather than recomputed with a new backend. Better Observed agreement therefore cannot by itself establish better completion.

CAPA makes 41,100 exposures (range 1–23 per frame), versus exactly 38,970 for Head-E10. Historical Head/CAPA runs used 20 L40S and 30 A6000 fields; FullFT used 25 of each. Timing includes processing overhead and is not hardware-matched. Per-field metrics and 900 provenance bindings were checked; this is not fresh reevaluation of every historical raw prediction.

NDF's paired $\delta_{1.05}$ gains over Frozen, CAPA and TestPromptDC have 95% confidence intervals of [0.17, 0.35], [0.78, 1.22] and [0.70, 0.93] percentage points. These use 10,000 paired-field bootstrap resamples: seed 20260923 for the historical Frozen/CAPA comparisons and 20260926 for the new

Table 6: Pixel-heldout results for retained Clean50 variants: accuracy in percent, MAE/RMSE in meters. All rows cover the same 50 scans. Unfavorable combinations are retained; E100 and FullFT are implementation/budget sensitivity results. Component ablations use the Head-E10 budget.

Method	$\delta_{1.05}$	$\delta_{1.10}$	$\delta_{1.15}$	$\delta_{1.25}$	MAE	RMSE	AbsRel
RGB-only + calibration	78.56	91.78	95.29	97.53	0.07846	0.18975	0.04229
DepthSensor Frozen	97.41	98.27	98.64	99.04	0.02258	0.10770	0.01201
CAPA-100	96.67	98.05	98.57	99.07	0.02452	0.10152	0.01256
TestPromptDC	96.85	98.08	98.56	99.03	0.02416	0.10433	0.01243
NDF-Head-E10	97.66	98.66	99.03	99.37	0.01980	0.08533	0.01004
NDF-Head-E100	97.66	98.65	99.03	99.38	0.01945	0.08402	0.00983
NDF-FullFT-E10	97.28	98.42	98.84	99.25	0.02178	0.09254	0.01109
NDF-FullFT-E100	97.29	98.44	98.87	99.28	0.02172	0.09057	0.01088
CAPA + Head-E10	96.65	98.03	98.56	99.07	0.02462	0.10154	0.01257
CAPA + Head-E100	96.61	98.02	98.55	99.07	0.02475	0.10139	0.01260
CAPA + FullFT-E10	96.50	98.00	98.55	99.07	0.02517	0.10182	0.01278
CAPA + FullFT-E100	96.76	98.15	98.66	99.15	0.02389	0.09737	0.01195
NDF, random decoder	94.54	97.55	98.56	99.35	0.03163	0.08891	0.01571
NDF, constant RGB	93.83	96.43	97.37	98.27	0.03842	0.13655	0.02011
NDF, no spatial prompt	49.92	75.65	87.13	95.34	0.13589	0.22915	0.07243

Table 7: Observed-point agreement on all 50 scans, from the same final metric predictions as the pixel-heldout table. Accuracy is in percent and MAE/RMSE in meters. All seven methods cover 100% of the 1,948,500 supplied points, with no invalid final predictions. Rounded 100.00 scores are not exact fits.

Method	$\delta_{1.05}$	$\delta_{1.10}$	$\delta_{1.15}$	$\delta_{1.25}$	MAE	RMSE	AbsRel
DepthSensor Frozen	98.66	98.97	99.16	99.39	0.01358	0.08111	0.00714
CAPA-100	99.91	99.95	99.96	99.98	0.00416	0.01641	0.00207
TestPromptDC	99.99	100.00	100.00	100.00	0.00138	0.00515	0.00073
NDF-Head-E10	99.64	99.79	99.84	99.90	0.00808	0.03412	0.00380
NDF, random decoder	97.40	98.99	99.46	99.79	0.02175	0.05610	0.01073
NDF, constant RGB	99.23	99.64	99.76	99.87	0.01214	0.04074	0.00602
NDF, no spatial prompt	50.78	76.30	87.48	95.46	0.13362	0.22645	0.07127

Table 8: Standalone recorded pipeline costs, including prerequisite CAPA once. E100 is cumulative and must not be added to E10. GPU-h are occupied-device elapsed proxies, not kernel-active time.

Method	Steps	Frame exposures	Recorded GPU-h	Peak GiB
RGB-only + calibration	0	0	0.603	2.04
DepthSensor Frozen	0	0	0.339	2.22
CAPA-100	5000	41100	2.670	6.05
NDF-Head-E10	38970	38970	0.471	1.69
NDF-Head-E100	389700	389700	1.809	1.69
NDF-FullFT-E10	38970	38970	6.885	10.83
NDF-FullFT-E100	389700	389700	41.354	10.83
CAPA + Head-E10	43970	80070	3.141	6.05
CAPA + Head-E100	394700	430800	4.414	6.05
CAPA + FullFT-E10	43970	80070	7.145	11.18
CAPA + FullFT-E100	394700	430800	42.756	11.18

TestPromptDC and component comparisons. They describe variation over these scans, not training-seed uncertainty. Clean50 has prior development exposure and is not a newly independent test set.

TestPromptDC audit and cost. All 15,588 saved prediction arrays pass hash, shape and finiteness checks. Zero-prompt identity with Frozen passes on the same device for 3,675 frames; 222 reuse earlier audited development/first-field results. Historical cross-device differences occur on 2,495 frames, and historical per-asset hashes are incomplete. These checks establish current implementation consistency, not bitwise reproduction of all historical predictions. The inherited warning-only deterministic setting also precludes a guarantee of bitwise training repeatability. Before clamping, non-prompt predictions contain no nonfinite or nonpositive values and 609 values above 10 m. Positive lower-bound and Observed pre-clamp counts were not fully logged; final coverage must not be read as a complete raw-clamp audit.

The port records 12.547 adaptation GPU-hours and 15.504 frame-loop GPU-hours including inference, checks and I/O, with 5.697 GiB peak allocated memory (6.108 GiB reserved). Formal wall time is 3.222 hours, using up to eight field workers across A6000, L40S and RTX4090 GPUs. Failed or migrated attempts add 4.330 summed process-hours of CPU/NFS waiting, not GPU-hours; one-off diagnostics lack complete GPU timing. These costs are not hardware-matched speed measurements. Relative to CAPA, TestPromptDC gains 0.179 points at $\delta_{1.05}$ but has higher RMSE, so neither port uniformly dominates the other.

Additional baseline compatibility. Frozen PromptDA (Lin et al., 2025b) was evaluated on all 50 scans with the same zero-filled 500-point maps, without densification. Its predictions collapse toward zero in this sparse-input setting: pixel-heldout $\delta_{1.05}$ is below 0.00001%, MAE is 1.84260 m and RMSE is 1.98021 m; Observed accuracy is 0%, MAE is 1.84139 m and RMSE is 1.98050 m. Strict loading of all 433 checkpoint tensors, meter units and preprocessing were checked; CPU replay reproduces the collapse, whereas the official example yields plausible depth. This supports an input-compatibility failure, not a general ranking of PromptDA. Final clamped coverage is 100%, but raw invalid counts are 51,304 non-prompt pixels and 83 observed points; 15,395 observed values are clamped, a potentially overlapping count. We retain this failed baseline separately from the main quality comparison.

Eff-TTO (Seo et al., 2026) cannot be directly ported with its official decoder parameterization: its LoRA wrapper targets 3×3, stride-1 convolutions, while InfiniDepth's implicit output head contains four linear layers and no matching convolution. Module inspection and positive/negative wrapper tests confirm zero eligible parameters. We report it as not applicable rather than inventing a different update module; no adaptation or quality score is produced. This does not rule out a separately designed linear-layer variant.

A.2 Component Ablations

The completed pilot uses historical development scans `scene0655_02`, `scene0153_00` and `scene0552_01`, with 27, 30 and 31 frames. On this same subset, $\delta_{1.05}$ is 97.894% for NDF, 93.740% with random decoder initialization, 94.711% without RGB content and 50.478% without spatial depth prompting. These pilot averages are not entered into the Clean50 table.

All three ablations are complete on the same 50 scans as Table 1, with frozen 500-point masks and scan-macro aggregation. Each uses the main NDF-Head configuration with exactly ten exposures per frame (38,970 steps), one decoder per scan and no intermediate reset. The 150 method-field units comprise nine audited pilot reuses and 141 new units; baseline predictions are reused rather than rerun. The historical pilot scans are not newly independent tests.

Full NDF exceeds each ablation at $\delta_{1.05}$ in all 50 scans. Its mean gains over random decoder, constant RGB and no spatial prompt are 3.12, 3.84 and 47.74 points, with paired 95% intervals [2.72, 3.56], [3.46, 4.22] and [44.56, 50.95]. The looser $\delta_{1.25}$ nearly ties random and pretrained decoders (99.35% versus 99.37%): random wins in 22 scans and the difference interval includes zero. The decoder-prior conclusion is therefore about tight-tolerance accuracy under this budget, not dominance at every threshold.

Random initialization changes only the implicit decoder, retaining the pretrained RGB backbone and depth-conditioning modules. The Clean50 extension retains the predeclared primary seed 20260926;

the other two seeds remain pilot-only. Across the three pilot seeds, mean $\delta_{1.05}$ is 93.827% (sample standard deviation 0.144 points), versus 97.894% for the pretrained decoder; $\delta_{1.25}$ remains close (99.200% versus 99.212%). This fixed-budget comparison does not establish the asymptotic performance of a randomly initialized decoder.

The RGB ablation substitutes the fixed preprocessing mean $(0.485, 0.456, 0.406)$, retaining the backbone. The prompt ablation uses the official all-false prompt-mask branch, bypassing spatial prompt feature updates. It retains all 500 sparse loss targets and their inverse median for external normalization and metric recovery. No RGB-only checkpoint is substituted. These are out-of-distribution input interventions on the same depth-conditioned model, not removal of all pretrained knowledge or all depth information.

The full-set audit checks all 23,382 raw/metric prediction arrays bytewise, alongside exact exposures, initialization and input-dependence checks. Baseline hashes match the separate 11,691-prediction audit. Both evaluated regions have 100% coverage and zero invalid final metric predictions. The inherited 10 m clamp remains: all-native-pixel macro clamp rates are 0.02234%, 0.00344% and 0.03144% for random decoder, constant RGB and no prompt; these are not invalid-prediction rates.

New work costs 0.9760, 0.9786 and 0.9849 recorded GPU-hours, respectively (2.9396 total); including reused units gives 1.0166, 1.0113 and 1.0175. Each variant peaks at about 2.22 GiB allocated memory. The 47 new fields per variant span 58.74, 56.56 and 54.90 minutes on four shared A6000 workers; these overlapping wall spans must not be summed. Costs include audit and I/O overhead. Historical baselines mix hardware, so these timings do not isolate training efficiency.

A.3 Sparse Observation Density

The density experiment retains all 50 Clean50 scans and 3,897 frames, with nested 100/250/500/1,000-point observations. A model-independent seed (20260926) subsamples the original 500 locations for the two smaller sets and adds 500 valid locations for the largest set. The original 500 locations and values remain unchanged. Input prompts, sparse losses, inverse-depth normalization and sparse-only calibration use only the current density. Each scan and density is initialized independently; NDF retains ten exact frame exposures, while CAPA retains its original 100 scene updates and frame-sampling schedule. No density-specific tuning is performed.

All twelve method–density cells use the same 1,057,992,422 valid pixels after excluding the maximum 1,000-point support. The 500-point predictions are reused and reevaluated on this common mask, not copied from the main table's averages. Final metric coverage is 100% with zero invalid predictions in every cell. Metrics are pooled within each scan, then macro-averaged across scans. Error bars use 10,000 paired scan-bootstrap samples with seed 20260926; the main chart shows accuracy differences in percentage points relative to Frozen at the same density. These intervals reflect scan variation, not repeated optimization seeds.

NDF's mean $\delta_{1.05}$ gains over CAPA at 100/250/500/1,000 points are 5.62/2.23/0.99/0.22 percentage points, with 50/50/49/43 scan wins. Against Frozen, the corresponding differences are -0.11/0.13/0.25/0.33 points. At 100 points, the paired interval $[-0.22, 0.003]$ includes zero and NDF wins 13/50 scans, although its MAE and RMSE improve. CAPA fits the 100 observed points at 99.98% accuracy but reaches 89.99% on common-heldout pixels; observed fit is not completion quality. All absolute metrics and unfavorable outcomes are retained in Table 9.

The 600 method–scan units comprise 450 new runs and 150 prediction reuses, totaling 15.3657 recorded new GPU-hours. New densities use A6000 GPUs; the 500-point predictions retain historical mixed L40S/A6000 execution. Hardware is therefore a residual confound for small density differences. Native pre-clamp values were not saved for the historical 500-point outputs, whereas new runs retain native and final predictions with clamp accounting. Per-unit prediction hashes are checked before completion; source, input and prediction-manifest hashes accompany the results. CAPA's calibration and budget still differ from NDF's. This is same-view completion on previously explored scans, not a strict 3D-unseen or equal-compute experiment.

Table 9: Sparse-depth density on Clean50, evaluated on a fixed common support excluding all 1,000 nested points. Accuracies are percentages and errors are meters; entries are scan-macro means. Bold marks the best mean at each density. The 500-point predictions are reused but reevaluated on this support, which differs from Table 1.

Points	Method	$\delta_{1.05}$	$\delta_{1.10}$	$\delta_{1.15}$	$\delta_{1.25}$	MAE	RMSE	AbsRel
100	Frozen	**95.72**	97.52	98.15	98.75	0.03071	0.12597	0.01650
100	CAPA	89.99	94.52	96.11	97.50	0.05215	0.16426	0.02789
100	NDF	95.61	**97.79**	**98.49**	**99.07**	**0.02910**	**0.10731**	**0.01502**
250	Frozen	96.84	98.00	98.45	98.92	0.02565	0.11628	0.01372
250	CAPA	94.75	97.02	97.83	98.59	0.03312	0.12380	0.01735
250	NDF	**96.97**	**98.35**	**98.82**	**99.25**	**0.02317**	**0.09535**	**0.01187**
500	Frozen	97.41	98.27	98.64	99.04	0.02258	0.10770	0.01201
500	CAPA	96.67	98.05	98.57	99.07	0.02452	0.10152	0.01256
500	NDF	**97.66**	**98.66**	**99.03**	**99.37**	**0.01980**	**0.08533**	**0.01004**
1000	Frozen	97.87	98.52	98.83	99.17	0.01994	0.09917	0.01051
1000	CAPA	97.98	98.76	99.07	99.38	0.01833	0.08272	0.00922
1000	NDF	**98.20**	**98.92**	**99.20**	**99.47**	**0.01704**	**0.07727**	**0.00858**

A.4 Spatial Heldout: Unified Rerun and Audit Limits

The existing dev12/locked38 split yields 31 eligible fields and seven geometry-based exclusions. Eligibility requires at least 20 train views, five selected heldout views, $1.5\,\mathrm{m}^2$ strict surface area and an 8% strict/heldout-visible area ratio. There are 1,540 train, 168 guard and 474 selected heldout views. Only 314 heldout views contribute valid scored pixels; two scans have three or four contributing views despite at least five selected views.

Masked camera-axis depths in meters are backprojected with dataset intrinsics and poses, fused and downsampled to 1 cm voxel centroids. Heldout mesh triangles are sampled at 10,000 points/m^2 (minimum 1,000); reference samples are not voxelized. Accuracy and completeness are directed Euclidean nearest-neighbor distances between predicted and reference points; Chamfer-L1 is half their sum. F-scores use inclusive 2/5/10 cm thresholds. No heldout-GT alignment, ICP, upper clipping of predicted depth or GT-distance outlier removal is used. All three rerun methods use the same frozen CPU evaluator, including its sampling seed and voxel convention.

Recovery and correctness. Historical split and mesh hashes matched. Historical mask hashes were unavailable, so a separately authorized check recomputed expected masks in memory from the frozen splits, meshes, archive-verified cameras and original mask-construction code. All 504 recovered masks across 31 formal and two original dev scans matched pixelwise, without modifying or replacing any mask. This is new geometric verification evidence, not recovered historical mask hashes. Two-dev-scene model smoke, 24 evaluator tests, mesh/sensor parity and camera roundtrips passed before formal execution. All 93 formal method–scan units completed, with intercepted model-execution reads reporting no guard/heldout adaptation or dense-GT access and exactly one scene reset. Prediction, mask, source and configuration hashes are retained.

Table 10: Additional unified-rerun metrics with training-only affine calibration. Steps and exposures count adaptation only, not calibration/inference. CAPA-100 and Head-E10 are different budgets.

Method	RMSE (m)	F@2cm (%)	F@5cm (%)	F@10cm (%)	Steps	Exposures
Frozen	0.4398	9.05	18.73	33.46	0	0
CAPA	0.3835	9.20	22.35	39.21	3,100	17,300
NDF	0.2966	13.04	29.90	50.00	15,400	15,400

Raw F-scores at 2, 5 and 10 cm are 1.26, 2.69 and 4.32% for Frozen; 0.48, 1.62 and 3.24% for CAPA; and 14.53, 31.47 and 50.26% for NDF. Raw Frozen/CAPA retain substantial scale error; CAPA's affine-invariant training loss does not directly fix raw metric scale. We keep training-only calibration as the main comparison rather than choosing the better prediction space per method. On

Table 11: Raw predictions evaluated separately with the identical backend and support. These uncalibrated numbers are not mixed into the calibrated main table.

	Raw 2D error ↓			Raw 3D distance (m) ↓		
Method	MAE (m)	RMSE (m)	AbsRel	Acc.	Comp.	CD-L1
Frozen	1.1172	1.1540	0.5120	0.7467	0.8610	0.8039
CAPA	1.1508	1.1825	0.5299	0.7702	0.8934	0.8318
NDF	0.2510	0.3063	0.1212	0.1890	0.0787	0.1339

Heldout, each space scores the same 21,985,645 valid-GT pixels in 314 of 474 selected views. All methods have 100% prediction coverage and zero nonfinite/nonpositive predictions on this support, in both spaces; the remaining 160 views have no eligible GT pixels, not failed predictions.

Overlap supplementary evaluation. Overlap uses the frozen Tier-A mesh triangles visible in both training and heldout views, excluding 5 cm voxels containing first-hit faces supervised by positive training sparse depth and their 10 cm dilation. It is not a strict-mask complement or a 3D-unseen region. Recomputed training visibility and sparse-supervision exclusions match the frozen arrays in all 31 formal and two original dev scans. Masks are materialized separately with the original first-hit projection and 3×3 erosion rules; these are new mask hashes, not recovered historical hashes. The masks have zero pixel intersection with Heldout. Existing whole-view raw and train-calibrated predictions are reused without new inference, calibration, clipping or GT-distance filtering; the evaluator and its 24 correctness tests are unchanged.

The same 31 scans and 474 selected views yield 3,749,336 eligible Overlap pixels in 463 contributing views, covering 44.9891 m^2 in total. The 11 zero-support views remain recorded; no scan is excluded. All methods have 100% coverage and zero invalid predictions in both prediction spaces. Effective views differ between regions, so their means are not a matched-support difficulty comparison. Against CAPA, NDF improves Overlap MAE by 0.1034 m [0.0285,0.2033] (20/31 wins) and CD-L1 by 0.0773 m [0.0270,0.1384] (22/31). Its F@5cm gain is 6.67 points [-1.82,14.85] (23/31); F@2cm also has an interval crossing zero. The result ledger retains all per-scan, raw and calibrated values with source/prediction/mask hashes. Added formal CPU evaluation takes 850.19 s wall time (1,103.01 s process time), excluding import and final hash auditing, with zero GPU-hours or repeated model cost. Historical tuning contact and the pipeline differences below still apply.

Retained implementation differences. VGGT NDF updates its 32,654,562-parameter depth head with AdamW, inverse-depth Smooth-L1 ($\beta = 0.05$), learning rate 3×10^{-5}, weight decay 10^{-4} and gradient clipping at 1. It uses one frame/update, shuffle-cycle sampling and exactly ten exposures/frame. Sequence CAPA uses VPT-16 initialization, learning rate 2×10^{-4}, 100 steps/scene and joint groups of up to four views in adaptation and inference. Its archived implementation inserts `prompt_tokens[0]` at every ViT block: 16,384 values directly affect the loss, although the allocated tensor contains 393,216 values. We retain this layer-shared variant rather than silently converting it to independent per-layer VPT. The new hash of the initializer is recorded, but no historical initializer hash was recovered.

Both pipelines resize these inputs to 518×392; Frozen/NDF explicitly disable RGB resize antialiasing, whereas CAPA retains torchvision's default enabled setting. Scene-level affine calibration uses training sparse observations only, but point sampling differs: Frozen/NDF sample per frame before fitting, while CAPA pools points before subsampling. Complete resolved configurations and every formal CAPA YAML hash were checked against the frozen source and smoke records. These are specified-pipeline comparisons, not a controlled ablation of trainable modules.

Eight scene workers complete the formal model stage in 26.3 minutes; CPU evaluation is separate. The common current runtime is Python 3.12 and PyTorch 2.11. The historical rental environment was unavailable, so the rerun is not bitwise reproduction of the old environment. Raw/calibrated predictions and adapted weights are now retained. Completed units resume by artifact hash; an interruption between runner completion and outer audit sealing requires explicit recovery rather than fabricated audit evidence.

Table 12: Unified-rerun costs on A6000 GPUs; each scan's three methods use the same device. GPU-h are unit wall-time proxies, including model loading, adaptation, calibration, export and runner bookkeeping, but excluding wrapper preflight/final artifact hashing. Elapsed spans include scheduling gaps and are not additive. Peak reserved GiB: 14.49/19.88/14.88.

Method	Occupied GPU-h	Sum unit wall (h)	Elapsed span (h)	Peak allocated GiB
Frozen	0.4403	0.4403	0.3433	13.08
CAPA	1.1305	1.1305	0.3594	18.39
NDF	1.0756	1.0756	0.3615	13.20

Paired results and persistent limits. Intervals use 5,000 paired scan-bootstrap resamples, seed 20260619, and reflect scan variation rather than repeated training seeds. Calibrated NDF improves CD-L1 over Frozen by 0.0825 m [0.0451,0.1235] (24/31 wins), and F@5cm by 11.17 percentage points [3.92,18.54] (24/31). Against CAPA, CD-L1 improves by 0.0691 m [0.0257,0.1157] (21/31), but F@5cm's 7.55-point gain has interval $[-3.15,17.45]$ (21/31). F@2cm intervals cross zero against both comparators. The result ledger includes per-scan values, medians, all intervals and unfavorable scenes.

Earlier full50 exploration included all 474 present heldout frame IDs. No heldout/guard input enters this rerun's adaptation or calibration, but the historical contact prevents a blind-selection claim. Query RGB remains available at inference. The protocol's discretized visibility exclusion also does not prove continuous-surface invisibility. Thus the result supports conditional prediction beyond adaptation-visible geometry, not visual-observation-free generation or a causal advantage of scene sharing.

The archival, non-unified tables remain in the result ledger, separate from these new tables. Relative to those old calibrated means, MAE changes by $+0.000002/-0.002477/-0.008652$ m for Frozen/CAPA/NDF; F@5cm changes by $+0.10/-0.03/+0.54$ percentage points. Old prediction binaries were unavailable, so these differences combine rerun hardware/runtime effects with backend changes and cannot be attributed solely to evaluator repair.

A.5 GS Depth Sources: Original Three and Added Five

Original fields: `scene0146_01`, `scene0423_02`, `scene0607_00`. Added fields: `scene0086_02`, `scene0609_01`, `scene0671_01`, `scene0686_02`, `scene0277_00`. Exactly five Clean50 families passed frozen completeness, camera and R/H feasibility checks after excluding historical GS development families. The requested seven additions were reduced to five rather than relaxing eligibility. The fixed selection seed is 20260922; no model performance enters selection.

All five added scans have earlier Clean50 exposure; `scene0671_01` and `scene0686_02` also have development/selection history. They are new to this GS comparison, not untouched throughout the project. All four arms cover both conditions on all eight fields: 12 audited historical units are reused and 52 new units are run. All 64 receive unified evaluation; all 12 historical numeric parity checks pass.

Frozen reconstruction protocol. Training uses 320×240 images, camera-axis depth, fixed topology, common initialization, 512 warm-up updates and three rounds of 256 updates. RGB loss is $0.8L_1 + 0.2(1 - \mathrm{SSIM})$; sparse and dense depth Smooth-L1 terms have weights 0.1 and 0.05, with metric $\beta = 0.05$. Camera order and common regularization are unchanged.

GS-only retains RGB and sparse-depth supervision without a dense teacher. Frozen and CAPA teachers remain fixed; NDF uses the prescribed E1/E2/E3 snapshots, not the Clean50 Head-E10 model.

Teacher support requires finite depth in $(0.001, 10)$ m, nearest-sparse-anchor relative residual at most 0.10, rendered opacity at least 0.01 and at least 128 valid samples. Prompt pixels are excluded. The same rule applies to all dense sources, but their resulting support may differ. No dense GT or

evaluation mask guides teacher fitting. NDF provides scene-specific E1/E2/E3 snapshots; there is no normal supervision or GS-to-NDF update.

Table 13: Eight-scan GS depth-source comparison, with R/H kept separate. Cross-view is world-point disagreement on the frozen near/medium view pairs, not mesh reconstruction accuracy. Geometry is evaluated on valid rendered support; overall depth coverage exceeds 99.99% for all rows. Reference-view appearance is not a novel-view test. Table 14 separates the original three and added five scans.

		Geometry error ↓			Reference-view appearance		
Cond.	Depth source	Hole MAE (m)	All MAE (m)	Cross-view (m)	PSNR↑	SSIM↑	LPIPS↓
R	None (GS-only)	0.03229	0.04536	0.01876	23.365	0.76396	0.31309
R	Frozen	0.03033	0.04119	0.01708	23.345	0.76379	0.31259
R	CAPA-100	0.02974	0.04120	0.01712	23.341	0.76367	0.31295
R	NDF E1/E2/E3	0.03015	0.04102	0.01708	23.344	0.76377	0.31263
H	None (GS-only)	0.07907	0.04572	0.01929	23.329	0.76364	0.31426
H	Frozen	0.05865	0.04139	0.01744	23.306	0.76344	0.31389
H	CAPA-100	0.06169	0.04144	0.01744	23.308	0.76344	0.31440
H	NDF E1/E2/E3	0.05817	0.04123	0.01741	23.308	0.76348	0.31379

Table 14: Original3/new5, separated from the combined8 results in Table 13. Geometry errors are meters; cross-view is overall world-point disagreement. Appearance uses reference cameras, not heldout novel views.

Subset	Cond.	Source	Hole MAE	All MAE	Cross-view	PSNR	SSIM	LPIPS
old3	R	None	0.03874	0.05910	0.02153	25.509	0.81177	0.23163
old3	R	Frozen	0.03683	0.05419	0.01960	25.491	0.81155	0.23136
old3	R	CAPA	0.03611	0.05446	0.01990	25.487	0.81155	0.23158
old3	R	NDF	0.03635	0.05404	0.01970	25.494	0.81163	0.23127
old3	H	None	0.08478	0.05972	0.02243	25.424	0.81131	0.23174
old3	H	Frozen	0.06692	0.05461	0.02022	25.393	0.81102	0.23140
old3	H	CAPA	0.07013	0.05470	0.02035	25.391	0.81089	0.23167
old3	H	NDF	0.06617	0.05443	0.02021	25.392	0.81098	0.23162
new5	R	None	0.02842	0.03711	0.01709	22.078	0.73527	0.36196
new5	R	Frozen	0.02642	0.03338	0.01556	22.058	0.73513	0.36133
new5	R	CAPA	0.02591	0.03325	0.01546	22.054	0.73495	0.36177
new5	R	NDF	0.02643	0.03321	0.01550	22.054	0.73506	0.36145
new5	H	None	0.07564	0.03732	0.01740	22.072	0.73504	0.36377
new5	H	Frozen	0.05368	0.03346	0.01577	22.054	0.73488	0.36337
new5	H	CAPA	0.05662	0.03348	0.01569	22.059	0.73497	0.36404
new5	H	NDF	0.05337	0.03331	0.01573	22.057	0.73498	0.36309

Table 15: Cross-view valid-support coverage (%). The result ledger also retains depth coverage by region, near/medium pair coverage and every per-field endpoint.

Subset	Cond.	None	Frozen	CAPA	NDF
old3	R	100.00	100.00	100.00	100.00
old3	H	100.00	100.00	100.00	100.00
new5	R	100.00	100.00	100.00	100.00
new5	H	100.00	100.00	100.00	100.00
combined8	R	100.00	100.00	100.00	100.00
combined8	H	100.00	100.00	100.00	100.00

For combined8 H, NDF's paired hole-MAE gains over Frozen and CAPA are 0.48 mm [0.08,0.87] and 3.51 mm [0.88,6.49], with 5/8 and 7/8 wins. For combined8 R: 0.18 mm [-0.45,0.69] and -0.41 mm [-1.16,0.10], with 6/8 and 5/8 wins. Positive means lower error. New5 H gains are

0.31 mm [-0.13,0.75] and 3.25 mm [0.72,6.26]; new5 R gains are -0.005 mm [-0.934,0.796] and -0.515 mm [-1.710,0.134]. Intervals use 10,000 paired-field bootstrap resamples, seed 20260903. These are small-sample, single-training-seed results.

New GS runs, shared warm-ups and teacher generation total 2.43 recorded GPU-hours. New training spans 34.3 minutes on eight A6000s, excluding preparation and subsequent CPU evaluation. Frozen/CAPA/new NDF teachers account for 0.186/1.008/0.427 GPU-hours. Historical reuse is costed separately, including an unused E4 generation stage in the old trajectory; E4 was not used by GS. New runs within a field share device/environment; historical RTX 4090 reuse precludes hardware-matched timing claims. Unified CPU evaluation totals 2,025 seconds of summed job wall time, with zero GPU-hours.

The versioned result ledger retains all per-field values, source hashes, regional/pair coverage, paired comparisons and non-overlapping costs. The evidence supports useful one-way dense-depth supervision, but only modest and condition-dependent NDF benefits over other depth sources.

CAPA has lower combined8 R-condition hole MAE, and NDF is essentially tied with Frozen on the new5 R subset. Reference-view appearance is also mixed, with slightly lower PSNR than GS-only. Most of the gain comes from adding dense-depth supervision; these comparisons do not establish uniform superiority over other depth sources or bidirectional improvement between depth fitting and reconstruction.

B Street Completion Details and Historical Results

B.1 Protocol and Implementation

The main Street table uses the frozen query-consistent InfiniDepth protocol. The 111 locked fields contain 5,378 training, 708 guard, and 1,521 heldout frames. DrivingStereo fields are clips, so field-level development/locked separation does not imply independent source drives. Development fields number 8, 10, and 10 for DrivingStereo, DDAD, and Waymo. View roles and evaluation masks remain fixed throughout fitting and evaluation.

RGB and prompt maps are padded on the right and bottom to multiples of 16; predictions are queried at native pixel coordinates and unpadded for scoring. The Street adapter uses 4 m per model-depth unit. Native predictions are converted back to meters without an extra affine fit. Evaluation keeps all valid non-prompt GT pixels in $(0.001, 200]$ m; predictions are not clipped to 200 m or filtered by their GT error. Nonfinite or nonpositive scored predictions cause evaluation failure rather than silently reducing support.

NDF. The main variant updates only the 1,451,329-parameter implicit decoder with one training frame per step and exactly ten exposures per frame. It uses inverse-depth Smooth-L1 supervision ($\beta = 0.05$ in the model target space), AdamW with learning rate 3×10^{-5}, weight decay 10^{-4}, and gradient clipping at 1. The backbone remains frozen. A development-only learning-rate comparison among 10^{-5}, 3×10^{-5}, and 5×10^{-5} selected 3×10^{-5} under the frozen tie rule, matching the ordinary NDF setting; these are not two independent methods. The main table uses the short-budget head variant consistently with the indoor comparison, rather than selecting a different variant for each locked dataset.

CAPA. Our InfiniDepth implementation injects VPT into the RGB backbone and shares one prompt across the sequence. Each of 100 updates samples $\lceil 0.1N \rceil$ of the N training frames and averages their losses. It uses sparse-aligned metric L1 supervision and AdamW with learning rate 2×10^{-4} and weight decay 0.01. This differs from NDF's decoder objective and exact per-frame exposure budget. The main table evaluates the native output of both methods; the separate sparse-affine-aligned evaluation is not mixed into it. In particular, alignment inside CAPA's training loss is not an evaluation-time fit to heldout GT. Both methods reset only at field boundaries, and held-out/guard views never enter adaptation.

Scope of the comparison. Heldout images still supply RGB and the permitted sparse depth prompt at inference. The result therefore concerns cross-view depth completion, not unobserved-surface synthesis or equal-compute attribution. The saved cross-view diagnostics use disparity errors

for DrivingStereo and temporal depth-change errors for DDAD/Waymo; these are distinct from the mesh-based strict 3D metrics in the indoor experiment. All raw predictions, source hashes, per-field results, and budget records are retained.

Qualitative examples. The DDAD and Waymo rows of Figure 5 use the first lexicographically ordered locked field and its middle heldout frame. The DrivingStereo row is an illustrative example selected after inspecting 12 frames from four locked fields shortlisted using existing accuracy gains; it highlights utility poles and overhead wires. It is not a random sample, and this qualitative selection does not change the full locked-set results. The plots use the saved native predictions, with one shared 0–80 m display scale. Saturation affects visualization only. GT is shown only at its valid non-prompt pixels, without densification or filling missing measurements.

B.2 Historical KITTI Comparisons

These retained experiments use a different backbone, sparse-input protocol, and aggregation from the main Street table. They are not pooled with the new three-dataset results. The full comparison uses the official 1,000-image KITTI Depth Completion validation selection (Uhrig et al., 2017), with valid GT in $(0.001, 80]$ m. Exported predictions share a pixel-weighted evaluator. Marigold-DC and VGGT use the deterministic sparse-depth filtering recipe of Marigold-DC; Eff-TTO uses its released KITTI configuration, while ProxyTTA uses raw KITTI sparse input and internal outlier removal.

Table 16: Historical KITTI validation comparison, preserved separately. Eff-TTO (Seo et al., 2026), Marigold-DC (Viola et al., 2025), and NDF use 40, 50, and 100 steps, respectively. †ProxyTTA (Park et al., 2024) is a single stochastic online trial using its official image-aggregated evaluator; it is excluded from best-value highlighting because no exported predictions are available for common re-evaluation.

Base estimator	Method	MAE (m) ↓	RMSE (m) ↓	AbsRel ↓
Depth Anything V2	Frozen + alignment	1.6833	3.3755	0.0983
Depth Anything V2	Eff-TTO	0.4187	**1.4223**	0.0231
Marigold-DC	Diffusion completion	0.5555	1.6889	0.0316
VGGT	Frozen + alignment	1.4778	3.8661	0.0683
VGGT	NDF	**0.3347**	1.4590	**0.0156**
MSG-CHN	ProxyTTA†	0.3268	1.2279	–

Table 17: Historical eight-image KITTI optimization diagnostic. CAPA uses its 100-step, 12,000-point setting. This subset was used for development and is not an additional independent test benchmark.

Method	Steps	MAE (m) ↓	RMSE (m) ↓	AbsRel ↓
VGGT Frozen	0	1.4560	3.5865	0.0622
VGGT NDF	10	0.7538	2.0774	0.0340
VGGT NDF	50	0.3875	1.4587	0.0169
VGGT NDF	100	**0.3471**	**1.4286**	**0.0151**
CAPA VGGT-VPT	100	0.4076	1.4886	0.0175

C Birmingham RGB Inpainting

C.1 Setting

The Birmingham satellite scene contains missing regions in both RGB and height. Before height-map inpainting, we complete the RGB canvas so that the depth estimator receives a dense visual observation. The RGB inpainting stage is designed for native-resolution city-scale completion, where local image quality and global writeback consistency must both be preserved.

We evaluate RGB completion across four representative region types: land interior holes, river continuity regions, sea or coastline regions, and estuary or river-mouth regions. These categories cover

the dominant failure modes in the Birmingham scene: missing urban texture, discontinuous waterways, coastline drift, and large-region color or structure mismatch.

C.2 Inpainting Agent

We use a typed inpainting agent built on Nano Banana 2 as the image generation backbone and GPT-5.5 as the reviewer. The target region is first covered by a DFS-style overlapping-window plan. Each exported window is assigned a semantic type, and each type is dispatched to its own prompt template and reviewer configuration. This type-aware design lets the pipeline use different instructions for land, river, sea, coastline, and estuary regions.

The reviewer screens generated candidates before writeback. It rejects outputs with mask leakage, remaining unfilled pixels, broken shoreline or river continuity, excessive color drift, or edits that alter valid context outside the target mask. The reviewer configuration is tuned through few-shot self-iteration on a small manually labeled set of success and failure cases.

C.3 Implementation Details

The Birmingham RGB canvas has native resolution 13673×14004, corresponding to approximately 0.1 m ground sampling distance. For the formal Stage-2 RGB completion run, the DFS planner produces 588 typed windows. After reuse and no-op pruning, 519 windows require actual generation.

Across the formal mainline runs, Nano Banana 2 evaluates 968 candidate images for the accepted 519-window solution, with 963 realized upstream generation attempts and 81 upstream failures. The accumulated active generation time from recorded per-attempt latencies is about 16.6 hours, with an average of 61.9 seconds per candidate. The full mainline campaign spans about 122.7 wall-clock hours from the first formal run to the last archived center-only pass.

The GPT-5.5 reviewer is invoked 882 times in the formal pipeline, keeping 521 candidates and rejecting 361. The production reviewer logs do not store exact API usage, so we estimate the reviewer cost conservatively at roughly one million tokens in total.

C.4 Model Selection

We compare four upstream image-editing backbones for typed Stage-2 RGB completion: Nano Banana, GPT Image 2, Nano Banana Pro, and Nano Banana 2. The comparison uses a fixed 16-window benchmark covering land, sea, and estuary cases.

GPT Image 2 consistently introduces global color drift relative to the surrounding valid context. Nano Banana often preserves local texture well, but is less reliable in geometric registration and complete mask removal. Nano Banana Pro improves some local completions but still exhibits unstable writeback behavior in difficult boundary cases. Nano Banana 2 gives the best overall balance among mask compliance, structural continuity, and color consistency, and is therefore used as the default RGB backend.

C.5 Reviewer Filtering

Reviewer filtering mainly removes three classes of failures. First, it rejects generations that modify regions outside the prescribed target mask. Second, it removes candidates with residual masked pixels or incomplete filling. Third, it catches water-boundary failures, including river tone drift, shoreline discontinuity, and mask-shaped artificial coastlines.

This filtering is important because many incorrect generations are locally plausible when viewed as isolated crops but become visible artifacts after global writeback. The reviewer therefore acts as a consistency gate between local image generation and city-scale canvas composition.

C.6 Overlap and Writeback

We use overlapping sliding windows rather than hard non-overlapping tiling. In a hard 2×2 tiling baseline, the same 2048×2048 crop is split into four independent 1024×1024 tiles, each completed

and stitched directly. Even when the four local outputs are individually plausible, direct stitching introduces visible seams along internal tile boundaries, especially when the invalid region crosses the middle horizontal or vertical split.

The overlapping planner gives adjacent windows shared context and writes back only the more reliable center region. This substantially reduces seam artifacts and makes the procedure practical for native-resolution city-scale completion.

C.7 Ablation

The RGB ablations focus on typed prompting and reviewer-based filtering. We compare variants without tile typing, with prompt typing only, with reviewer typing only, and with the full typed RGB pipeline. Prompt typing mainly improves local semantic plausibility, while reviewer typing rejects candidates that remain locally plausible but break global consistency after writeback. The full typed agent is necessary to simultaneously preserve region-specific appearance and suppress large-canvas seam artifacts.

D Birmingham Height Inpainting Details

The canvas contains $177{,}875{,}345$ valid and $13{,}680{,}747$ invalid height pixels. The valid subset is split into $176{,}140{,}546$ training/prompt pixels and $1{,}734{,}799$ evaluation-only pixels. Observations are local RGB-height windows; the relative-depth InfiniDepth variant takes 672×512 RGB input, and the metric variant additionally takes depth prompts. Heights are converted using $z = H_{\mathrm{cam}} - h + 5$, with H_{cam} the camera height.

Each training window uses 25–3,000 prompt-depth points and 100,000 dense query points. We train for 10 epochs at batch size 16, taking about one hour on one NVIDIA RTX PRO 6000 Blackwell GPU. The learning rates are 2×10^{-5} for the DINOv3 ViT-L/16 backbone and 10^{-4} for the remaining parameters. The objective combines MAE with an overlap-window loss. Transferred-mask augmentation copies invalid masks from other windows and excludes their pixels from depth prompting.

The dummy baseline uses Navier–Stokes inpainting. The augmentation ablation removes both mask transfer and overlap loss; the prior ablation randomly initializes the depth head; the depth-estimator ablation removes RGB input. Metric evaluation reports height MAE in meters, AbsRel, and δ_{1+x}, the fraction of evaluated pixels whose relative depth error is within x. Relative-depth evaluation uses the corresponding scale-shift-invariant metrics. Heldout evaluates reserved heights, Observed evaluates known training heights, and Overlap compares predictions from overlapping windows in invalid regions.

D.1 Qualitative Results

Figures 6 and 7 compare individual windows and their stitched predictions. The former highlights local structures; the latter shows agreement across overlapping windows.

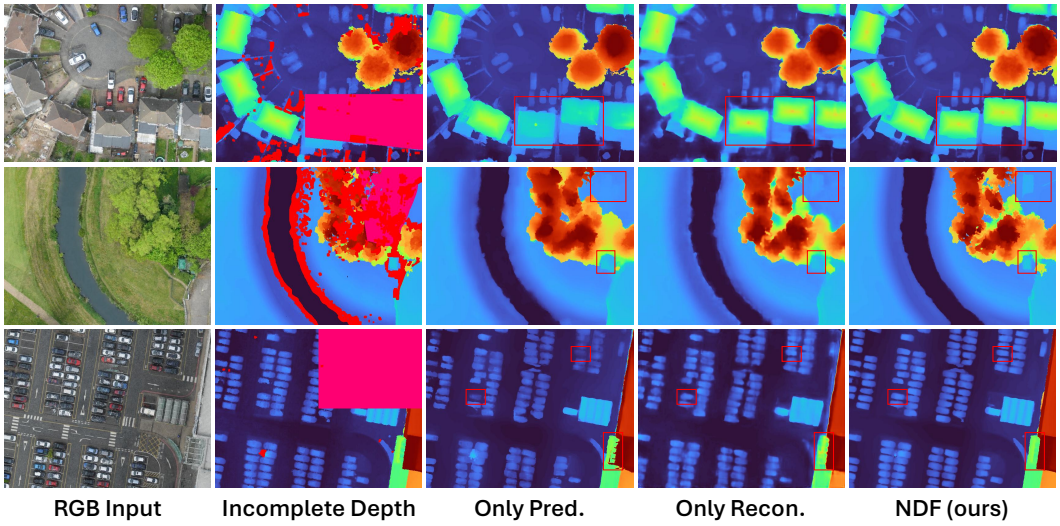


Figure 6: Birmingham window completion. *Only Pred./Recon.*: pretrained prediction/reconstruction alone. Pink/red indicate invalid/heldout pixels. NDF preserves finer structures.

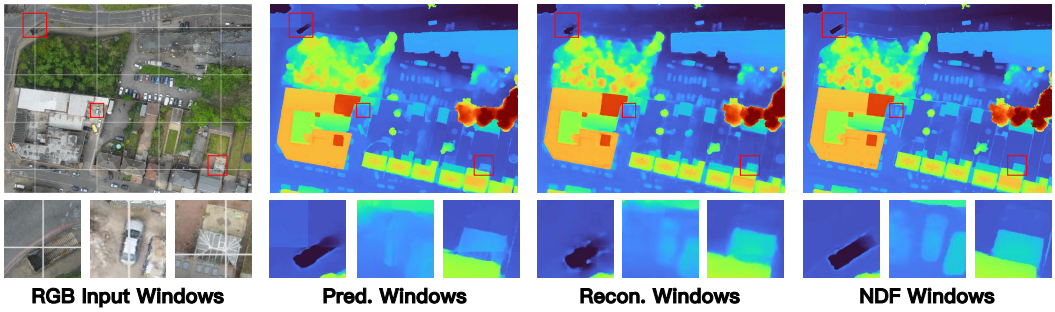


Figure 7: Stitched windows. NDF reduces cross-window inconsistencies while retaining sharper geometry than reconstruction alone.