PatchRefiner: Leveraging Synthetic Data for Real-Domain High-Resolution Monocular Metric Depth Estimation

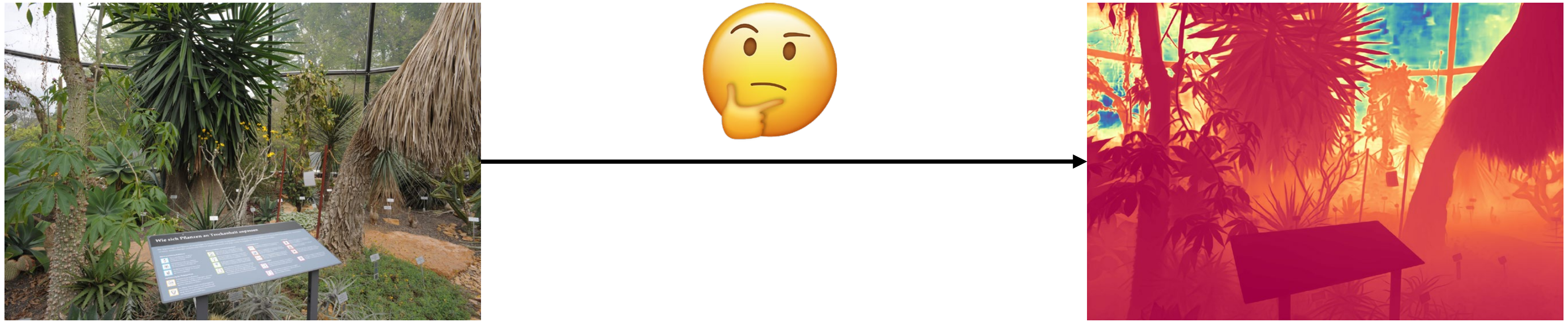
Zhenyu Li, Shariq Farooq Bhat, Peter Wonka

King Abdullah University of Science and Technology (KAUST)

King Abdullah University of Science and Technology

EUROPEAN CONFERENCE ON COMPUTER VISION

Goal

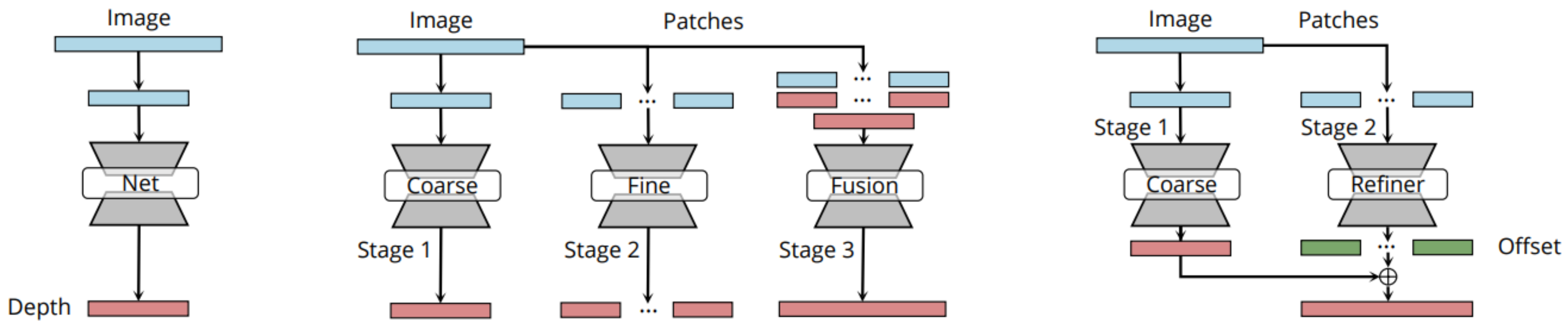


How to do ***high resolution*** metric depth estimation for real domain?

Two Contribution

- Better framework compared with PatchFusion
- High resolution model training strategy for real domain

Formulation



(a) LR Depth **(b)** Fusion-Based Framework **(c)** Refiner-Based Framework

Fig. 1: Framework Comparison. (a) Low resolution depth estimation framework with single forward pass. (b) Fusion-based high-resolution framework combining the best of coarse and fine depth predictions [30, 37]. (c) Our refiner-based framework predicts a residual to refine the coarse prediction.

Regard the high-resolution estimation as a refinement process

Framework

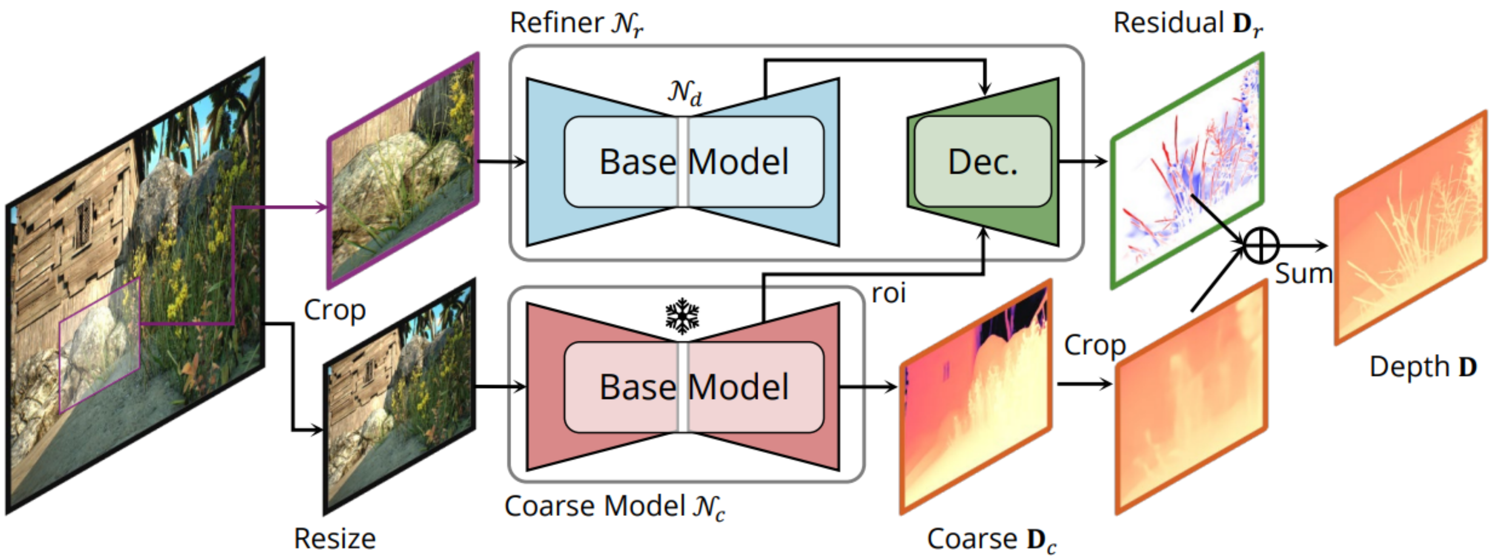


Fig. 2: Architecture Illustration. PatchRefiner contains a pre-trained frozen coarse depth estimation model $\mathcal{N}_c$ and a refiner model $\mathcal{N}_r$ predicts residual depth map $\mathcal{D}_r$ to refine the coarse depth $\mathcal{D}_c$. The refiner contains one base depth model $\mathcal{N}_d$ that has the same architecture as $\mathcal{N}_c$, and a light-weight decoder to aggregate information and make the final prediction.

Effectiveness

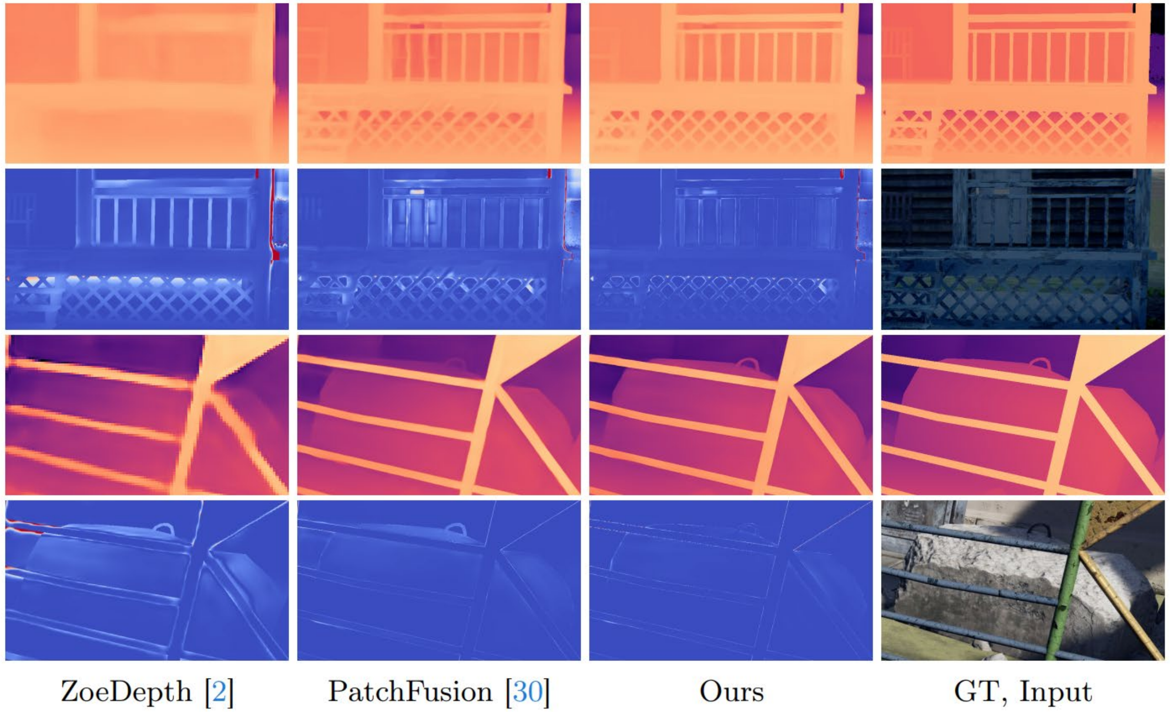


Fig. 5: Qualitative Comparison on UnrealStereo4K. We show the depth prediction and corresponding error map, respectively. The qualitative comparisons showcased here indicate our framework outperforms counterparts [2, 30] with sharper edges and lower error around boundaries. We show individual patches in all images to emphasize details near depth boundaries.

Real-domain challenge

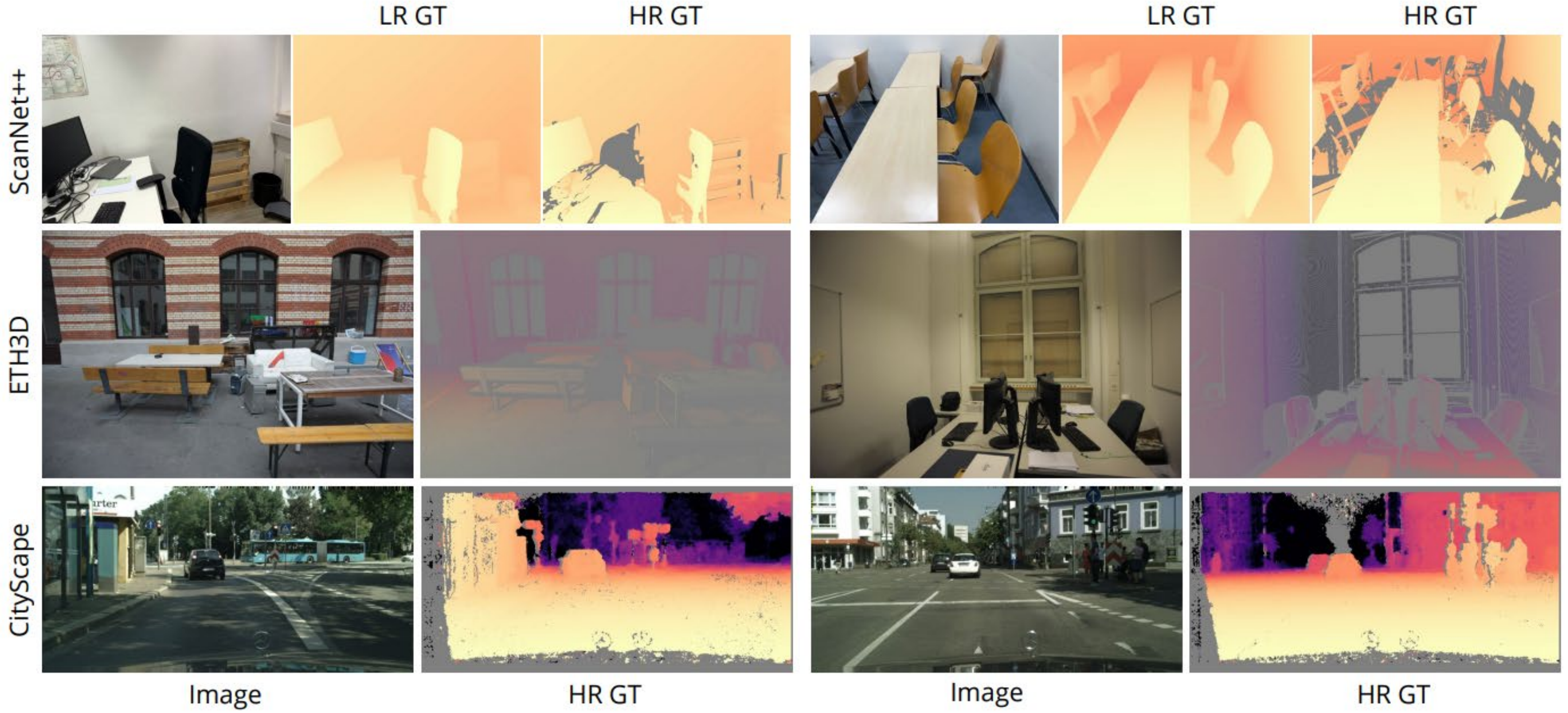


Fig. 3: Visualization of Real-Domain Data Pairs. Points lacking ground-truth data are depicted in gray. Due to sparse annotations near edges, models trained on real-domain data exhibit blurred boundary estimations.

King Abdullah University of Science and Technology

Real-domain solution

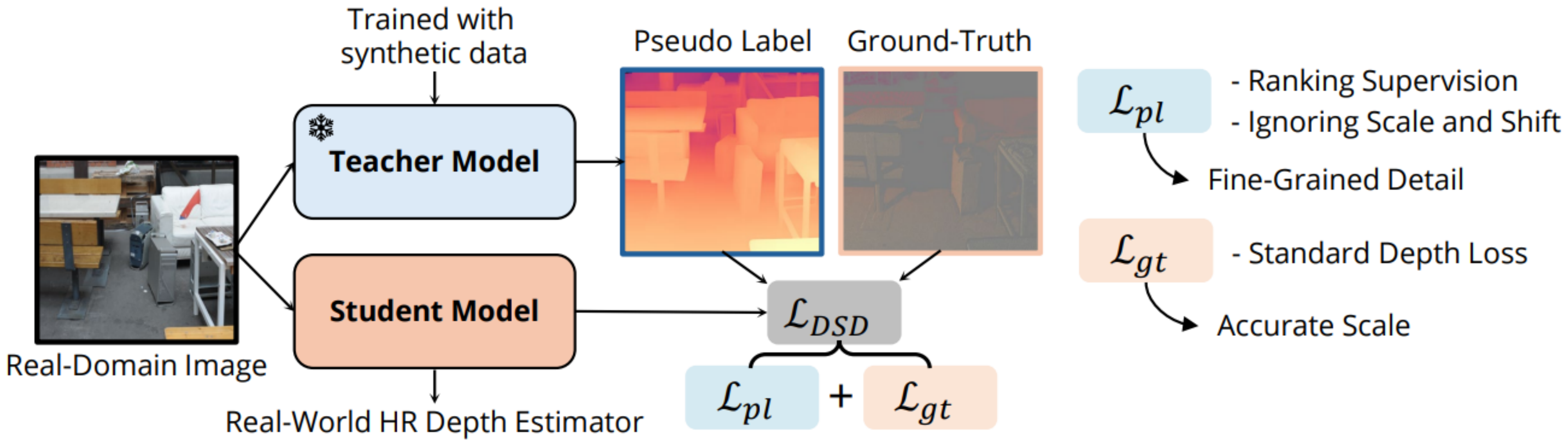


Fig. 4: Enhancing Real-Domain Learning with Synthetic Data. A teacher model trained on synthetic data produces pseudo labels for real-domain training. The student model benefits from a DSD dual-supervision approach: loss on pseudo labels for detail enhancement and loss on ground truth for scale accuracy. This method ensures detailed depth perception without compromising scale accuracy.

King Abdullah University of Science and Technology

Effectiveness

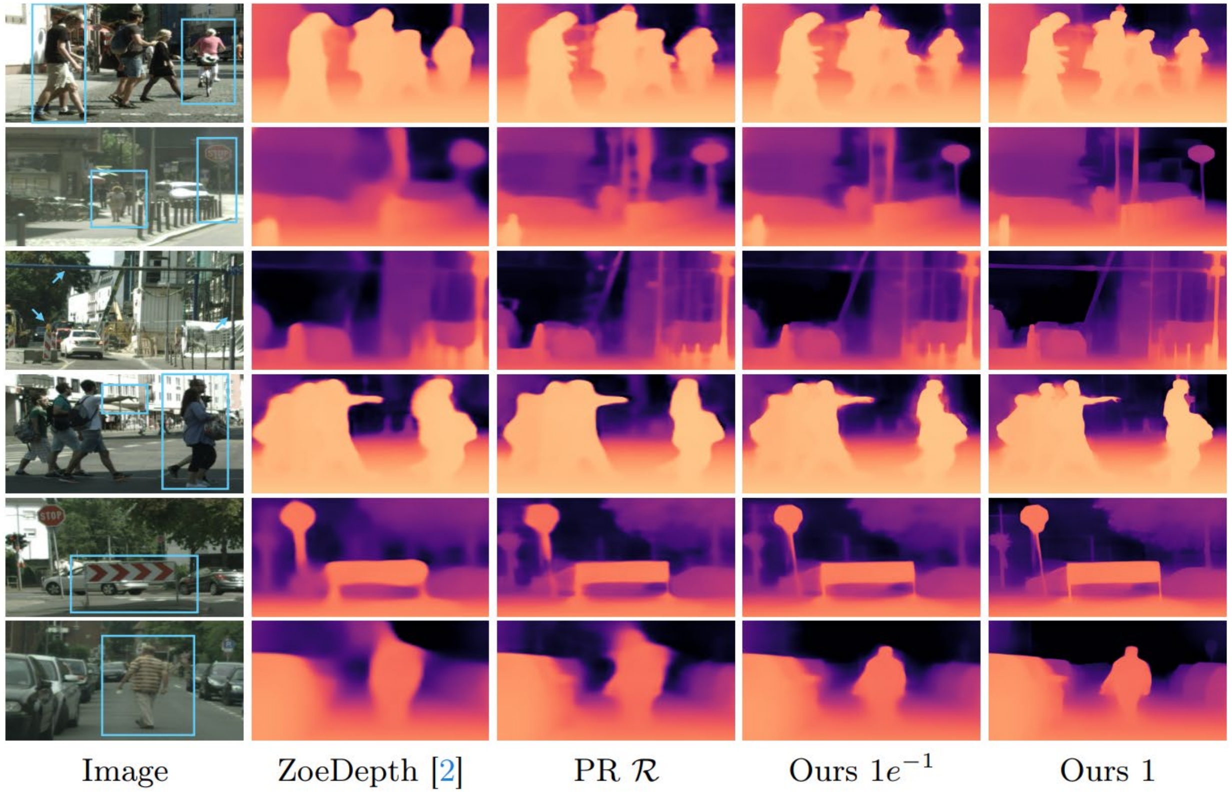


Fig. 7: Qualitative Comparison on CityScapes. This figure illustrates depth estimation comparisons between the base ZoeDepth model, PatchRefiner (PR) trained on CityScapes, and our method. We display outcomes under varying levels of $\mathcal{L}_{pl}$ supervision ($\lambda_1 = \lambda_2 = 1e^{-1}$ or 1), featuring zoomed-in sections of each image to highlight detail fidelity near depth discontinuities.

King Abdullah University of Science and Technology

EUROPEAN CONFERENCE ON COMPUTER VISION

Results

More Results and Interactive Images?

Check our paper: https://arxiv.org/pdf/2406.06679

Github: https://github.com/zhyever/PatchRefiner

Thank You

King Abdullah University
of Science and Technology