Multi-Resolution Monocular Depth Map Fusion by Self-Supervised Gradient-Based Composition

Yaqiao Dai*, Renjiao Yi*, Chenyang Zhu†, Hongjun He, Kai Xu†

National University of Defense Technology
chenyang.chandler.zhu, kevin.kai.xu@gmail.com

Abstract

Monocular depth estimation is a challenging problem on which deep neural networks have demonstrated great potential. However, depth maps predicted by existing deep models usually lack fine-grained details due to convolution operations and down-samplings in networks. We find that increasing input resolution is helpful to preserve more local details while the estimation at low resolution is more accurate globally. Therefore, we propose a novel depth map fusion module to combine the advantages of estimations with multi-resolution inputs. Instead of merging the low- and high-resolution estimations equally, we adopt the core idea of Poisson fusion, trying to implant the gradient domain of high-resolution depth into the low-resolution depth. While classic Poisson fusion requires a fusion mask as supervision, we propose a self-supervised framework based on guided image filtering. We demonstrate that this gradient-based composition performs much better at noisy immunity, compared with the state-of-the-art depth map fusion method. Our lightweight depth fusion is one-shot and runs in real-time, making it 80X faster than a state-of-the-art depth fusion method. Quantitative evaluations demonstrate that the proposed method can be integrated into many fully convolutional monocular depth estimation backbones with a significant performance boost, leading to state-of-the-art results of detail enhancement on depth maps. Codes are released at https://github.com/yuinsky/gradient-based-depth-map-fusion.

Introduction

Depth is an essential information in a wide range of 3D vision applications, bridging 2D images to 3D world. Prior methods of image-based depth estimation rely on multi-view geometry or other scene priors and constraints. In real-life scenarios, multi-view images or additional inputs are not always accessible and monocular depth estimation (depth from a single image) is the most common case. Estimating depth from a single image is a challenging and ill-posed problem, where deep learning demonstrated great potential, by which priors are automatically learnt from training data.

As many other vision tasks based on deep-learning, the predictions of networks are more blurry than the inputs, as shown in the second column "448" (denotes resolution of 448×448) in Fig. 2. Details are losing due to the down-sampling, max-pooling or convolution operations in the network. Furthermore, although fully convolutional networks take images at arbitrary resolutions as inputs, the networks are overfitted to the fixed resolution of training images. With a testing image at the training resolution, the estimated depth map is more accurate in values. Considering the memory and training time, networks for monocular depth estimation are usually trained in a relatively low resolution. In two most recent monocular depth estimation methods SGR (Xian et al. 2020) and LeRes (Yin et al. 2021), training images are at the resolution of 448×448. Testing images at training low-resolution lead to more accurate depth predictions than those at different resolutions, as demonstrated in Fig. 2. Testing images at original high-resolutions lead to inaccurate predictions of depth values, but keep much more local details. It motivates our explorations in multi-resolution depth map fusion, to combine the depth values and details predicted at different resolutions. It is studied that visual perceptions in natural images include various visual "levels" (Hubel and Wiesel 1962), while different visual "levels" should be considered at the same time to get an overall plausible result.

Figure 1: We propose a multi-resolution depth map fusion method to recover high-resolution depth maps. Compared with the single-resolution depth estimation network, the proposed method recovers wide levels of details.

*Co-first authors.
†Corresponding authors.

Instead of treating low- and high-resolution depth predictions equally by symmetry feature encoders, we adopt the core idea of Poisson fusion and propose a novel gradient-based composition network, fusing the gradient domain at high resolution into the depth map at low resolution.

Poisson fusion is not fully differentiable and requires an additional manually labeled mask as input. In order to achieve an end-to-end automatic and differentiable pipeline, we propose a network conceptually inspired by Poisson fusion, without requiring a fusion mask. Based on the observation of higher value accuracy of low-resolution depth, and better texture details of high-resolution depth, we fuse the values of low-resolution depth and gradients of higher-resolution depth by guided filter (He, Sun, and Tang 2010), which is a gradient-preserving filter, to get rid of the requirement of fusion masks. In guided filters, there are two parameters, the window size r is set to adjust the receptive field, along with an edge threshold ϵ.

Here, we fix both parameters for the whole dataset and select high quality data as training set to get a reasonable guided filtered depth map. This depth map is used as the supervision of the gradient domain. By adopting the guided filter during training, our model learn to preserve the detail automatically without the help of guided filters while testing.

In details, the self-supervision is constrained by two separate losses, an image level normalized regression loss (ILNR loss) at the depth domain between the low-resolution depth map and the network prediction, and a novel ranking loss at the gradient domain between guided filtered depth map and the network prediction, as illustrated in Fig. 3. By this self-supervised framework, no labeled data is needed in training.

With most fully convolutional monocular depth estimation methods as backbone, our method effectively enhances their depth maps as in Fig. 1. Details are very well recovered, with the original depth accuracy preserved. Our detailed pipeline is described in Fig. 3. In the depth map fusion network (the gradient-based composition module in Fig. 3), we firstly use a 1-layer convolution to get the approximate gradient map of the depth of higher resolution, then the depth map of lower resolution and the approximated gradient map of high-resolution input are fused in each layer of the 10-layer network to reconstruct the fused depth map. The network structure is designed specifically for a gradient-based composition inspired by Poisson fusion.

Miangoleh et al. (Miangoleh et al. 2021) describe similar observations and adopt GAN to merge low- and high-resolution depth maps of selected patches. Their method (BMD) effectively enhance the details in final depth maps but image noises would confuse their method to decide the high- and low-frequency patches. BMD is also time-consuming with the iterative fashion. Our solutions for multi-resolution depth fusion has a good robustness to image noises, runs in real-time, and fully self-supervised in training. In experiments, we demonstrated that the performance of our method is more stable for different levels of noises, while the state-of-the-art depth map fusion method BMD (Miangoleh et al. 2021) degenerates when the noise variance is high. Our method benefits from the one-shot depth fusion by the lightweight network design, running at 5.4 fps while BMD (Miangoleh et al. 2021) running at 0.07 fps in the same environment. Comprehensive evaluations and a large amount of ablations are conducted, proving the effectiveness of every critical modules in the proposed method.

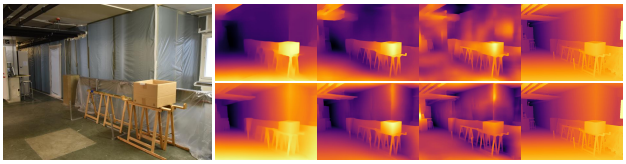

Figure 2: Different input sizes lead to different depth predictions. Low-resolution inputs recover more accurate depth values while higher-resolution inputs lead to more details. The numbers denote the input sizes to the network after resizing. GT denotes ground truth depth maps. The top and bottom rows show depth maps predicted by SGR (Xian et al. 2020) and LeRes (Yin et al. 2021) respectively.

Our contributions are summarized as follows:

- A portable network module is proposed to improve fully convolutional monocular depth estimation networks through a multi-resolution gradient-based fusion approach. Our method take advantages of the depth predictions of different resolution inputs, preserving the details while maintaining the overall accuracy.
- A self-supervised framework is introduced to find the optimal fused depth prediction. No labeled data is required.
- The method has good robustness to various image noises, and runs in real-time, while state-of-the-art depth map fusion method degenerates significantly with noises increasing, and takes seconds for each data.

Related Work

Monocular depth estimation is an essential step for many 3D vision applications, such as autonomous driving and SLAM (Bailey and Durrant-Whyte 2006) systems. Estimating depth from one single image is a challenging ill-posed problem, while traditional methods usually require multiple images to explore depth cues. For example, structure from motion (Levinson et al. 2011) is based on the feature correspondences among multi-view images. With only one image, it is infeasible to solve the ambiguities.

For ill-posed problems, deep neural networks show a good superiority. Deep-learning based methods can be categorized by supervision styles. A most straight-forward solution is supervised learning. Eigen et al. (2014) proposes the first supervised work to solve monocular depth estimation, by defining Euclidean losses between predictions and ground truths to train a two-component network (global and local network). Mayer et al. (2016) solves scene flow in a supervised manner. Monocular depth estimation, along with optical flow estimation, are working as sub-problems of scene flow estimation. Recently, pretrained layers in ResNet (He

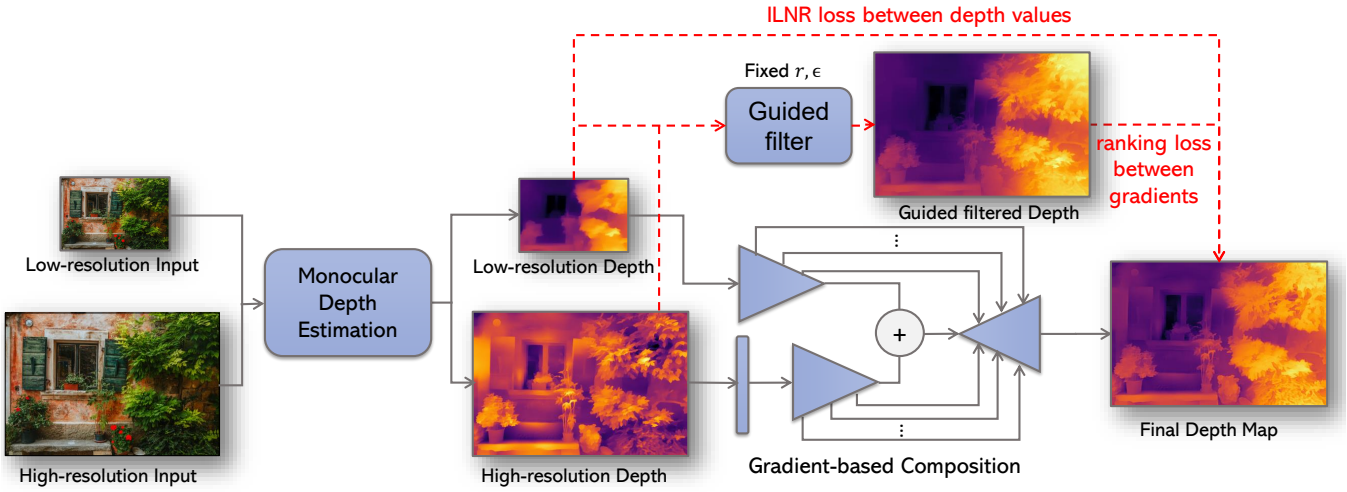

Figure 3: Pipeline of the self-supervised multi-resolution depth map fusion method. With input images at low- and high-resolution, depth maps in two resolutions are predicted by the backbone monocular depth estimation network. The proposed gradient-based composition fuses two depth maps into a plausible one, training in a self-supervised fashion by a ILNR loss in the depth domain and a ranking loss in the gradients domain, supervised by the low-resolution depth map and the guided filtered depth map respectively. The end-to-end pipeline can be integrated with most fully convolutional backbone networks. Red dashed lines denote the procedures included in training phase only.

et al. 2016) are widely used in monocular depth estimation networks (Xian et al. 2018; Ranftl et al. 2019; Yin et al. 2021) to speed up the training. Semi-supervised methods (Smolyanskiy, Kamenev, and Birchfield 2018; Kuznietsov, Stuckler, and Leibe 2017; Amiri, Loo, and Zhang 2019) training from stereo pairs are proposed to soften the requirement of direct supervision. They estimate disparity between two stereo images, and define a consistency between one input image, and the re-rendered image by estimated inverse depth, disparity, camera pose and the other input. Kuznietsov et al. (2017) use sparse supervision from LIDAR data and incorporate with berHu norm (Zwald and Lambert-Lacroix 2012). Following works based on LIDAR data (He et al. 2018; Wu et al. 2019) propose similar semi-supervised training pipelines. Unsupervised methods (Godard et al. 2019; Casser et al. 2019; Bozorgtabar et al. 2019; Zhu et al. 2018) are mostly relying on the constraint of reprojections between neighboring frames in their training image sequences. By getting rid of supervisions, these methods suffer from many problems such as scale ambiguities and scale inconsistencies. Other than different supervisions, various losses or constraints are proposed to better constrain the problem, such as Berhu loss (Heise et al. 2013; Zhang et al. 2018), conditional random fields (Li et al. 2015; Liu et al. 2015; Wang et al. 2015; Yuan et al. 2022), transformers (Ranftl, Bochkovskiy, and Koltun 2021; Yuan et al. 2022) or generative adversarial networks (GAN) (Feng and Gu 2019; Jung et al. 2017; Gwn Lore et al. 2018).

A common issue of deep-learning methods is the depth details lost in network outputs. Depth predictions are blurry with inaccurate details at object boundaries. This issue exists in many vision problems, such as image segmentation. Deep-learning methods generate more blurry results than non-CNN methods. Although replacing input images with higher resolutions generate predictions with higher resolutions, it leads to inaccurate depth values, as shown in Fig. 2. It motivates our work to fuse depth maps of different resolutions to get an overall plausible one. Traditional image fusion methods such as Poisson fusion (Pérez, Gangnet, and Blake 2003), Alpha blending (Porter and Duff 1984) require additional inputs (alpha weights or masks), requiring manual labeling. The proposed method automatically decides which regions from two-resolution depth maps have to be fused and how to fuse them. A recent depth map fusion method (Miangoleh et al. 2021) uses GAN to fuse the low- and high-resolution depths. Our method solves two-resolution depth fusion by self-supervised gradient-domain composition, achieving better robustness on image noises and real-time performance.

Method

Overview

Our key observation is that the local details are preserved in the gradient domain of depth estimation from a high-resolution input, while the global value accuracy is better estimated with low-resolution input. In other words, a convolutional neural network (CNN) can focus on different levels of details when dealing with input images of different resolutions. Therefore, fusing the predictions of multi-resolution inputs is a straightforward choice to enhance the depth estimation. The goal of our method is finding the optimal fusion operation $\oplus$ for d_{high} and d_{low}, which are depth predictions of the input image I at high- and low-resolutions:

$$f = d_{\text{high}} \oplus d_{\text{low}}, \tag{1}$$

where f is the fused depth estimation with enhanced details.

Inspired by Poisson fusion, the fusion operation $\oplus$ should transplant the gradient domain of d_{high} to d_{low} for detail-preserving. Thus, we formulate the optimization as,

$$\min_{\oplus} \iint_{\Omega} |\nabla f - \nabla d_{\text{high}}| \partial\Omega + \iint_{I-\Omega} |f - d_{\text{low}}| \partial\Omega, \quad (2)$$

where ∇ denotes the gradients of an image. The optimization of $\oplus$ only focus on the gradient domain within Ω and value domain among other areas $I - \Omega$, where Ω is the area d_{high} has better details than d_{low}. We propose a self-supervised neural network to find the fusion operation $\oplus$ based on Eq. 2.

Note that the classic Poisson fusion is not differentiable while calculating the fused gradient around boundaries of fusion area Ω. We first introduce a multi-level gradient-based fusion network module to approximate the Poisson fusion. Since we have no supervision of Ω to train this fusion module, we propose a self-supervised framework based on the supervision of guided filtering with a novel training loss. Therefore, the fusion module in our pipeline is fully differentiable and capable of preserving gradient details of proper fusion area Ω while maintaining overall consistency and the training is fully self-supervised. At last, the evaluations demonstrates the superiority of the method on depth estimation accuracy and detail preservation over the state-of-the-art alternatives with better efficiency, and robustness to image noises and complicated textures.

Multi-Level Gradient-Based Depth Fusion

Monocular depth estimation. Our multi-level gradient-based depth fusion requires monocular depth estimation with different resolutions as inputs. LeRes (Yin et al. 2021) is a novel and state-of-the-art monocular depth estimation network that can provide good depth initials. We adopt LeRes as the backbone to produce the depth initialization with different resolutions. For each input I, we adopt two different resolutions $I_{\text{low}}, I_{\text{high}}$, and the corresponding predictions are $d_{\text{low}}, d_{\text{high}}$ respectively. Our method also work with other monocular depth estimation backbones. The evaluations with other three monocular single-resolution backbones (Xian et al. 2020; Yuan et al. 2022; Ranftl, Bochkovskiy, and Koltun 2021) are in Section Experiments.

Differentiable gradient-domain composition. Poisson fusion could be a good candidate for constructing the depth fusion module since it takes the gradient consistency into optimization. To ensure that the whole framework of our method can be trained in an end-to-end manner, we need to find a differentiable approximation to deal with the truncation of gradient backward around the merging boundaries.

To avoid the gradient truncation, we adopt an encoder-decoder framework to formulate the fusion module which can utilize Ω implicitly in the latent space. Then we can rewrite Eq. 1 as below:

$$f = \mathcal{D}(\mathcal{E}_l(d_{\text{low}}) + \mathcal{E}_h(d_{\text{high}}), \Omega), \quad (3)$$

where $\mathcal{D}$ is the decoder module, and $\mathcal{E}_l$, $\mathcal{E}_h$ are the encoder modules for low- and high-resolution depth estimations.

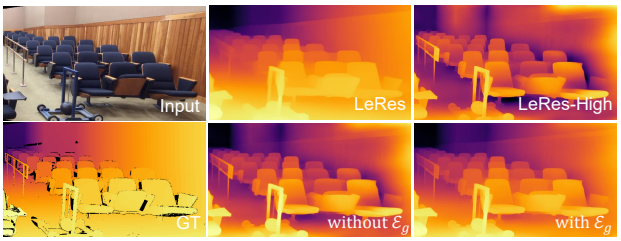

Figure 4: Visual comparison upon $\mathcal{E}_g$. LeRes and LeRes-High are depth estimation results given by LeRes(Yin et al. 2021) with low- and high resolution inputs.

However, this formulation has two problems in implementation. First, Ω requires supervision but there are no such datasets. Thus, a self-supervised framework is proposed to solve this problem in the next subsection. Second, training an encoder-decoder framework is highly inefficient if $\mathcal{E}_l$ and $\mathcal{E}_h$ do not share hyperparameters, however, sharing all hyperparameters will degrade the performance. Since we want to extract gradients from d_{high} and the depth values from d_{low}, it is natural to formulate $\mathcal{E}_h$ based on $\mathcal{E}_l$ with an additional one-level convolution layer $\mathcal{E}_g$ to extract gradients.

$$\mathcal{E}_h(\cdot) = \mathcal{E}_l(\mathcal{E}_g(\cdot)), \quad (4)$$

$\mathcal{E}_g(d_{\text{high}})$ can be considered as a varied approximation with tunable parameters of ∇d_{high}. Eq. 3 can be rewritten as,

$$f = \mathcal{D}(\mathcal{E}_l(d_{\text{low}}) + \mathcal{E}_l(\mathcal{E}_g(d_{\text{high}})), \Omega). \quad (5)$$

This solution is simple yet effective. Sharing most hyperparameters between $\mathcal{E}_l$ and $\mathcal{E}_h$ is highly efficient in the training stage. Our evaluation also demonstrates that the fusion performance will drop significantly if $\mathcal{E}_g$ is absent in the Eq. 5, which means this design is critical and essential. A visual comparison upon $\mathcal{E}_g$ is presented in Fig. 4 which support our claimant as well.

Multi-level fusion framework. To fully utilize d_{low} and d_{high} with neural networks, a simple one-step fusion with Eq. 5 is not enough. A pyramid-style framework is introduced to fuse depth at different resolutions. We formulate Eq. 5 with the multi-level encoder $\mathcal{E}_l^{2^i}$ as below:

$$f = \mathcal{D}(\sum_{i=2}^{11} \mathcal{E}_l^{2^i}(d_{\text{low}}) + \mathcal{E}_l^{2^i}(\mathcal{E}_g(d_{\text{high}})), \Omega) \quad (6)$$

where $\mathcal{E}_l^{2^i}$ is multi-layer fully convolution module with resolution 2^i. For implementation of Eq. 6, we take d_{low} and $\mathcal{E}_g(d_{\text{high}})$ into a multi-level convolution-based encoder individually. The convoluted outputs of d_{low} and $\mathcal{E}_g(d_{\text{high}})$ from each level are then skip connected and be supplied as input for layered upsampling and convolution modules to reconstruct the final depth map.

Self-Supervised Framework of Depth Fusion

Self-supervision with guided filtering. Image fusion based on Poisson equations introduces us to an interesting idea to fuse predictions of different resolution images. However, the

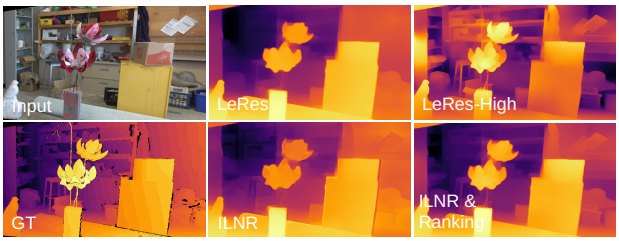

Figure 5: Visual comparison upon l_{mILNR} and l_{rank}. LeRes and LeRes-High are depth results by LeRes(Yin et al. 2021) with low- and high-resolution inputs. ILNR is the fused result training only by l_{mILNR} while ours adopts both.

classic Poissons-based fusion requires the manual labeling of the fusion area Ω. To get rid of the manual steps we still need a proper Ω for depth fusion, training under the supervision of existing datasets is the most straightforward way. Unfortunately, no datasets are available for providing such information. We introduce a self-supervision mechanism driven by guided filtering to deal with this problem.

The guided filtering is an edge-preserving filtering, it has the gradient smoothing property. This filter fuses d_{low} and d_{high} together while keeping the gradient details of d_{high} without manual mask labels like Ω. This character makes it a perfect supervision for training our network. However, the guided filtering requires additional parameters to control fusion quality, and the tunable parameters are data dependent. Fortunately, we find that in the gradient domain of the fused image, a preset parameter works for the training set. Therefore, we can rewrite Eq. 2 as below,

$$\min_{\oplus} \iint |\nabla f - \nabla d_{\text{gf}}| + |f - d_{\text{low}}| \partial\Omega, \tag{7}$$

where d_{gf} is the fused result of d_{low} and d_{high} through guided filtering with a set of fixed parameters. The ablations in Section demonstrate that d_{gf} provide effective supervision for training $\oplus$. Our fused module outperforms guided filtering as well in evaluations.

Self-supervised loss. The training objective Eq. 7 is then optimized through the self-supervised loss l_{fusion} as below,

$$l_{\text{fusion}}(f, d_{\text{gf}}, d_{\text{low}}) = l_{\text{mILNR}}(f, d_{\text{low}}) + l_{\text{rank}}(f, d_{\text{gf}}), \tag{8}$$

where $l_{\text{mILNR}} = \sum_r l_{\text{ILNR}}(f^r, d^r_{\text{low}})$ is multi-resolution image-level normalized regression (ILNR) loss (Yin et al. 2021) which constrains the value domain of fused result f to be similar as d_{low} at every resolution r levels. Another term l_{rank} is a novel ranking loss inspired by (Xian et al. 2020) which constrains the gradient domain of f being close to d_{gf}. Given a pair of points i, j, $p^i_f \in f, p^i_{\text{gf}} \in d_{\text{gf}}$ are pixels on f and d_{gf} respectively. $l_{\text{rank}} = \frac{1}{N}\sum_{i,j} E(p^i_f, p^j_f, p^i_{\text{gf}}, p^j_{\text{gf}})$ is formulated based on these sampled pixel pairs, and N is the number of pixel pairs. Point pair i, j are sampled based on edge areas M extracted from ∇d_{gf} and ∇d_{high} by Canny.

$$E(p^i_f, p^j_f, p^i_{\text{gf}}, p^j_{\text{gf}}) = \begin{cases} \log(1 + e^{-\frac{1}{|(p^i_f - p^j_f) - (p^i_{\text{gf}} - p^j_{\text{gf}}) + \sigma|}}), & z_{ij} = 1, \\ |p^i_f - p^j_f|^2, & z_{ij} = 0, \end{cases} \tag{9}$$

where z_{ij} is an indicator, $z_{ij} = 1$ means pixel i, j are located on different sides of an edge of M while $z_{ij} = 0$ means they located on the same side. σ is a regular term for robust compuatation. A visual comparison upon our loss terms l_{mILNR} and l_{rank} is presented in Fig. 5, which demonstrate the our design of losses is effective.

Experiments

In this section, we first introduce the benchmark datasets and metrics. Next, we compare with several state-of-the-art monocular depth estimation and refinement methods in aspects of several error metrics, robustness to noises, and running time. Lastly, we conduct several ablations to study the effectiveness of critical designs in the pipeline.

Benchmark Datasets and Evaluation Metrics

To evaluate the depth estimation ability of our method, we adopt several commonly used zero-shot datasets, which are Multiscopic (Yuan et al. 2021), Middlebury2021 (Scharstein et al. 2014) and Hypersim (Roberts et al. 2021). All benchmark datasets are unseen during training.

We test on the whole set of Middlebury2021 (Scharstein et al. 2014), including 24 real scenes. For Multiscopic (Yuan et al. 2021), we evaluate on synthetic test data, containing 100 high resolution indoor scenes. For Hypersim (Roberts et al. 2021), we evaluate on three subsets of 286 tone-mapped images generated by the released codes. Several error metrics are adopted. $SqRel$ and rms are the square relative error and the root mean square error, respectively. The mean absolute logarithmic error log_{10} is defined as $log_{10} = \frac{1}{N}\sum \|log(d^*_i) - log(d_i)\|$. The error metric δ_k describes the percentage of pixels satisfying $\delta = max(\frac{d^*_i}{d_i}, \frac{d_i}{d^*_i}) < 1.25^k$. D^3R measures the detected edge error, which is introduced in (Miangoleh et al. 2021). Here, d^*_i is the ground truth depth value and d_i denote the predict value at pixel i. Due to the scale ambiguity in monocular depth estimation, we follow Ranftl et al. (2019) to align the scale and shift using the least squares before computing errors.

Comparisons to State-of-the-Arts

Quantitative evaluation. We compare our method with two state-of-the-art fusion based monocular depth estimation alternatives, BMD (Miangoleh et al. 2021), and 3DK (Niklaus et al. 2019), in Tab. 1. To demonstrate that our self-supervised framework is effective, we also present the performance of monocular depth estimation methods applying with guided filter (He, Sun, and Tang 2010) as a baseline. In the comparison, we adopt the trained models released by the authors and evaluate with their default configuration.

Tab. 1 demonstrate that our method outperform all the other fusion based depth estimation alternatives on most benchmarks and error metrics. Our method also produce better fusion results than guided filtering which is our training

Methods	Multiscopic				Middlebury2021				Hypersim			
	SqRel↓	rms↓	log10↓	δ_1 ↑	SqRel↓	rms↓	log10↓	δ_1 ↑	SqRel↓	rms↓	log10↓	δ_1 ↑
SGR	9.161	14.031	0.086	0.745	0.846	3.948	0.067	0.773	0.593	1.536	0.102	0.612
NeWCRFs	11.031	14.658	0.088	0.749	0.829	3.724	0.058	0.830	0.513	1.322	0.088	0.694
DPT	4.021	9.781	0.059	**0.841**	0.700	3.698	0.060	0.827	0.327	1.145	0.083	0.734
LeRes	9.168	13.12	0.082	0.776	0.464	3.042	0.052	0.847	0.319	1.011	**0.071**	0.768
SGR-GF	9.314	14.107	0.087	0.743	0.844	3.9	0.067	0.773	0.600	1.539	0.103	0.612
NeWCRFs-GF	10.601	14.407	0.087	0.751	0.796	3.549	0.058	0.832	0.518	1.320	0.088	0.694
DPT-GF	4.142	10.060	0.060	0.835	0.685	3.653	0.060	0.825	0.332	1.149	0.083	0.733
LeRes-GF	9.01	13.063	0.082	0.776	0.457	2.976	0.052	0.849	0.324	1.011	0.072	0.769
3DK	9.379	14.879	0.077	0.73	0.911	4.18	0.069	0.745	0.718	1.521	0.108	0.610
LeRes-BMD	9.259	13.101	0.083	0.773	0.487	3.014	0.055	0.844	0.312	**0.993**	0.072	0.769
SGR-Ours	9.144	14.0	0.086	0.746	0.816	3.858	0.067	0.776	0.605	1.549	0.103	0.609
NeWCRFs-Ours	10.405	14.299	0.087	0.752	0.786	3.587	0.057	0.832	0.520	1.324	0.089	0.687
DPT-Ours	**3.998**	**9.759**	**0.058**	**0.841**	0.647	3.533	0.059	0.832	**0.311**	1.109	0.081	0.745
LeRes-Ours	8.833	12.921	0.081	0.781	**0.444**	**2.963**	**0.051**	**0.853**	0.315	0.999	**0.071**	**0.77**

Table 1: The quantitative evaluations on three benchmark datasets. Bold numbers denote the best result while underlined numbers are second best. Depth map fusion methods on monocular depth estimation backbones are presented as "backbone-fusion method". Our method achieve the best performance on $11/12$ metrics in these 3 datasets.

Method	Multiscopic	Middlebury2021
SGR	0.576	0.735
DPT	0.594	0.613
NeWCRFs	0.767	0.737
LeRes	0.570	0.719
Ours	**0.542**	**0.589**

Table 2: Evaluation of our fusion module on edge correctness of depth by D^3R metric, where lower values are better.

supervision. It means that our network takes advantages of both the low-resolution input and the guided filtered fusion result but not strictly constrained. It should be noticed that our method was trained with the monocular depth estimation backbone LeRes (Yin et al. 2021). The trained fusion module can be directly integrated with other fully convolutional monocular depth estimation networks such as SGR (Xian et al. 2020), NeWCRFs (Yuan et al. 2022) and DPT (Ranftl, Bochkovskiy, and Koltun 2021), without any fine-tuning. The proposed method is portable and can be easily incorporated into many state-of-the-art depth estimation models.

We also evaluate the details recovered by our method. We present the comparison between our method and two backbone methods with D^3R metric in Tab. 2, which measure the edge correctness of the estimation depth details.

Visual comparisons. The qualitative evaluations are presented in Fig. 6. As discussed earlier, most monocular depth estimation methods suffer from blurry predictions. The improvement of detail-preserving by our multi-resolution depth map fusion is significant. Most details missing in the backbone methods are successfully recovered, while keeping the original depth values correct.

Anti-noise evaluation. One of the main advantages of our method over the state-of-the-art depth fusion method BMD (Miangoleh et al. 2021) is the noise robustness enabled by our gradient-based fusion. Since BMD determines fusion areas explicitly by edge detection, its performance will drop significantly while the input images include noises.

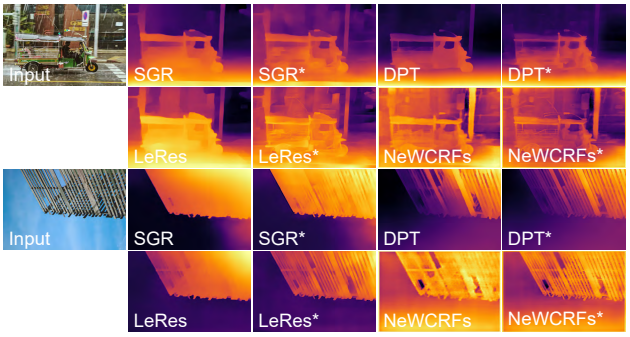

Figure 6: Qualitative comparisons on unseen natural images from the Internet. Our method (*) successfully boosts the performances of the backbone monocular depth estimation methods, and recovers the details in depth maps.

We compare the anti-noise ability of our methods with BMD. For evaluation, we add Gaussian noises or Pepper noises to the input image. The mean value of Gaussian noises is 0, the variances changes from 0.001 to 0.009 for Gaussian noises, and the signal-noise ratio changes from 100% to 95% for Pepper noises. Fig. 8 shows the variety of δ_1 value when adopting different levels of Gaussian (left) or Pepper noises (right).

Fusion based methods such as BMD can also be easily influenced by the complicated textures. We split Middlebury2021 benchmark into two sub-sets based on the edge number detected in the ground truth depth images. Less edges on depth maps means more depth independent textures exist. The plot at the left of Fig. 7 demonstrates that our method outperforms BMD more significantly on the difficult sub-set. At the right of Fig. 7, a visual comparison on a difficult example is shown, where BMD suffers from the complicated textures and extract many texture details into the depth map (e.g. paintings on the whiteboard).

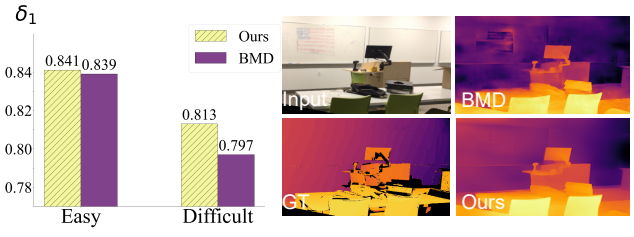

Figure 7: Left: Quantitative evaluation between our method and BMD on easy and difficult datasets regarding to texture complexities. Ours outperform BMD more significantly on the difficult subset. Right: Visual comparisons with BMD.

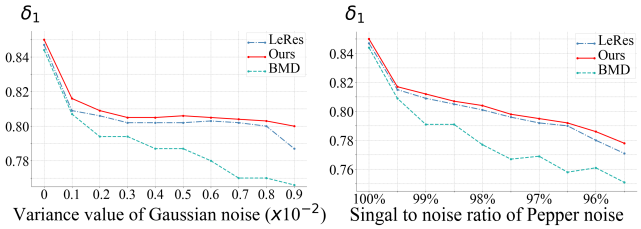

Figure 8: Quantitative evaluation of anti-noise ability of LeRes, BMD and ours. The vertical axis denotes the value of δ_1, which is the higher the better.

Running time. Our fusion module is one-shot and requires no additional complex processing, it is highly efficient and only takes 68% processing time comparing with LeRes for depth estimation of high resolution input. The fusion processing of our method is 10X faster than guided filter while present better performance. The whole pipeline of our method, including the depth estimation time of low- and high-resolution inputs, is more than 80X faster than BMD.

Ablation Studies

We evaluate the effectiveness of critical designs in our method. All ablation alternatives are trained for 30 epochs, under identical training configurations.

We first investigate the effects on model performance of different type of loss functions, including ILNR loss, gradient loss (Li and Snavely 2018) and ranking loss. We construct the ablation study with 7 different setting of training losses as shown in Tab. 3 on Middlebury2021. We adopt ILNR loss to supervise the value domain in our fusion network comparing with low resolution result and guided fused result, and adopt gradient loss(Li and Snavely 2018), original ranking loss (Xian et al. 2020) or the proposed ranking loss to supervise the gradient domain comparing with high resolution depth and guided fused result. Tab. 3 shows that our configuration outperform other alternatives on D^3R metric. We also evaluate the validation of our differential gradient-domain composition design. We compare the depth estimation performance with and without $\mathcal{E}_g$. The results in Tab. 4 demonstrate that this simple design is critical and can significantly improve the detail preserving performance. The visual comparison is presented in Fig. 4.

Loss functions				Result
ILNR	Gradient	SGR Ranking	Our Ranking	Error
Low-res	High-res	×	×	0.734
Low-res	×	High-res	×	0.722
Low-res	×	×	High-res	0.688
Guided	Guided	×	×	0.723
Guided	×	Guided	×	0.714
Guided	×	×	Guided	0.711
Low-res	×	×	Guided	**0.684**

Table 3: Ablation study on different training loss settings. The last row is the setting of our method.

Method	SqRel↓	rms↓	log10↓	δ_1 ↑	δ_2 ↑
w/	0.468	2.977	0.051	0.833	0.952
w/o	0.791	3.811	0.062	0.799	0.952

Table 4: Comparison of our architecture with and without the first convolution layer of high-resolution depth.

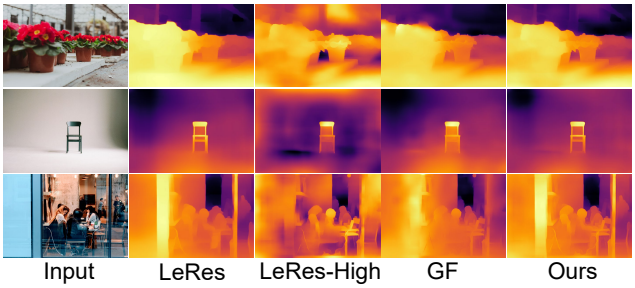

Figure 9: Three types of challenging cases, which are out-of-focus images, featureless regions, and transparent surfaces.

Limitations

Our method benefits from details in high-resolution images. If the resolutions of original images are low, the improvements by our method may be minor. Furthermore, we demonstrate several challenging cases in Fig. 9. The first one is out-of-focus regions, which are blurred, and the details cannot be recovered by high resolution depth map. The second one is featureless regions, such as the white floor and wall in the figure, where monocular depth estimation backbones cannot predict the accurate depth. The third case is transparent or reflective materials such as water and glasses. In the figure, the depth enhancement results contain many artifacts due to the glass wall in front of the scene.

Conclusion

We introduce a multi-resolution gradient-based depth map fusion pipeline to enhance the depth maps by monocular depth estimation backbones. Depth maps with a wide level of details are recovered, which are helpful for many following tasks such as 3D scene reconstruction. Comparing with prior works, the proposed method has a great robustness to image noises, and runs in real time. Comprehensive experiments are conducted, proving the effectiveness of every critical modules in the method.

Acknowledgements

We thank the anonymous reviewers for their valuable comments. This work is supported in part by the National Key Research and Development Program of China (2018AAA0102200), NSFC (62132021, 62002375, 62002376), Natural Science Foundation of Hunan Province of China (2021JJ40696, 2022RC1104) and NUDT Research Grants (ZK19-30, ZK22-52).

References

Amiri, A. J.; Loo, S. Y.; and Zhang, H. 2019. Semi-supervised monocular depth estimation with left-right consistency using deep neural network. In *2019 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, 602–607. IEEE.

Bailey, T.; and Durrant-Whyte, H. 2006. Simultaneous localization and mapping (SLAM): Part II. *IEEE robotics & automation magazine*, 13(3): 108–117.

Bozorgtabar, B.; Rad, M. S.; Mahapatra, D.; and Thiran, J.-P. 2019. Syndemo: Synergistic deep feature alignment for joint learning of depth and ego-motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4210–4219.

Casser, V.; Pirk, S.; Mahjourian, R.; and Angelova, A. 2019. Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 8001–8008.

Eigen, D.; Puhrsch, C.; and Fergus, R. 2014. Depth map prediction from a single image using a multi-scale deep network. *arXiv preprint arXiv:1406.2283*.

Feng, T.; and Gu, D. 2019. Sganvo: Unsupervised deep visual odometry and depth estimation with stacked generative adversarial networks. *IEEE Robotics and Automation Letters*, 4(4): 4431–4437.

Godard, C.; Mac Aodha, O.; Firman, M.; and Brostow, G. J. 2019. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3828–3838.

Gwn Lore, K.; Reddy, K.; Giering, M.; and Bernal, E. A. 2018. Generative adversarial networks for depth map estimation from RGB video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 1177–1185.

He, K.; Sun, J.; and Tang, X. 2010. Guided image filtering. In *European conference on computer vision*, 1–14. Springer.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

He, L.; Chen, C.; Zhang, T.; Zhu, H.; and Wan, S. 2018. Wearable depth camera: Monocular depth estimation via sparse optimization under weak supervision. *IEEE Access*, 6: 41337–41345.

Heise, P.; Klose, S.; Jensen, B.; and Knoll, A. 2013. Pm-huber: Patchmatch with huber regularization for stereo matching. In *Proceedings of the IEEE International Conference on Computer Vision*, 2360–2367.

Hubel, D. H.; and Wiesel, T. N. 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1): 106–154.

Jung, H.; Kim, Y.; Min, D.; Oh, C.; and Sohn, K. 2017. Depth prediction from a single image with conditional adversarial networks. In *2017 IEEE International Conference on Image Processing (ICIP)*, 1717–1721. IEEE.

Kuznietsov, Y.; Stuckler, J.; and Leibe, B. 2017. Semi-supervised deep learning for monocular depth map prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6647–6655.

Levinson, J.; Askeland, J.; Becker, J.; Dolson, J.; Held, D.; Kammel, S.; Kolter, J. Z.; Langer, D.; Pink, O.; Pratt, V.; et al. 2011. Towards fully autonomous driving: Systems and algorithms. In *2011 IEEE intelligent vehicles symposium (IV)*, 163–168. IEEE.

Li, B.; Shen, C.; Dai, Y.; Van Den Hengel, A.; and He, M. 2015. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1119–1127.

Li, Z.; and Snavely, N. 2018. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2041–2050.

Liu, F.; Shen, C.; Lin, G.; and Reid, I. 2015. Learning depth from single monocular images using deep convolutional neural fields. *IEEE transactions on pattern analysis and machine intelligence*, 38(10): 2024–2039.

Mayer, N.; Ilg, E.; Hausser, P.; Fischer, P.; Cremers, D.; Dosovitskiy, A.; and Brox, T. 2016. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4040–4048.

Miangoleh, S. M. H.; Dille, S.; Mai, L.; Paris, S.; and Aksoy, Y. 2021. 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*, 9685–9694.

Niklaus, S.; Mai, L.; Yang, J.; and Liu, F. 2019. 3d ken burns effect from a single image. *ACM Transactions on Graphics (TOG)*, 38(6): 1–15.

Pérez, P.; Gangnet, M.; and Blake, A. 2003. Poisson image editing. In *ACM SIGGRAPH 2003 Papers*, 313–318.

Porter, T.; and Duff, T. 1984. Compositing digital images. In *Proceedings of the 11th annual conference on Computer graphics and interactive techniques*, 253–259.

Ranftl, R.; Bochkovskiy, A.; and Koltun, V. 2021. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12179–12188.

Ranftl, R.; Lasinger, K.; Hafner, D.; Schindler, K.; and Koltun, V. 2019. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *arXiv preprint arXiv:1907.01341*.

Roberts, M.; Ramapuram, J.; Ranjan, A.; Kumar, A.; Bautista, M. A.; Paczan, N.; Webb, R.; and Susskind, J. M. 2021. Hypersim: A Photorealistic Synthetic Dataset for Holistic Indoor Scene Understanding. In *International Conference on Computer Vision (ICCV) 2021*.

Scharstein, D.; Hirschmüller, H.; Kitajima, Y.; Krathwohl, G.; Nešić, N.; Wang, X.; and Westling, P. 2014. High-resolution stereo datasets with subpixel-accurate ground truth. In *German conference on pattern recognition*, 31–42. Springer.

Smolyanskiy, N.; Kamenev, A.; and Birchfield, S. 2018. On the importance of stereo for accurate depth estimation: An efficient semi-supervised deep neural network approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 1007–1015.

Wang, P.; Shen, X.; Lin, Z.; Cohen, S.; Price, B.; and Yuille, A. L. 2015. Towards unified depth and semantic prediction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2800–2809.

Wu, Z.; Wu, X.; Zhang, X.; Wang, S.; and Ju, L. 2019. Spatial correspondence with generative adversarial network: Learning depth from monocular videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7494–7504.

Xian, K.; Shen, C.; Cao, Z.; Lu, H.; Xiao, Y.; Li, R.; and Luo, Z. 2018. Monocular relative depth perception with web stereo data supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 311–320.

Xian, K.; Zhang, J.; Wang, O.; Mai, L.; Lin, Z.; and Cao, Z. 2020. Structure-guided ranking loss for single image depth prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 611–620.

Yin, W.; Zhang, J.; Wang, O.; Niklaus, S.; Mai, L.; Chen, S.; and Shen, C. 2021. Learning to recover 3d scene shape from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 204–213.

Yuan, W.; Gu, X.; Dai, Z.; Zhu, S.; and Tan, P. 2022. NeWCRFs: Neural Window Fully-connected CRFs for Monocular Depth Estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Yuan, W.; Zhang, Y.; Wu, B.; Zhu, S.; Tan, P.; Wang, M. Y.; and Chen, Q. 2021. Stereo Matching by Self-supervision of Multiscopic Vision. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 5702–5709. IEEE.

Zhang, Z.; Cui, Z.; Xu, C.; Jie, Z.; Li, X.; and Yang, J. 2018. Joint task-recursive learning for semantic segmentation and depth estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 235–251.

Zhu, C.; Xu, K.; Chaudhuri, S.; Yi, R.; and Zhang, H. 2018. SCORES: Shape composition with recursive substructure priors. *ACM Transactions on Graphics (TOG)*, 37(6): 1–14.

Zwald, L.; and Lambert-Lacroix, S. 2012. The berhu penalty and the grouped effect. *arXiv preprint arXiv:1207.6868*.