ONCE: PATCH-FREE HIGH-RESOLUTION DEPTH REFINEMENT VIA ONE-PASS CALIBRATION AND ENHANCEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent high-resolution metric depth refinement methods often recover fine geometry through repeated patch-wise inference and subsequent fusion. However, recovering spatial detail must also account for systematic metric errors in the coarse prediction. We present ONCE (One-pass Network for Calibration and Enhancement), a patch-free depth booster that refines the output of a frozen monocular metric-depth backbone at 4K resolution in a single full-frame pass. ONCE combines a learned Corrector for image-level metric recalibration with an RGB-guided Refiner for spatial detail recovery. The Corrector predicts scale and shift directly from coarse-depth features and applies them to the coarse metric prediction. The Refiner fuses RGB and depth features through a dual-encoder architecture and predicts bounded log-relative residuals. These residuals provide spatially varying updates whose absolute magnitude scales with the calibrated depth, allowing corrections to adapt to different depth ranges. Both components are jointly optimized through the final depth prediction, without patch extraction, repeated overlapping inference, or prediction aggregation. On UnrealStereo4K with ZoeDepth, ONCE_{L} achieves $\delta_1 = 99.052\%$, RMSE $= 0.784$, and SEE $= 0.608$, outperforming all compared ZoeDepth-based refiners across the five reported depth metrics. Compared with PRV2_{C}, it reduces RMSE by 11.3% and SEE by 22.7% while using 21.5% fewer refinement parameters. Ablations demonstrate that learned calibration improves final metric accuracy and that log-relative refinement outperforms additive and linear-relative alternatives under the tested settings. Results with DAv2 further demonstrate applicability to another coarse backbone, while Cityscapes experiments show improvements after target-domain fine-tuning. Together, these results demonstrate that a single full-frame correction can surpass repeated patch-wise refinement in both metric accuracy and boundary recovery with a smaller refinement network.

1 INTRODUCTION

Monocular metric depth estimation supports 3D reconstruction, autonomous driving, and robotics. Recent models improve prediction across scenes and cameras through large-scale training and camera-aware modeling (Bhat et al., 2023; Yin et al., 2023; Hu et al., 2024; Piccinelli et al., 2024). However, their operating resolution often remains below that of modern cameras. Resizing high-resolution images preserves much of the scene layout but can obscure thin structures such as wires and fence bars and blur fine object boundaries. Recovering these details while maintaining metric accuracy remains a central challenge in high-resolution depth estimation.

Depth boosting addresses this resolution gap using additional image guidance or learned priors. BoostingDepth and PatchFusion combine predictions across resolutions (Miangoleh et al., 2021; Li et al., 2024b). PatchRefiner applies a shared refiner to multiple RGB patches, predicting additive residuals from local image evidence and coarse geometric context (Li et al., 2024c). PatchRefiner V2 (PRV2) improves efficiency through lightweight encoders and coarse-to-fine guidance, while retaining patch-wise refinement (Li et al., 2026). Full-image methods also exist: SharpDepth uses diffusion distillation to recover detail while maintaining consistency with the coarse metric esti-

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.

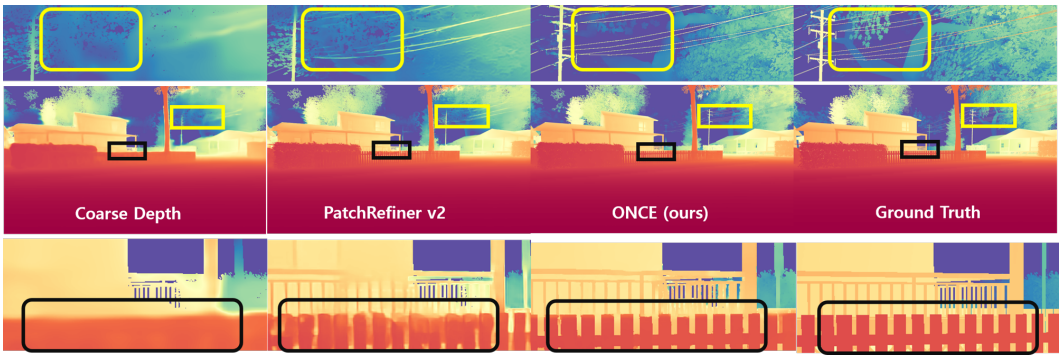

Figure 1: **Metric-depth refinement with ONCE.** From left to right: coarse depth, PatchRefiner V2, ONCE, and ground truth. The yellow regions highlight background depth discrepancies that remain after patch-wise refinement and are reduced in the ONCE prediction. The black regions show clearer separation of fence bars and gaps.

mate (Pham et al., 2025). These approaches raise a practical question: *can a single full-frame correction outperform repeated patch-wise refinement?*

We answer this question with **ONCE** (**O**ne-pass **N**etwork for **C**alibration and **E**nhancement). ONCE refines the output of a frozen metric-depth backbone at 4K resolution in one full-frame pass. It consists of a *Corrector*, which predicts an image-level scale and shift to recalibrate the coarse metric depth, and a *Refiner*, which predicts an RGB-guided bounded log-relative residual. Together, they adjust the metric reference and recover spatial detail without patch extraction, repeated local refinement, or prediction aggregation.

A coarse prediction may capture the scene layout while systematically overestimating or underestimating depth. Sharper boundaries alone do not resolve these metric errors. The Corrector therefore learns an affine adjustment directly from coarse-depth features and applies it to the coarse prediction itself. The Refiner addresses the remaining spatial correction in log-depth space. This choice accounts for the different meaning of absolute errors across depths: a 1 m error is a 100% relative error at 1 m, but only 10% at 10 m. Its multiplicative update has a bounded relative range, while the absolute correction scales with depth. Figure 1 illustrates the two aspects of the task. In the yellow regions, PRV2 recovers additional structure but retains background depth discrepancies visible in the coarse prediction. ONCE produces background depth closer to ground truth. In the black regions, ONCE also recovers clearer fence boundaries and intervening gaps. These observations motivate correcting the metric reference alongside spatial detail recovery. Calibration and residual ablations separately evaluate the corresponding design choices.

On UnrealStereo4K with ZoeDepth, ONCE_{L} achieves $\delta_1 = 99.052\%$ and $\text{SEE} = 0.608$, compared with 98.863% and 0.787 for PRV2_{C}, using 21.5% fewer refinement parameters. ONCE_{M} also improves accuracy over PRV2_{E} with 25.7% fewer parameters. Experiments with a shared DAv2 coarse checkpoint further evaluate ONCE with another backbone. Cityscapes results after target-domain fine-tuning are provided in Appendix A.6.

Our contributions are threefold: (i) a learned Corrector that predicts image-level scale and shift adjustments directly from coarse metric-depth features; (ii) ONCE, a single-pass full-frame refinement network that outperforms PRV2 on UnrealStereo4K with fewer refinement parameters; and (iii) an RGB-guided Refiner with bounded log-relative updates that scale with depth and outperform additive and linear-relative alternatives in our ablation study.

2 Related Work

2.1 Monocular and Metric Depth Estimation

Monocular depth estimation has advanced through multi-scale prediction, depth discretization, and dense prediction architectures (Eigen et al., 2014; Fu et al., 2018; Bhat et al., 2021; Ranftl et al., 2021; Yuan et al., 2022; Piccinelli et al., 2023). Large-scale training supports transferable relative-

depth representations (Ranftl et al., 2022; Yang et al., 2024a;b), while metric adaptation and camera-aware modeling address absolute depth (Bhat et al., 2023; Yin et al., 2023; Hu et al., 2024; Piccinelli et al., 2024). Generative priors and multi-scale architectures further improve spatial detail (Ke et al., 2024; Bochkovskiy et al., 2025). Related approaches include RGB-guided depth super-resolution (Hui et al., 2016; de Lutio et al., 2022; Metzger et al., 2023; Zhong et al., 2023), continuous reconstruction (Chen et al., 2021; Tosi et al., 2021; Yu et al., 2026), and sparse-depth-guided prediction (Lin et al., 2025). ONCE refines a frozen monocular model's metric-depth prediction using the original RGB image.

2.2 Multi-Resolution Depth Fusion

Depth fusion combines estimates with complementary geometry and spatial detail through content-adaptive merging, gradient-based composition, or Poisson fusion with self-distillation (Miangoleh et al., 2021; Dai et al., 2023; Li et al., 2024a). PatchFusion combines coarse and fine predictions through global-to-local guidance and consistency-aware fusion (Li et al., 2024b). These methods primarily combine multiple depth estimates. ONCE instead learns metric recalibration and dense relative correction from a single coarse depth map and the RGB image.

2.3 Depth-Conditioned Refinement

PatchRefiner applies a shared refiner to RGB patches, predicting additive residuals with frozen coarse-model features and detail-and-scale supervision (Li et al., 2024c). PRV2 reduces refinement cost through lightweight encoders and coarse-to-fine feature guidance, while retaining patch-wise residual prediction (Li et al., 2026). Other methods further improve patch processing and cross-region consistency (Kwon & Kim, 2025; Cui et al., 2026). Full-image alternatives use diffusion priors (Zhang et al., 2024; Pham et al., 2025) or gated soft-boundary correction (Zhang et al., 2026). SharpDepth preserves metric consistency through reconstruction supervision and alignment to the coarse estimate (Pham et al., 2025).

Affine correction has also been studied in relative pose estimation, where two-view correspondences constrain depth correction and camera pose (Yu et al., 2025). ONCE's Corrector instead predicts image-level scale and shift directly from coarse-depth features, without cross-view correspondences at inference. It adjusts the coarse metric reference itself, while the Refiner learns a bounded log-relative update whose absolute magnitude scales with depth. Both components operate in one full-frame pass, in contrast to the patch-wise additive refinement of PatchRefiner and PatchRefinerV2.

3 Method

ONCE learns to correct the coarse metric reference and recover spatial detail within one full-frame network. Figure 2 contrasts this formulation with PatchRefiner-style additive refinement. The *Corrector* predicts an image-level affine adjustment, and the *Refiner* predicts a dense log-relative correction. Their outputs are combined to obtain the final metric depth.

3.1 Overall Architecture

Given a high-resolution RGB image $I \in \mathbb{R}^{3\times H\times W}$, a monocular backbone predicts metric depth at its native input resolution. We resize this prediction to obtain $D_0 \in \mathbb{R}^{H\times W}$. The backbone is first adapted to the target dataset and then frozen. Its predictions are precomputed during ONCE training. At inference, the complete pipeline consists of one backbone evaluation followed by one ONCE refinement pass.

Figure 3 shows the architecture. Separate encoders extract RGB and depth features. The Corrector uses the globally pooled depth bottleneck to predict one scale and shift per image, producing calibrated depth D_{cal}. The Refiner combines spatial depth features with multi-scale RGB features to predict a dense residual. This residual is bounded in log-depth space and applied multiplicatively to D_{cal}.

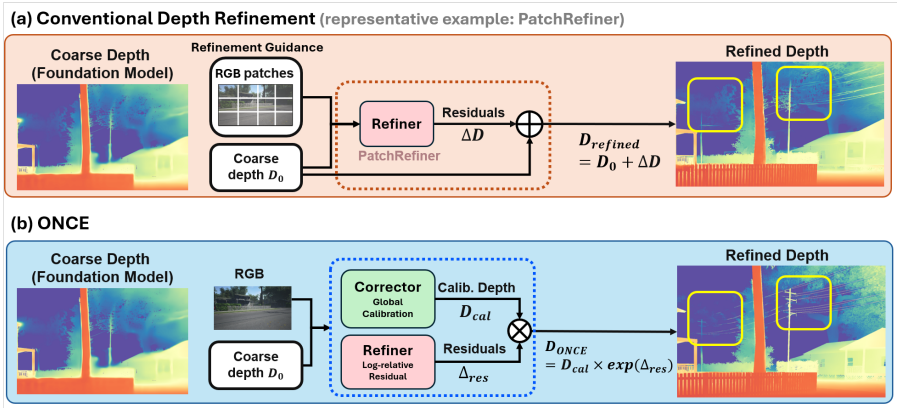

Figure 2: **Patch-wise refinement and ONCE.** (a) PatchRefiner represents a coarse-guided approach that applies a shared additive refiner to RGB patches. (b) ONCE combines a Corrector for image-level metric recalibration with a Refiner for dense log-relative updates in one full-frame pass. The calibrated depth is multiplied by $\exp(\Delta_{\text{res}})$. Feature pathways, crop indexing, patch aggregation, and final clipping are omitted for clarity.

The two branches share depth features and are trained jointly from the final prediction. ONCE uses the coarse depth map without accessing intermediate features of the frozen backbone. All refinement is performed over the full image, without patch extraction or aggregation.

3.2 Corrector: Learned Metric Recalibration

Metric-depth predictions may retain systematic scale and shift errors even when the scene layout is accurate. The Corrector models this error component through an image-dependent affine transformation before clipping. Its parameters are shared across all pixels, providing a common metric adjustment before spatially varying refinement.

For the five-level configuration, let d_5 denote the depth encoder's bottleneck feature. A lightweight head operates on its globally pooled representation:

$$(g_s, g_b) = \text{Conv}_1\left(\text{ReLU}\left(\text{Conv}_1\left(\text{GAP}(d_5)\right)\right)\right), \tag{1}$$

where Conv_1 denotes a 1×1 convolution. The outputs parameterize a bounded affine correction:

$$\begin{aligned} s &= \exp\left(s_{\max} \tanh(g_s)\right), \\ b &= b_{\max} \tanh(g_b), \\ D_{\text{cal}} &= \text{clip}\left(sD_0 + b,\, d_{\min}, d_{\max}\right). \end{aligned} \tag{2}$$

We use $s_{\max} = 0.4$, $b_{\max} = 5.0\,\text{m}$, and $[d_{\min}, d_{\max}] = [10^{-3}, 80]\,\text{m}$. The clipping operation clamps depth to this interval.

The exponential keeps s positive, preserving depth ordering before clipping. The bounds limit the magnitude of the global adjustment, and $g_s = g_b = 0$ gives the identity transformation. The Corrector is trained through the final depth loss, without separate supervision for scale or shift.

The key distinction is that the coarse metric reference itself is adjusted by a learned image-wide transformation. The calibration ablation in Table 3 shows that learning both scale and shift achieves the highest δ_1 and lowest RMSE, outperforming no calibration and either parameter alone on these metrics.

3.3 Refiner: Bounded Log-Relative Correction

The Refiner predicts a dense residual r_{res} from RGB and depth features. RGB features supply boundary and appearance cues, while depth features provide geometric context. The residual is

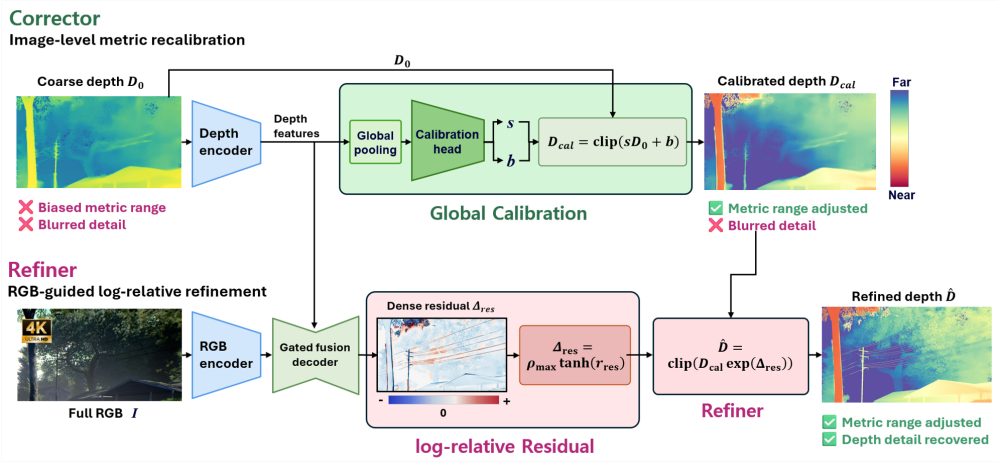

Figure 3: **ONCE architecture.** The Corrector predicts image-level scale and shift from globally pooled depth features. The Refiner fuses spatial depth and RGB features to predict a bounded log-relative correction, which is applied to the calibrated depth. Both branches operate within a single forward pass. The depth maps and residual visualization schematically illustrate metric adjustment and detail recovery.

converted into a bounded log-depth update:

$$\begin{aligned} \Delta_{\text{res}} &= \rho_{\max} \tanh(r_{\text{res}}), \\ \widetilde{D} &= D_{\text{cal}} \exp(\Delta_{\text{res}}), \\ \hat{D} &= \text{clip}\left(\widetilde{D}, d_{\min}, d_{\max}\right). \end{aligned} \tag{3}$$

All operations are element-wise. Before final clipping, this is equivalent to $\log \widetilde{D} = \log D_{\text{cal}} + \Delta_{\text{res}}$. With $\rho_{\max} = 0.5$, the update satisfies

$$e^{-0.5} \leq \frac{\widetilde{D}}{D_{\text{cal}}} \leq e^{0.5}. \tag{4}$$

Its absolute correction is $D_{\text{cal}}\big(\exp(\Delta_{\text{res}}) - 1\big)$. Thus, a given residual produces the same relative change across depths, while its absolute magnitude scales with the calibrated depth. A bounded additive residual instead imposes a fixed correction range in metric units.

Spatially varying residuals allow neighboring pixels to receive different corrections, supporting the recovery of thin structures and depth boundaries. The log-relative parameterization controls the magnitude of these updates; boundary localization is learned from the fused RGB and depth features. Alternative parameterizations and their bounds are detailed in Appendix A.4, with comparisons in Section 4.4.

3.4 Dual-Encoder Gated Fusion

The RGB and depth encoders separately process I and D_0. In the five-level configuration, the RGB encoder produces (r_1, r_2, r_3, r_4) at resolutions 1, $1/2$, $1/4$, and $1/8$ of the input. The depth encoder produces $(d_1, \ldots, d_5)$, with d_5 at $1/16$ resolution. The first block operates at full resolution, followed by stride-2 downsampling blocks. Feature widths are C, $2C$, $4C$, $8C$, and $8C$, with $C = 96$. Dropout with probability 0.2 is applied at the deepest stages.

The Refiner's decoder starts from d_5 and progressively upsamples the depth representation. At each of four scales, a gated-fusion block combines the decoder feature with the corresponding RGB feature. The gate is computed from joint RGB–depth information and modulates the decoder features. The final fusion output is passed to a residual head that predicts r_{res} at the target resolution.

The depth bottleneck supports both branches: its pooled representation feeds the Corrector, while its spatial representation initializes the Refiner's decoder. The Refiner uses features extracted from the original coarse depth D_0. The calibrated depth D_{cal} enters the final multiplicative update, so no second depth encoding is required. The two branches are jointly optimized and are not constrained to correct disjoint error components. Model capacity is adjusted through the number of refinement levels.

3.5 Training Objective

Given ground-truth metric depth D^* and a valid-pixel mask M, we train ONCE with

$$\mathcal{L} = \lambda_1 \mathcal{L}_1 + \lambda_{\text{smooth}} \mathcal{L}_{\text{smooth}} + \lambda_{\text{grad}} \mathcal{L}_{\text{grad}}. \tag{5}$$

We set $\lambda_1 = 1$, $\lambda_{\text{smooth}} = 0.5$, and $\lambda_{\text{grad}} = 0.8$. The pixel-wise terms supervise metric accuracy, gradient matching encourages structural consistency, and RGB-guided smoothness regularizes local depth variation. The objective supervises the final refined depth and jointly trains the encoders, Corrector, and Refiner. The foundation model for coarse depth remains frozen. Loss definitions, weights, and domain-specific training details are provided in Appendix A.3.

4 Experiments

We evaluate whether ONCE can improve metric accuracy and recover fine structure in a single full-frame refinement pass. We compare its accuracy and refinement cost with patch-based methods, evaluate two coarse backbones, and examine the contributions of the Corrector, the Refiner's log-relative formulation, and native-resolution fine-tuning.

4.1 Experimental Setup

We evaluate on UnrealStereo4K (Tosi et al., 2021) at its native 3840×2160 resolution, following the protocol of PatchFusion and PatchRefiner (Li et al., 2024b;c). Predictions are evaluated without ground-truth scale or shift alignment. We report δ_1 (in percent), absolute relative error (REL), root mean squared error (RMSE), scale-invariant logarithmic error (SiLog), and Soft Edge Error (SEE).

ONCE uses frozen metric-depth predictions from ZoeDepth or Depth Anything V2 (DAv2). The S, M, and L variants use 3, 5, and 7 refinement levels, respectively. Training consists of crop-based learning followed by native-4K fine-tuning; inference uses one full-frame ONCE pass after coarse depth estimation. Baseline accuracy scores are taken from the cited papers. We do not include SharpDepth (Pham et al., 2025) in the quantitative comparison. SharpDepth is designed for zero-shot generalization to real-world images and refines the prediction of a UniDepth backbone without target-domain training, whereas the compared methods and ONCE refine a coarse backbone adapted to UnrealStereo4K. Its scores under our protocol would therefore reflect differences in backbone and training data rather than in the refinement design.

Parameter counts and inference times cover only the refinement component and exclude the coarse backbone. All timings are measured on a single NVIDIA A100 under the same evaluation setting. Training and evaluation details are provided in Appendix A.1, and additional results after target-domain fine-tuning on Cityscapes are reported in Appendix A.6.

4.2 Quantitative Results

Table 1 compares ONCE with existing high-resolution depth estimation and refinement methods. ONCE_{L} achieves the best results on all five depth metrics among the listed ZoeDepth-based refiners. Compared with PRV2_{C}, it reduces RMSE from 0.884 to 0.784 (11.3%) and SEE from 0.787 to 0.608 (22.7%), while increasing δ_1 by 0.189 percentage points. These gains are obtained with 192.9M refinement parameters rather than 245.8M (21.5% fewer), and an inference time of 0.54 s rather than 0.62 s. Thus, a single full-frame correction improves both metric accuracy and boundary quality over repeated patch-wise refinement in this comparison.

ONCE_{M} also improves all five depth metrics over PRV2_{E}, with 25.7% fewer refinement parameters (53.6M vs. 72.1M). RMSE decreases from 0.948 to 0.831 and SEE from 0.816 to 0.666, while

Table 1: Comparison on UnrealStereo4K. Best depth accuracy in each metric is shown in **bold**. † denotes the re-evaluation reported by PatchRefiner V2. #param counts only parameters added for high-resolution refinement; the coarse depth backbone is excluded. T denotes refiner-only inference time for one 4K image. All timings use a single NVIDIA A100 under the same evaluation setting.

Method	$\delta_1(\%)\uparrow$	REL↓	RMSE↓	SiLog↓	SEE↓	#param↓	T↓	Reference
iDisc Piccinelli et al. (2023)	96.940	0.053	1.404	8.502	1.070	–	–	CVPR 2023
SMD-Net Tosi et al. (2021)	97.774	0.044	1.282	7.389	0.883	–	–	CVPR 2021
Graph-GDSR de Lutio et al. (2022)	97.932	0.044	1.264	7.469	0.872	–	–	CVPR 2022
BoostingDepth Miangoleh et al. (2021)	98.104	0.044	1.123	6.662	0.939	–	–	CVPR 2021
ZoeDepth Bhat et al. (2023)	97.717	0.046	1.289	7.448	0.914	–	–	–
ZoeDepth + PF Li et al. (2024b)	98.419	0.040	1.088	6.212	0.838	432.7M	3.44s	CVPR 2024
ZoeDepth + PF† Li et al. (2024b)	98.369	0.039	1.064	6.342	0.855	432.7M	3.44s	
ZoeDepth + PR Li et al. (2024c)	98.821	0.033	0.892	5.417	0.750	369.0M	1.45s	ECCV 2024
ZoeDepth + PR† Li et al. (2024c)	98.680	0.034	0.941	5.614	0.771	369.0M	1.45s	
ZoeDepth + PRV2_M Li et al. (2026)	98.610	0.034	1.003	5.760	0.832	47.0M	0.32s	ICLR 2026
ZoeDepth + PRV2_E Li et al. (2026)	98.728	0.034	0.948	5.579	0.816	72.1M	0.57s	
ZoeDepth + PRV2_C Li et al. (2026)	98.863	0.032	0.884	5.281	0.787	245.8M	0.62s	
ZoeDepth + ONCE_S **(Ours)**	98.353	0.037	1.066	6.308	0.799	5.1M	0.42s	**Ours**
ZoeDepth + ONCE_M **(Ours)**	98.965	0.033	0.831	5.300	0.666	53.6M	0.51s	
ZoeDepth + ONCE_L **(Ours)**	**99.052**	**0.031**	**0.784**	**5.123**	**0.608**	192.9M	0.54s	

Table 2: Performance and relative improvements over the corresponding coarse backbone. Parentheses show relative percentage changes, including for δ_1; these are not percentage-point differences.

Method	$\delta_1(\%)\uparrow$	RMSE↓	SEE↓
ZoeDepth	97.717	1.289	0.914
+ PRV2_C (Li et al., 2026)	98.863 (+1.17%)	0.884 (-31.42%)	0.787 (-13.89%)
+ ONCE_M (Ours)	98.965 (+1.28%)	0.831 (-35.53%)	0.666 (-27.13%)
+ ONCE_L (Ours)	**99.052 (+1.37%)**	**0.784 (-39.19%)**	**0.608 (-33.44%)**
DAv2	98.072	1.144	0.908
+ PRV2_C (Li et al., 2026)	98.929 (+0.87%)	0.859 (-24.91%)	0.779 (-14.21%)
+ ONCE_M (Ours)	99.103 (+1.05%)	0.780 (-31.82%)	0.639 (-29.63%)
+ ONCE_L (Ours)	**99.130 (+1.08%)**	**0.748 (-34.62%)**	**0.603 (-33.59%)**

refinement time decreases from 0.57 s to 0.51 s. ONCE_S provides a smaller alternative with 5.1M parameters, although PRV2_M has the lowest latency in the table. The parameter and latency results therefore describe different aspects of refinement cost.

Table 2 reports gains over each coarse backbone. With ZoeDepth, ONCE_L reduces RMSE by 39.2% and SEE by 33.4%. With DAv2, both ONCE_M and ONCE_L outperform PRV2_C on all three reported metrics. ONCE_L reaches $\delta_1 = 99.130\%$, RMSE $= 0.748$, and SEE $= 0.603$. The DAv2 comparisons use a shared coarse checkpoint (Appendix A.1), supporting the applicability of ONCE to another metric-depth backbone.

4.3 Qualitative Results

Figure 4 compares ONCE_M with PRV2_E and their coarse ZoeDepth input. Both refiners recover detail missing from the coarse prediction, but ONCE provides clearer separation of thin structures and their surroundings. The second row shows more distinct wires around the utility pole, while the fourth row shows finer palm-frond boundaries. The example in Figure 1 also illustrates reduced background depth discrepancies in the yellow regions and clearer fence bars and gaps in the black regions. Together, these examples show improvements in metric structure and boundary recovery. The following ablations evaluate the Corrector and residual formulation separately. Additional UnrealStereo4K examples are provided in Appendix A.5.

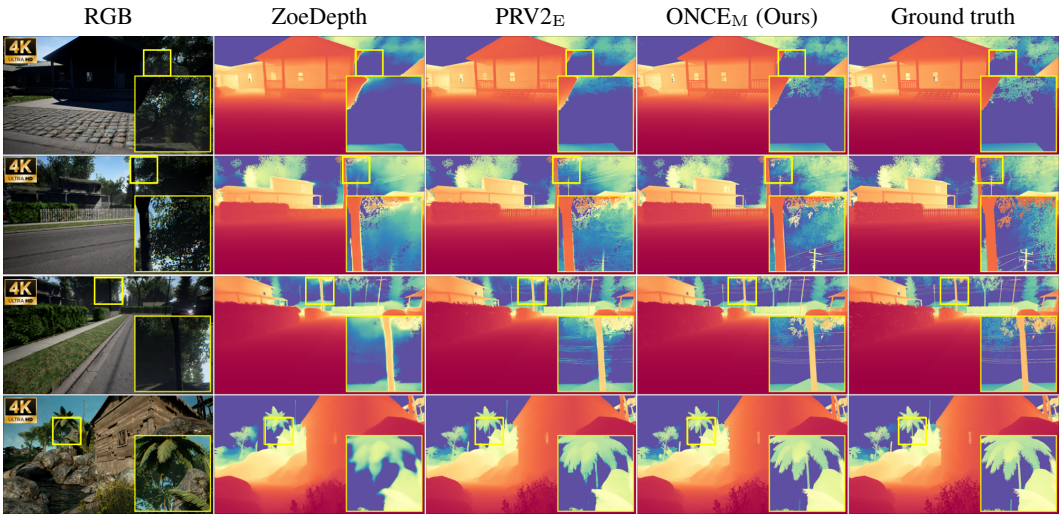

Figure 4: Qualitative depth results on UnrealStereo4K. From left to right: RGB, coarse ZoeDepth, PRV2_E, ONCE_M, and ground-truth depth. Yellow boxes mark enlarged regions containing foliage, thin structures, and depth discontinuities. Both refinement methods use ZoeDepth as the coarse backbone.

Table 3: Ablation of global metric calibration on UnrealStereo4K using ZoeDepth-based ONCE_M after Stage 2. s_θ and b_θ denote learned image-level scale and shift parameters. Best results are shown in bold.

Configuration	s	b	$\delta_1(\%)\uparrow$	RMSE↓	SiLog↓
No calibration	1	0	98.771	0.967	5.359
Scale only	s_θ	0	98.840	0.906	5.544
Shift only	1	b_θ	98.854	0.884	**5.292**
Scale + shift (full)	s_θ	b_θ	**98.965**	**0.831**	5.300

Table 4: Refiner residual-parameterization ablation on UnrealStereo4K using ZoeDepth-based ONCE_M after Stage 2. Bold indicates the best result for each metric.

Formulation	$\delta_1(\%)\uparrow$	$\delta_2(\%)\uparrow$	$\delta_3(\%)\uparrow$	AbsRel↓	SqRel↓	RMSE↓	RMSElog↓	$\log_{10}\downarrow$	SiLog↓	SEE↓
Log-relative (ours)	**98.965**	**99.641**	**99.835**	**0.0325**	**0.0590**	**0.8311**	**0.0580**	**0.0142**	**5.300**	**0.6660**
Absolute	98.634	99.554	99.815	0.0349	0.0782	0.9820	0.0623	0.0152	5.752	0.8149
Linear-relative	98.722	99.578	99.823	0.0347	0.0684	0.9315	0.0621	0.0153	5.669	0.7701

4.4 Ablation Studies

Corrector: global metric calibration. We compare four calibration configurations using ZoeDepth-based ONCE_M after Stage 2 (Table 3). The frozen coarse checkpoint, Refiner, log-relative formulation, training objective, initialization protocol, and training budget are held fixed. Without calibration, $s = 1$ and $b = 0$; scale-only and shift-only variants learn one parameter while fixing the other to its identity value. The full Corrector learns both parameters.

Joint scale-and-shift calibration achieves the highest δ_1 and lowest RMSE. Compared with no calibration, it increases δ_1 from 98.771% to 98.965% and reduces RMSE from 0.967 to 0.831, a 14.1% reduction. Learning both parameters also improves these two metrics over either parameter alone. SiLog decreases from 5.359 to 5.300 relative to no calibration, although shift-only calibration obtains the lowest SiLog of 5.292. The Corrector therefore improves final metric accuracy even when the Refiner already predicts dense depth-relative corrections, with the strongest gains in δ_1 and RMSE.

Refiner: residual parameterization. We compare additive, linear-relative, and log-relative corrections using ZoeDepth-based ONCE_{M} under the same Stage 2 training and evaluation protocol. The Corrector, network architecture, and training objective are unchanged. The formulations, correction amplitudes, and clipping convention are given in Appendix A.4.

As shown in Table 4, log-relative refinement achieves the best results across all ten reported metrics. Compared with additive correction, it reduces RMSE from 0.9820 to 0.8311 (15.4%) and SEE from 0.8149 to 0.6660 (18.3%). It also outperforms linear-relative correction, reducing RMSE by 10.8% and SEE by 13.5%. These results support the log-relative formulation for both depth accuracy and boundary recovery under the tested settings. Since the three formulations use different correction ranges, this comparison evaluates their specified parameterizations, rather than the exponential mapping under matched output bounds.

Table 5: Effect of the native-4K fine-tuning stage on ZoeDepth-based ONCE_{M}. Changes are relative for error metrics and in percentage points (pp) for δ_1.

Metric	Stage 1 (1024^2 crops)	Stage 2 (native 4K)	Change
$\delta_1(\%) \uparrow$	98.751	**98.965**	$+0.214$ pp
REL$\downarrow$	0.041	**0.033**	-19.5%
RMSE$\downarrow$	0.948	**0.831**	-12.3%
SiLog$\downarrow$	6.098	**5.300**	-13.1%
SEE$\downarrow$	0.756	**0.666**	-11.9%

Native full-resolution fine-tuning. Table 5 compares the crop-trained Stage 1 checkpoint with the Stage 2 checkpoint of ZoeDepth-based ONCE_{M}, using the same architecture and training objective. Native-4K fine-tuning improves all five metrics, increasing δ_1 by 0.214 percentage points and reducing RMSE and SEE by 12.3% and 11.9%, respectively. The second stage benefits both metric accuracy and boundary recovery. These gains reflect the complete fine-tuning stage, including its additional optimization steps, rather than the effect of training resolution alone.

The loss-component ablation in Appendix A.3.1 supports the use of ℓ_1, RGB-guided smoothness, and gradient matching: this combination achieves the highest δ_1, lowest RMSE, and tied-lowest AbsRel among the tested objectives.

5 Conclusion

We presented ONCE, a patch-free metric-depth booster that combines a learned Corrector and an RGB-guided Refiner in a single full-frame pass. The Corrector adjusts the coarse metric reference through image-level scale and shift, while the Refiner recovers spatial detail through bounded log-relative updates. On UnrealStereo4K, ONCE_{L} outperforms the compared ZoeDepth-based refiners across all five depth metrics, using 21.5% fewer refinement parameters than PRV2_{C}. The calibration ablation shows that joint scale-and-shift learning improves δ_1 and RMSE over no calibration and either parameter alone. The residual ablation favors log-relative correction over additive and linear-relative alternatives under the tested settings. Native-resolution fine-tuning further improves accuracy, with gains reflecting both full-frame training and additional optimization. Results with a shared DAv2 coarse checkpoint and Cityscapes fine-tuning support applicability beyond the primary setting. Together, these findings establish full-frame metric correction as an effective and parameter-efficient alternative to repeated patch-wise refinement.

References

Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4009–4018, 2021.

Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023.

Alexey Bochkovskiy, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. In *International Conference on Learning Representations (ICLR)*, 2025.

Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8628–8638, 2021.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3213–3223, 2016.

Wenqing Cui, Zhenyu Li, Mykola Lavreniuk, Jian Shi, Ramzi Idoughi, Xiangjun Tang, and Peter Wonka. Any resolution any geometry: From multi-view to multi-patch. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12576–12585, 2026.

Yaqiao Dai, Renjiao Yi, Chenyang Zhu, Hongjun He, and Kai Xu. Multi-resolution monocular depth map fusion by self-supervised gradient-based composition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

Riccardo de Lutio, Alexander Becker, Stefano D'Aronco, Simone Russo, Jan Dirk Wegner, and Konrad Schindler. Learning graph regularisation for guided super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 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 (NeurIPS)*, 2014.

Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2002–2011, 2018.

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. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Tak-Wai Hui, Chen Change Loy, and Xiaoou Tang. Depth map super-resolution by deep multi-scale guidance. In *European Conference on Computer Vision (ECCV)*, pp. 353–369, 2016.

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9492–9502, 2024.

Byeongjun Kwon and Munchurl Kim. One look is enough: Seamless patchwise refinement for zero-shot monocular depth estimation on high-resolution images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8077–8087, 2025.

Jiaqi Li, Yiran Wang, Jinghong Zheng, Zihao Huang, Ke Xian, Zhiguo Cao, and Jianming Zhang. Self-distilled depth refinement with noisy poisson fusion. In *Advances in Neural Information Processing Systems*, 2024a.

Zhenyu Li, Shariq Farooq Bhat, and Peter Wonka. Patchfusion: An end-to-end tile-based framework for high-resolution monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024b.

Zhenyu Li, Shariq Farooq Bhat, and Peter Wonka. Patchrefiner: Leveraging synthetic data for real-domain high-resolution monocular metric depth estimation. In *European Conference on Computer Vision (ECCV)*, 2024c.

Zhenyu Li, Wenqing Cui, Shariq Farooq Bhat, and Peter Wonka. Patchrefiner v2: Fast and lightweight real-domain high-resolution metric depth estimation. In *International Conference on Learning Representations (ICLR)*, 2026.

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 (CVPR)*, pp. 17070–17080, 2025.

Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Guided depth super-resolution by deep anisotropic diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18237–18246, 2023.

S. Mahdi H. Miangoleh, Sebastian Dille, Long Mai, Sylvain Paris, and Yağız Aksoy. Boosting monocular depth estimation models to high-resolution via content-adaptive multi-resolution merging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Duc-Hai Pham, Tung Do, Phong Nguyen, Binh-Son Hua, Khoi Nguyen, and Rang Nguyen. Sharpdepth: Sharpening metric depth predictions using diffusion distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17060–17069, 2025.

Luigi Piccinelli, Christos Sakaridis, and Fisher Yu. idisc: Internal discretization for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10106–10116, 2024.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12179–12188, 2021.

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1623–1637, 2022.

Fabio Tosi, Yiyi Liao, Carolin Schmitt, and Andreas Geiger. Smd-nets: Stereo mixture density networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

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*, pp. 10371–10381, 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 (NeurIPS)*, volume 37, pp. 21875–21911, 2024b.

Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3D: Towards zero-shot metric 3D prediction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9043–9053, 2023.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Yifan Yu, Shaohui Liu, Rémi Pautrat, Marc Pollefeys, and Viktor Larsson. Relative pose estimation through affine corrections of monocular depth priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16706–16716, 2025.

Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. NeW CRFs: Neural window fully-connected CRFs for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Xiang Zhang, Bingxin Ke, Hayko Riemenschneider, Nando Metzger, Anton Obukhov, Markus Gross, Konrad Schindler, and Christopher Schroers. BetterDepth: Plug-and-play diffusion refiner for zero-shot monocular depth estimation. In *Advances in Neural Information Processing Systems*, 2024.

Xiang Zhang, Yang Zhang, Lukas Mehl, Markus Gross, and Christopher Schroers. Guardians of the hair: Rescuing soft boundaries in depth, stereo, and novel views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Zhiwei Zhong, Xianming Liu, Junjun Jiang, Debin Zhao, and Xiangyang Ji. Guided depth map super-resolution: A survey. *ACM Computing Surveys*, 2023.

A Appendix

A.1 Experimental Details

A.1.1 Data preparation and evaluation

UnrealStereo4K (Tosi et al., 2021) contains 3840×2160 stereo frames from eight synthetic indoor and outdoor environments. We convert disparity to metric depth using the calibrated focal length and baseline of each scene. We follow the data splits and SSIM-based filtering procedure of PatchFusion and PatchRefiner (Li et al., 2024b;c) to exclude samples with invalid disparity annotations. UnrealStereo4K evaluation uses positive-disparity pixels with ground-truth depth in $[10^{-3}, 80)$ m. Predictions are evaluated at native resolution without fitting a scale or shift to ground truth.

We report δ_1 in percent, REL (absolute relative error), RMSE, SiLog, and SEE using the same metric definitions as the cited protocol. Published baseline scores are retained in Table 1. The dagger-marked PatchFusion and PatchRefiner entries are the re-evaluations reported by PRV2 (Li et al., 2026), with the non-public MiDaS pretraining stage removed from the fine/refiner branch. The dagger does not denote ground-truth alignment. Cityscapes uses the separate evaluation protocol described in Appendix A.6.

A.1.2 Backbones and training protocol

We use metric-adapted ZoeDepth and DAv2 as coarse backbones and freeze their parameters during refinement training. Coarse predictions are precomputed and resized to the target resolution. Stage 1 uses random 1024×1024 crops, photometric augmentation, and horizontal flipping. Stage 2 continues training on native, uncropped 4K images. The architecture and objective are held fixed within each reported Stage 1–Stage 2 comparison. At inference, one coarse-depth prediction is followed by one full-frame ONCE refinement pass.

Training uses AdamW with weight decay 5×10^{-2} and a cosine-annealing warm-restart scheduler with $T_0 = 10$ and $T_{\text{mult}} = 2$. Stage 1 runs for 150 epochs with initial learning rate 3.46×10^{-5} and effective batch size 80. Stage 2 runs for 150 epochs with initial learning rate 1.062×10^{-5} and effective batch size 8. The additional Stage 2 training changes both resolution and optimization budget; its comparison with Stage 1 does not isolate the effect of resolution alone.

For the DAv2 comparison, ONCE_{M}, ONCE_{L}, and PRV2_{C} use the same frozen, metric-adapted DAv2 checkpoint. The coarse DAv2 row in Table 2 is evaluated from this shared checkpoint, and all corresponding relative improvements use that row as their reference. This controls the coarse-checkpoint choice, while the refinement methods retain their respective training procedures.

A.1.3 Inference cost and measurement conditions

The reported refinement latencies for ONCE_{M} and ONCE_{L} are 0.51 s and 0.54 s per 4K image, respectively; the corresponding PRV2_{E} and PRV2_{C} latencies are 0.57 s and 0.62 s. These timings refer to refinement on a single NVIDIA A100 and exclude coarse-backbone inference. Parameter

counts likewise include only the added refinement component. They therefore describe the cost of boosting a coarse prediction, rather than the complete RGB-to-depth pipeline.

A.2 Calibration and Refinement Implementation

A.2.1 Corrector: global metric calibration

For the five-level ONCE_{M} configuration, the Corrector predicts two scalars per image from the globally pooled depth-encoder bottleneck d_5:

$$(g_s, g_b) = \text{Conv}_1(\text{ReLU}(\text{Conv}_1(\text{GAP}(d_5)))) \,. \tag{6}$$

Here, GAP denotes global average pooling and Conv_1 a 1×1 convolution. The calibrated depth is

$$\begin{aligned} s &= \exp(s_{\max} \tanh(g_s)) \,, \qquad b = b_{\max} \tanh(g_b), \\ D_{\text{cal}} &= \text{clip}(sD_0 + b, d_{\min}, d_{\max}). \end{aligned} \tag{7}$$

The same s and b are broadcast to every pixel. For UnrealStereo4K, we use $s_{\max} = 0.4$, $b_{\max} = 5.0$ m, and $[d_{\min}, d_{\max}] = [10^{-3}, 80]$ m. Positive scale preserves depth ordering before clipping; clipping can flatten values at the interval boundaries. Both calibration and dense refinement are learned through the final depth objective, without fitting calibration parameters to ground truth at inference.

A.2.2 Dual-encoder architecture and gated fusion

For ONCE_{M}, the RGB encoder produces (r_1, r_2, r_3, r_4) at resolutions 1, $1/2$, $1/4$, and $1/8$ of the input. The depth encoder produces $(d_1, \ldots, d_5)$, with d_5 at $1/16$ resolution. The first block operates at full resolution, followed by stride-2 downsampling blocks. The reference architecture uses base width $C = 96$ and depth-encoder widths $(C, 2C, 4C, 8C, 8C)$. Dropout with probability 0.2 is applied at the deepest stages.

For an RGB feature r and a decoder feature u at the same resolution, gated fusion is defined as

$$\begin{aligned} q = \text{Conv}_3(r), \quad k = \text{Conv}_3(u), \quad v = \text{Conv}_3(u), \\ w = \sigma(\text{Conv}_1(\text{ReLU}(\text{BN}(\text{Conv}_3([q; k]))))) \,, \\ \text{Gate}(r, u) = \text{Conv}_3(v \odot w). \end{aligned} \tag{8}$$

Here, Conv_3 denotes a 3×3 convolution, $[\cdot;\cdot]$ channel concatenation, $\odot$ element-wise multiplication, and σ the sigmoid function. The decoder starts from d_5, successively upsamples its features, and fuses them with r_4, r_3, r_2, r_1. A residual head maps the final feature to r_{res}. The Refiner predicts this residual from RGB and coarse-depth features and forms the final prediction as

$$\begin{aligned} \Delta_{\text{res}} &= \rho_{\max} \tanh(r_{\text{res}}), \\ \hat{D} &= \text{clip}(D_{\text{cal}} \exp(\Delta_{\text{res}}), d_{\min}, d_{\max}) \,, \end{aligned} \tag{9}$$

with $\rho_{\max} = 0.5$. The decoder uses features of the original coarse depth D_0; D_{cal} enters the final multiplicative update without a second depth-encoding pass. The Corrector and Refiner are trained jointly and need not correct disjoint error components.

A.3 Training Objective Details

The final UnrealStereo4K objective combines metric-depth supervision, gradient matching, and RGB-guided smoothness:

$$\mathcal{L} = \lambda_1 \mathcal{L}_1 + \lambda_{\text{smooth}} \mathcal{L}_{\text{smooth}} + \lambda_{\text{grad}} \mathcal{L}_{\text{grad}}. \tag{10}$$

We use $\lambda_1 = 1$, $\lambda_{\text{grad}} = 0.8$, and $\lambda_{\text{smooth}} = 0.5$. AbsRel and SiLog penalties are evaluated as alternatives in the loss ablation but are not included in the final objective. AbsRel remains an evaluation metric.

Let M be the set of valid ground-truth pixels, D^* the metric ground-truth depth, and $\hat{D}$ the final refined prediction. The pixel-wise term is

$$\mathcal{L}_1 = \frac{1}{|M|} \sum_{p \in M} |\hat{D}_p - D^*_p|. \tag{11}$$

Table 6: Loss-component ablation on UnrealStereo4K using ZoeDepth-based ONCE_{M} after Stage 2. Each row lists the complete set of active loss terms. The final objective uses ℓ_1, RGB-guided smoothness, and gradient matching, without AbsRel or SiLog supervision. AbsRel in the metric columns is the same metric as REL. Bold indicates the best reported value, including ties.

Configuration	$\delta_1(\%)\uparrow$	RMSE↓	AbsRel↓	SiLog↓	SEE↓
ℓ_1 only	98.952	0.834	0.0326	5.305	0.671
ℓ_1 + SiLog	98.937	0.835	0.0333	5.266	0.689
ℓ_1 + AbsRel	98.941	0.837	0.0331	**5.262**	0.690
ℓ_1 + RGB Smooth	98.954	0.838	**0.0325**	5.327	**0.664**
ℓ_1 + RGB Smooth + Grad (ours)	**98.965**	**0.831**	**0.0325**	5.300	0.666

It directly supervises metric depth and trains both the Corrector and Refiner through their combined output.

The gradient term penalizes differences between first-order depth derivatives. Writing M_j for valid derivative locations in direction $j \in \{x, y\}$, it is defined as

$$\mathcal{L}_{\text{grad}} = \sum_{j\in\{x,y\}} \text{mean}_{p\in M_j} \left|\partial_j \hat{D}_p - \partial_j D^*_p\right|. \tag{12}$$

Only locations supported by valid ground-truth pixels contribute to this term.

The RGB-guided smoothness term is

$$\mathcal{L}_{\text{smooth}} = \sum_{j\in\{x,y\}} \text{mean}\left(e^{-\,\text{mean}_c(|\partial_j I|)}|\partial_j \hat{D}|\right), \tag{13}$$

where mean_c averages over RGB channels and the outer mean averages over spatial derivative locations. Strong image gradients reduce the smoothness penalty, allowing depth boundaries to follow image evidence. All three terms operate on metric depth; the log-relative parameterization determines how the refined prediction is formed.

A.3.1 Loss-component ablation

We use ZoeDepth-based ONCE_{M} and report Stage 2 results for all configurations. Table 6 compares an ℓ_1 baseline with additional SiLog, AbsRel, or RGB-guided smoothness supervision, followed by the final combination of ℓ_1, RGB-guided smoothness, and gradient matching. The alternatives are separate configurations; the rows do not represent cumulative additions of all preceding terms.

The final objective achieves the highest δ_1 and lowest RMSE, and ties with ℓ_1 plus RGB-guided smoothness for the lowest AbsRel. Relative to ℓ_1 alone, it increases δ_1 from 98.952% to 98.965% and reduces RMSE from 0.834 to 0.831. Adding gradient matching to the RGB-smoothness configuration improves RMSE from 0.838 to 0.831 and SiLog from 5.327 to 5.300, with a small increase in SEE from 0.6644 to 0.666.

The alternatives show a trade-off between metrics. Adding AbsRel or SiLog supervision to ℓ_1 improves SiLog but worsens δ_1, RMSE, AbsRel, and SEE relative to ℓ_1 alone. The ℓ_1 plus AbsRel configuration obtains the lowest SiLog, while ℓ_1 plus RGB-guided smoothness obtains the lowest SEE. We use the three-term objective for its best δ_1 and RMSE and tied-best AbsRel, while retaining low boundary error. The differences are small, so the table supports this choice under the reported setting rather than uniform superiority on every metric.

A.4 Residual Parameterizations for Ablation

The residual ablation compares three transformations of the same predicted residual r_{res}. Let $\tilde{D}$ denote depth before the final clipping operation:

$$\begin{aligned} \text{Absolute:} \quad & \tilde{D}_{\text{abs}} = D_{\text{cal}} + a_{\text{abs}} \tanh(r_{\text{res}}), \\ \text{Linear-relative:} \quad & \tilde{D}_{\text{lin}} = D_{\text{cal}} \left(1 + a_{\text{rel}} \tanh(r_{\text{res}})\right), \\ \text{Log-relative:} \quad & \tilde{D}_{\text{log}} = D_{\text{cal}} \exp\left(\rho_{\max} \tanh(r_{\text{res}})\right). \end{aligned} \tag{14}$$

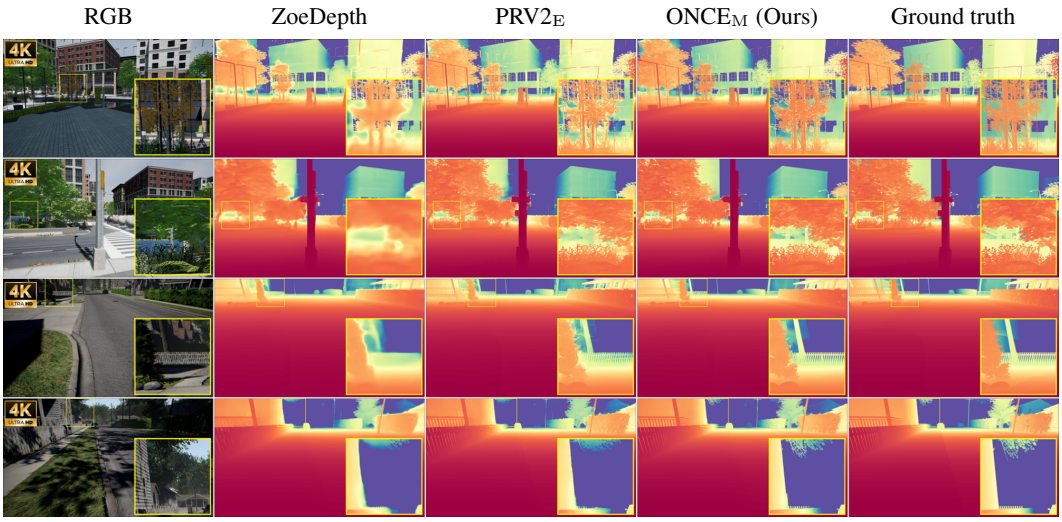

Figure 5: Additional qualitative results on UnrealStereo4K. Columns show RGB, coarse ZoeDepth, PRV2_E, ONCE_M, and ground-truth depth. Yellow boxes mark regions shown in the enlarged insets. Both refiners use ZoeDepth as the coarse backbone.

For each variant v, the evaluated prediction is $\hat{D}_v = \text{clip}(\tilde{D}_v, d_{\min}, d_{\max})$. We use $a_{\text{abs}} = 10.0$ m, $a_{\text{rel}} = 0.5$, and $\rho_{\max} = 0.5$. The absolute amplitude has metric-depth units; the other two amplitudes are dimensionless.

Table 4 reports Stage 2 results for all three formulations. The variants use the same network architecture, training objective, and training budget, with the residual formulation changed as specified above.

Since $-1 \leq \tanh(r_{\text{res}}) \leq 1$, the log-relative update obeys

$$e^{-0.5} \leq \frac{\tilde{D}_{\log}}{D_{\text{cal}}} \leq e^{0.5}. \tag{15}$$

Thus, its multiplicative range is approximately $[0.607, 1.649]$, whereas the linear-relative range is $[0.5, 1.5]$. The absolute variant uses an additive range of $[-10, 10]$ m. These are bounds before clipping. The comparison evaluates the specified parameterizations and amplitudes; it does not isolate the exponential mapping under identical correction bounds.

A.5 Additional Qualitative Results on UnrealStereo4K

Figure 5 supplements the main-paper comparison with additional UnrealStereo4K scenes. The five columns show RGB, coarse ZoeDepth, PRV2_E, ONCE_M, and ground-truth depth. The first two rows highlight foliage and small gaps between foreground vegetation and background structures. The last two rows show fence elements, narrow vertical structures, and vegetation near depth boundaries. Both refiners recover detail that is blurred in the coarse prediction. In the enlarged regions, ONCE retains fine geometric structure and separates foreground objects from their surroundings. These examples complement the metric and boundary evaluations; they show the complete model's output rather than isolating the Corrector or Refiner.

A.6 Real-World Evaluation on Cityscapes

We evaluate ONCE_M on Cityscapes (Cordts et al., 2016) after target-domain fine-tuning, using ZoeDepth as the frozen coarse backbone. We follow the official PRV2 evaluation procedure, including the segmentation- and depth-based reference boundaries for EdgeComp and F1. Table 7 includes the published PRV2_E result with SSIGM and offline pseudo-labels (Li et al., 2026). This comparison evaluates the reported systems with their respective training objectives rather than isolating the refinement architecture.

The reported ONCE checkpoint is fine-tuned for 60 epochs at a learning rate of 5×10^{-6}, with batch-normalization statistics frozen. The base objective is the three-term UnrealStereo4K objective defined in Appendix A.3:

$$\mathcal{L}_{\text{base}} \equiv \mathcal{L} = \mathcal{L}_1 + 0.8\mathcal{L}_{\text{grad}} + 0.5\mathcal{L}_{\text{smooth}}. \tag{16}$$

For real-world adaptation, we retain these terms and add a boundary-focused depth loss to emphasize geometric discontinuities.

Let M denote the valid ground-truth pixel set. We compute the Sobel gradient magnitude of ground-truth log-depth and select boundary pixels using a per-image quantile:

$$\begin{aligned} g(p) &= \left\|\nabla_{\text{Sobel}} \log D^*(p)\right\|_2, \\ B &= \left\{p \in M : g(p) > Q_{0.85}(\{g(q) : q \in M\})\right\}. \end{aligned} \tag{17}$$

Here, $Q_{0.85}$ denotes the 0.85 quantile over valid pixels. The additional loss and the resulting Cityscapes objective are

$$\begin{aligned} \mathcal{L}_{\text{boundary}} &= \frac{1}{|B|} \sum_{p \in B} \left|\hat{D}_p - D_p^*\right|, \\ \mathcal{L}_{\text{City}} &= \mathcal{L}_{\text{base}} + \zeta \mathcal{L}_{\text{boundary}}, \qquad \zeta = 8.0. \end{aligned} \tag{18}$$

The additional term concentrates metric-depth supervision on ground-truth depth discontinuities, complementing the three-term UnrealStereo4K objective. The training mask B is distinct from the segmentation- and depth-based reference boundaries used to evaluate EdgeComp and F1.

Table 7: Cityscapes results after target-domain adaptation. PRV2_{E} uses SSIGM with offline pseudo-labels. ONCE_{M} uses the three-term UnrealStereo4K objective and adds boundary-focused supervision as defined in Equation 18. Bold indicates the best reported accuracy in each column. Parameter counts exclude the coarse backbone.

Method	$\delta_1(\%) \uparrow$	RMSE↓	EdgeComp↓	F1(%) ↑	Params
ZoeDepth	92.246	9.095	41.340	18.920	–
PRV2_{E} + SSIGM	**92.889**	**8.530**	**16.970**	**34.850**	72.1M
ONCE_{M} (Ours)	92.712	8.827	24.522	26.414	53.6M

ONCE improves the coarse ZoeDepth baseline on all four accuracy metrics, reducing RMSE by approximately 3.0% and EdgeComp by 40.7%. PRV2_{E} + SSIGM remains stronger on all four metrics, whereas ONCE uses 25.7% fewer refinement parameters. These results support applicability after real-domain adaptation while retaining a clear accuracy gap to PRV2 under the reported settings. The comparison evaluates the complete fine-tuned model and does not separately establish the contribution of the additional boundary loss.

Figure 6 compares the coarse ZoeDepth prediction with both refined outputs. The coarse maps capture the main road, vehicle, and building layout but smooth local depth variation around vegetation and narrow structures. The first two rows highlight vegetation boundaries and utility structures, where ONCE introduces more distinct local depth variation in selected regions. The third and fourth rows show roof and roadside object boundaries, with some contours remaining similarly smooth across the predictions. The examples illustrate local refinement of the coarse depth, while the aggregate comparison is reported in Table 7.

The reference depth panels contain incomplete regions and irregular boundaries, particularly around vehicles and vegetation. Such imperfections can weaken fine-detail supervision and complicate pixel-wise evaluation near discontinuities (Li et al., 2024c). However, the present experiments do not isolate their contribution to ONCE's errors or explain its performance gap to PRV2. In particular, the boundary evaluation combines segmentation and depth information; the quality of the displayed depth maps alone therefore does not explain the EdgeComp and F1 differences.

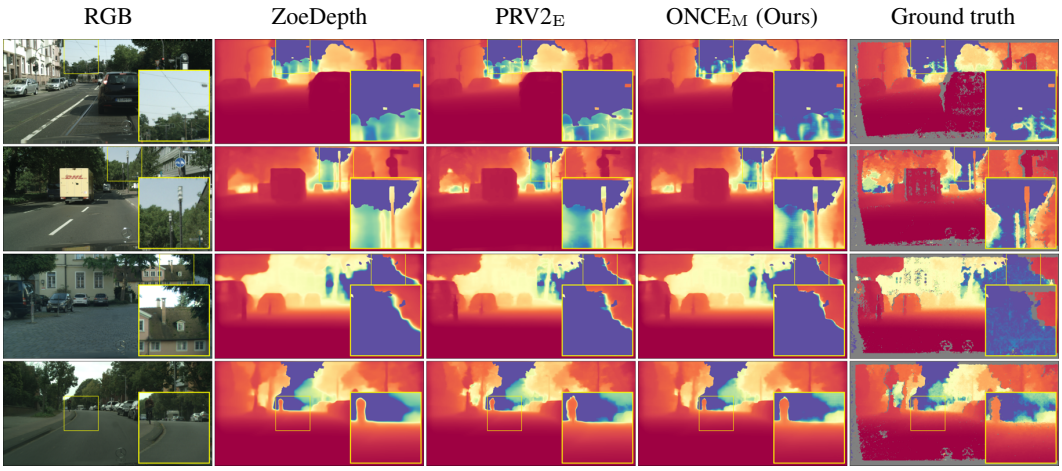

Figure 6: Qualitative results on Cityscapes. From left to right: RGB, coarse ZoeDepth, PRV2_{E}, ONCE_{M}, and ground-truth depth. PRV2_{E} uses SSIGM, and the checkpoints match Table 7. Yellow boxes mark regions shown in the enlarged insets.

AI USE DISCLOSURE

AI tools were used to identify additional references for the "Monocular and Metric Depth Estimation" subsection of Related Work and to check the grammar of the Introduction and Method sections. The authors reviewed the AI-assisted content and take full responsibility for the paper.