JUSTC

Article

http://justc.ustc.edu.cn

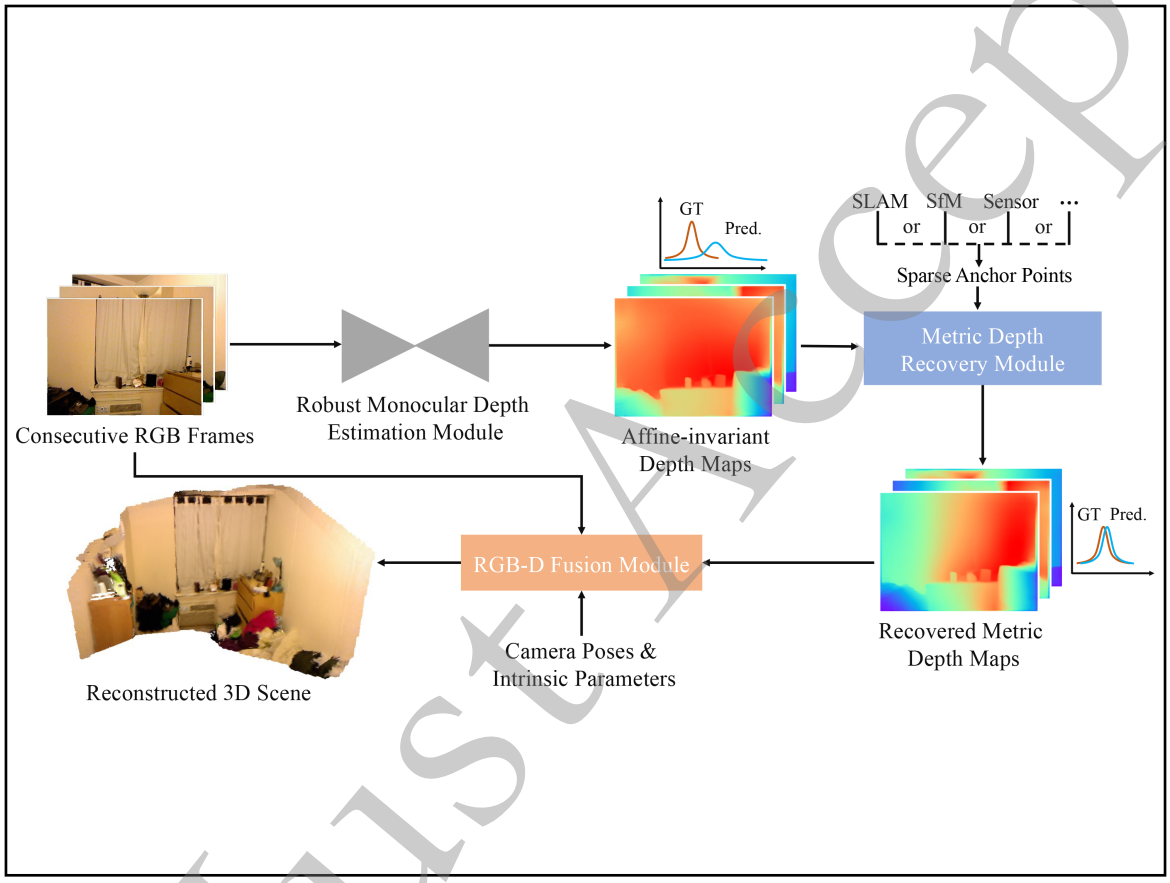
Towards 3D Scene Reconstruction from Locally Scale-Aligned Monocular Video Depth

Guangkai Xu[1], and Feng Zhao[1] ✉

National Engineering Laboratory for Brain-inspired Intelligence Technology and Application, University of Science and Technology of China, Hefei 230026, China

✉Correspondence: Feng Zhao, E-mail: fzhao956@ustc.edu.cn

Graphical abstract



Our pipeline for dense 3D scene reconstruction is composed of a robust monocular depth estimation module, a metric depth recovery module, and an RGB-D fusion module. With our robust depth model trained on enormous data and the proposed locally weighted linear regression method, we can achieve robust and accurate 3D scene shapes.

Public summary

■

■

■

Citation: Xu G K, Zhao F. Towards 3D Scene Reconstruction from Locally Scale-Aligned Monocular Video Depth. *JUSTC*, **2023**, 53(0): . DOI: 10.52396/JUSTC-2023-0061

JUSTC

Article

http://justc.ustc.edu.cn

Towards 3D Scene Reconstruction from Locally Scale-Aligned Monocular Video Depth

Guangkai Xu[1], and Feng Zhao[1] ✉

National Engineering Laboratory for Brain-inspired Intelligence Technology and Application, University of Science and Technology of China, Hefei 230026, China

✉Correspondence: Feng Zhao, E-mail: fzhao956@ustc.edu.cn

Cite This: *JUSTC*, **2023**, 53(X): (9pp) Read Online SI Supporting Information

Abstract: Monocular depth estimation methods have achieved excellent robustness on diverse scenes, usually by predicting affine-invariant depth, up to an unknown scale and shift, rather than metric depth in that it is much easier to collect large-scale affine-invariant depth training data. However, in some video-based scenarios such as video depth estimation and 3D scene reconstruction, the unknown scale and shift residing in per-frame prediction may cause the predicted depth to be inconsistent. To tackle this problem, we propose a locally weighted linear regression method to recover the scale and shift map with very sparse anchor points, which ensures the consistency along consecutive frames. Extensive experiments show that our method can drop the Rel error of existing state-of-the-art approaches by 50% at most over several zero-shot benchmarks. Besides, we merge 6.3 million RGBD images to train robust depth models. By locally recovering scale and shift, our produced ResNet50-backbone model even outperforms the state-of-the-art DPT ViT-Large model. Combined with geometry-based reconstruction methods, we formulate a new dense 3D scene reconstruction pipeline, which benefits from both the scale consistency of sparse points and the robustness of monocular methods. By performing simple per-frame prediction over a video, the accurate 3D scene geometry can be recovered.

Keywords: 3D scene reconstruction; monocular depth estimation; locally weighted linear regression

CLC number: **Document code:** A

1 Introduction

Dense monocular depth estimation[1–9] is a fundamental task in computer vision, which is important for a few downstream applications, such as autonomous driving[10, 11], virtual/augmented reality (VR/AR)[12, 13], 3D scene understanding[14–16] and reconstruction[17, 18]. Existing supervised approaches[1, 2, 19] and unsupervised methods[9, 20–22] have made tremendous progress in accuracy and robustness. To solve the generalization issue over diverse scenes, current state-of-the-art methods, such as MiDaS[3], LeReS[1], ominidata[19], and DPT[2], propose to merge a large-scale multi-domains data and predict an affine-invariant depth/inverse depth[7]. Although strong generalizability has been achieved, the predicted depth/inverse depth is up to an unknown scale and shift. However, in many video-based scenarios, such as autonomous driving and VR/AR, losing metric information significantly limits its application. The variant scale and shift in per-frame prediction will cause the depth inconsistency and the failure of accurate 3D reconstruction over consecutive frames.

How to ensure the depth consistency over consecutive frames is receiving increasing attention. Bian *et al.*[9, 20] employs an unsupervised paradigm and models the predicted depth prediction as scale-invariant. They propose a geometric consistency loss to implicitly learn the scale consistency over consecutive frames. Similarly, CVD[23] employs an unsupervised method but does the inference training to ensure consistency. By contrast, RCVD[24] takes the MiDaS model as the depth prior and estimates consistent dense depth maps and camera poses from a monocular video. They have achieved promising visual consistency of depth maps. However, we obverse that their reconstructed point clouds are still not so satisfactory. Please see the supplemental material for more detailed analyses. In this work, instead of the visual consistency, we focus on geometric consistency, *i.e.*, achieving the 3D scene reconstruction from consecutive frames.

Following previous methods, we enforce the model to predict an affine-invariant depth, thus recovering scale and shift for the prediction is the main barrier for 3D scene reconstruction over a monocular video. Existing methods[1–3] propose to directly compute a scale and shift value from least-squares fitting with the ground truth (GT), i.e. global recovery strategy. However, we observe that the optimal scale and shift are always heteroscedastic. During practical application, the global fitting does not consider the distribution difference between affine-invariant depth and ground-truth depth, and fails to effectively align the local regions. In Figure 2(a), we show an example error map visualization between the ground truth and such global scale-shift recovered depth, and there exists a low-frequency error clearly. Such a coarse alignment method cannot recover a high-quality metric depth for reconstruction.

Motivated by this observation, we propose a local recov-

DOI: 10.52396/JUSTC-2023-0061
JUSTC, **2023**, 53(X):

ery strategy to recover the locally aligned metric depth. Concretely, we employ the locally weighted linear regression as a novel metric depth alignment method, and demonstrate that heteroscedastic scale and shift maps can be recovered with as few as 25 points by enforcing spatial smoothness. Rather than fitting a unified scale map and shift map, *i.e.*, sharing the same scale and shift values over all coordinates of the image, our local recovery strategy can retrieve a location-related scale map and shift map to adjust the distribution of prediction. Experiments show that our method can significantly improve metric accuracy. Although some sparse anchor points are required, compared with existing state-of-the-art depth completion methods[25–28], which generally take more than hundreds to thousands of sparse points, our strategy requires much fewer points than them. Also experimentally we show that although completion methods have taken sparse depth measurements as inputs, our method can further boost their performance with our local recovery strategy.

Besides boosting performance, the second benefit of our local recovery strategy is to better analyze the weakness of all existing depth estimation methods, and guide the design and choice of loss functions. The depth error can be decoupled into two parts: the coarse misalignment error and the detail-missing error. Compared with the error of global fitting, the alleviated error achieved by the local recovery is the coarse misalignment error, while the remaining one is the detail-missing error. Through our re-local-alignment analytical experiments in Table 3, we observe that current state-of-the-art depth estimation methods, including supervised learning metric depth, unsupervised learning scale-invariant depth, supervised learning affine-invariant depth, and depth completion, all suffer from noticeable misalignment issue *w.r.t.* the ground truth.

To achieve 3D scene reconstruction from a monocular video, retrieving a strong and robust monocular depth prediction model is also essential. We collect over 6.3 million images from the existing RGB-D datasets for training models using backbones such as ResNet50[29] and Swin-L[30], and investigate how much accuracy can we benefit from a large-scale dataset. With the robust monocular depth and local recovery strategy, the metric depth can be recovered by locally aligning with some sparse points.

The last challenge is how to obtain accurate sparse anchor points as the metric guidance. For analytical evaluation, we leverage the ground-truth depth to decouple and analyze the composition of errors. For practical application, aligning with ground-truth depth can be the upper bound of our local recovery strategy. We perturb the ground truth manually and analyze the error to simulate the inaccuracy of sparse anchor points. Also, The SfM[31] can be employed to recover sparse depth points on distinguishable feature points in practice. Through per-frame alignment with such accurate guidance, we can achieve geometrically consistent depth and perform robust 3D scene reconstruction. Although existing geometry-based methods, *e.g.*, multi-view stereo reconstruction[32–34], also takes the similar paradigms of structure from motion (SfM), their performance may suffer from inaccurate correspondences in low-texture regions. By contrast, our per-frame prediction comes from our well-trained strong monocular depth estimation model, which is much more robust in the low-texture regions.

Finally, through applying the local recovery strategy, our ResNet50 model drops the depth absolute relative error up to 50% on current affine-invariant depth evaluation benchmarks. For 3D scene reconstruction, our pipeline significantly improves both the accuracy and consistency, and achieves better performance than related works on five NYU[15] videos. To summarize, our main contributions are as follows:

- We propose a novel and effective metric depth recovery strategy, *i.e.*, locally weighted linear regression, which significantly improves the accuracy of the recovered metric depth with a very sparse set of anchor points. Extensive experiments show that the depth absolute relative error of state-of-the-art methods can drop up to 50% with our proposed method.
- Our local recovery strategy can be an analytical tool for subsequent depth prediction works, enabling decoupling the prediction errors and analyzing the weakness of their models.
- We train a robust monocular depth estimation model on large diverse data that contains 6.3 million images in total. We provide detailed analyses of its performance *w.r.t.* the training dataset size using our analytical tool.
- Aiming at the video-based scenarios, by combining our strong monocular depth estimation model with a geometry-based method for retrieving high-confidence anchor points, we design a new pipeline for robust and dense 3D scene reconstruction.

2 Materials and methods

The pipeline for our dense 3D scene reconstruction method is shown in Figure 1. Overall, our pipeline contains robust data-driven monocular depth estimation, a novel metric depth recovery, and RGB-D fusion[35].

2.1 Our Pipeline for Dense 3D Scene Reconstruction

Robust Monocular Depth Estimation Module. Retrieving robust and accurate depth maps from 2D images is significant for 3D scene reconstruction. In the supplementary material, we analyzed that the unsupervised depth estimation methods suffer from the weak supervision of photometric loss, while inaccurate correspondences may degrade the accuracy and robustness of MVS-based methods. Thus, the supervised monocular depth estimation method is employed in our pipeline, whose promising robustness and accuracy have been demonstrated in recent works[1, 3].

To retrieve strong monocular depth estimation models, we collect over 6.3 million data from 14 diverse datasets, which cover a wide range of scenes, camera poses, and camera intrinsic parameters. Following previous works, we enforce the network to learn the affine-invariant depth. Several scale-shift invariant losses are employed during training for better learning the inherent geometric information of depth maps, including the pair-wise normal (PWN) loss[1], image-level normalized regression (ILNR) loss[1], and multi-scale gradient (MSG) loss[5] as follows.

DOI: 10.52396/JUSTC-2023-0061
JUSTC, **2023**, 53(X):

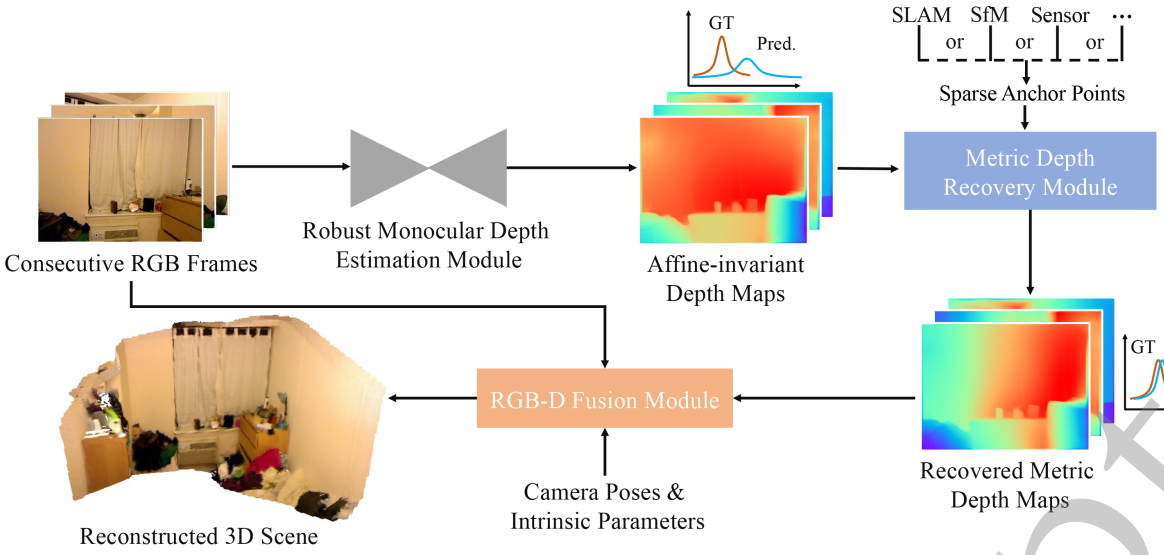


Fig. 1. The pipeline for dense 3D scene reconstruction. The robust monocular depth estimation model trained on 6.3 million images, locally weighted linear regression strategy, and TSDF fusion[35] are the main components of our method.

$$
\begin{aligned}
L_{\text{PWN}} &= \frac{1}{M}\sum_{i=1}^{M}|\boldsymbol{n}_{Ai}\cdot\boldsymbol{n}_{Bi}-\boldsymbol{n}^{*}_{Ai}\cdot\boldsymbol{n}^{*}_{Bi}| \\
L_{\text{INLR}} &= \frac{1}{N}\sum_{i=1}^{N}|d_i-\hat{d}^{*}_i|+|\tanh(d_i/100)-\tanh(\hat{d}^{*}_i/100)| \\
L_{\text{MSG}} &= \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{N}|\nabla^{k}_{x}d_i-\nabla^{k}_{x}\hat{d}^{*}_i|+|\nabla^{k}_{y}d_i-\nabla^{k}_{y}\hat{d}^{*}_i| \\
L &= L_{\text{PWN}}+\alpha L_{\text{ILNR}}+\beta L_{\text{MSG}}
\end{aligned}
\tag{1}
$$

Here, $\boldsymbol{n}_{Ai}$ and $\boldsymbol{n}_{Bi}$ represent the surface normal of sampled point pair (Ai, Bi). $\boldsymbol{n}^{*}_{Ai}$ and $\boldsymbol{n}^{*}_{Bi}$ are the corresponding ground truth. $\hat{d}^{*}_i=(d^{*}_i-\mu_{trim})/\sigma_{trim}$ is the Z-score normalized ground-truth depth, and μ_{trim} and σ_{trim} represent the mean and standard deviation of the trimmed ground-truth depth map, which removes the nearest and farthest 10% values in advance. The ∇^{k}_{x} and ∇^{k}_{y} stand for the gradient at k-th scale along x and y axis separately. The PWN loss samples M paired points on edge, planar, and random regions to supervise the surface normal information. The ILNR loss is introduced to reduce the average element-wise difference of N pixels between the predicted depth d_i and the normalized ground truth $\hat{d}^{*}_i$. The MSG loss ensures the accurate depth gradients at K scales. The loss function is balanced by hyperparameters α and β, which are set to 1 and 0.2 individually in our experiments.

Metric Depth Recovery Module. In the supplemental material, we analyzed the shape distortion and duplication caused by inaccurate shifts and inconsistent scales, which ensures the importance of recovering accurate and consistent scale-shift values along consecutive frames. Existing approaches recover a scale and a shift value for the depth map using least-squares fitting with some anchor points, which neglects the information of the distribution difference between affine-invariant prediction and anchor points. In contrast, we propose to perform locally weighted linear regression to recover the metric depth (Refer to Section 2.2 in detail). Compared to the global least-squares fitting, our local recovery strategy generates a scale and a shift map for each depth map, which can not only recover metric depth, but also correct the overall depth maps and ensure the accuracy and consistency of 3D reconstruction.

Our proposed local recovery strategy leverages the sparse anchor points obtained from SLAM system[36], SfM algorithms[31, 37], or some low-quality sensors, and can not only boost the performance of depth estimation, but also be an analytical tool to decouple the prediction depth errors into the coarse misalignment error and the detail missing error. Please see Section 3.2 in detail.

RGB-D Fusion Module. Through per-frame depth estimation and local recovery, locally scale-aligned monocular video depths are obtained. But the subtle details may still remain partially inconsistent between frames, which can cause outliers and duplication if we simply unproject it to 3D space without post-processing. Therefore, we propose to fuse multi-frame information with an RGB-D fusion module, which takes the RGB frames, depth maps, camera poses, and intrinsic parameters as inputs. It balances the difference between frames, filters out outliers and inconsistent regions between frames, and outputs the fused dense 3D mesh or point cloud.

In this work, we employ the TSDF fusion[35] to fuse multiple depth maps into a projective truncated signed distance function (TSDF) voxel volume during reconstruction. The sparse guided points used for local alignment can be obtained from various SLAM[36] systems, SfM[31, 37] algorithms, and some low-quality sensors such as ToF sensors of mobile phones. Two strong and robust monocular depth estimation models are trained based on ResNet50[29] and Swin-L[30] backbones, respectively.

2.2 Metric Depth Recovery

Monocular depth estimation methods[1–3, 7] have achieved promising results on diverse scenes. The problem is that their predicted depth/inverse depth is scale-shift-invariant, namely, affine-invariant depth/inverse depth[7]. Here we take the affine-invariant depth as an example. To recover the metric depth, it should be scaled and shifted, *i.e.*, $\hat{\boldsymbol{D}}=s\boldsymbol{D}+\theta\boldsymbol{J}$, where $\hat{\boldsymbol{D}}$, $\boldsymbol{D}$, s, θ and $\boldsymbol{J}$ are the recovered metric depth, predicted affine-invariant depth, scale, shift and all-ones matrix respectively.

DOI: 10.52396/JUSTC-2023-0061
JUSTC, **2023**, 53(X):

Some methods propose to obtain them through a global least-squares fitting method with ground-truth depth:

$$\begin{aligned}
&\min_{\boldsymbol{\beta}} \ (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}),\ \boldsymbol{X}=[\boldsymbol{d},\mathbf{1}] \ \in \mathcal{R}^{n\times 2}\\
&\boldsymbol{\beta}=[s,\theta]^{\mathrm{T}}\in\mathcal{R}^{2\times 1}, \boldsymbol{d},\boldsymbol{y}\in\mathcal{R}^{n\times 1}\\
&\hat{\boldsymbol{\beta}}=(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}\\
&\hat{\boldsymbol{D}}=s\boldsymbol{D}+\theta\boldsymbol{J},\ \text{with}\ \hat{\boldsymbol{D}},\boldsymbol{D},\boldsymbol{J}\in\mathcal{R}^{H\times W}
\end{aligned}\tag{2}$$

where $\boldsymbol{y}$ is the flattened ground-truth metric depth, $\boldsymbol{X}$ is the homogeneous representation of the flattened predicted depth $\boldsymbol{d}$, $n=H\times W$ represents the flattened length of depth map with a shape of (H,W). $\boldsymbol{\beta}$ is composed of scale value s and shift value θ, and $\hat{\boldsymbol{\beta}}$ is the optimized value of $\boldsymbol{\beta}$. Note that the scale value s and shift value θ can be regarded as a scale map S and a shift map Θ shared on the whole map.

However, such a globally scaling and shifting method often fails to reduce the spatially heteroscedastic errors, which follow a rather simple pattern. For example, we visualize the pixel-wise absolute relative error map between the ground truth and the globally recovered predicted depth in Figure 2(a), and observe the existence of the low-frequency spatial error. We see that the left part has a higher error than that of the right part. Motivated by this observation, we propose to leverage a locally recovering method, *i.e.*, *locally weighted linear regression (LWLR)*, to recover a scale and a shift map. Guided by very sparse ground-truth points, we can fix and quantify these low-rank spatial errors which are common in depth estimation tasks.

Locally Weighted Linear Regression. We thus employ a locally weighted regression method, which is:

$$\begin{aligned}
&\min_{\boldsymbol{\beta}_{u,v}} \ (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}_{u,v})^{\mathrm{T}}\boldsymbol{W}_{u,v}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}_{u,v}),\ \ \boldsymbol{d},\boldsymbol{y}\in\mathcal{R}^{m\times 1}\\
&\boldsymbol{X}=[\boldsymbol{d}^{\mathrm{T}},\mathbf{1}]\ \in\mathcal{R}^{m\times 2}, \boldsymbol{W}_{u,v}=diag(w_1,w_2,...,w_m)\\
&\boldsymbol{\beta}_{u,v}=[s_{u,v},\theta_{u,v}]^{\mathrm{T}}\in\mathcal{R}^{2\times 1}\\
&\hat{\boldsymbol{\beta}}_{u,v}=(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_{u,v}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_{u,v}\boldsymbol{y}\\
&\hat{\boldsymbol{D}}=\boldsymbol{S}\odot\boldsymbol{D}+\boldsymbol{\Theta},\ \text{with}\ \hat{\boldsymbol{D}},\boldsymbol{S},\boldsymbol{D},\boldsymbol{\Theta}\in\mathcal{R}^{H\times W}
\end{aligned}\tag{3}$$

where $\boldsymbol{y}$ is the sampled sparse ground-truth metric depth (we use around 25∼100 points in practice), $\boldsymbol{X}$ is the homogeneous representation of the sampled sparse predicted depth $\boldsymbol{d}$, m stands for the number of sampled points.

Different from recovering a scale-shift value or a globally shared scale-shift map by the global least-squares fitting method, we recover a location-aware scale-shift map. For each 2D coordinate (u,v), the predicted depth $\boldsymbol{d}$ can be fitted to the ground-truth depth $\boldsymbol{y}$ by minimizing the squared locally weighted ℓ_2 distance, which is re-weighted by a diagonal weight matrix $\boldsymbol{W}_{u,v}$. It pays more attention to sparse points closer to the estimated location, based on the idea that points near each other in the explanatory variable space are more likely to be related in a simple way. By iterating over the whole image, the scale map $\boldsymbol{S}$ and shift map $\boldsymbol{\Theta}$ can be generated composed of the scale values $s_{u,v}$ and shift values $\theta_{u,v}$ of each location (u,v). Finally, the locally recovered metric depth $\hat{\boldsymbol{D}}$ equals to the shift map $\boldsymbol{\Theta}$ plus the Hadamard product ($\odot$, known as element-wise product) of the affine-invariant depth $\boldsymbol{D}$ and the scale map $\boldsymbol{S}$. In our implementation, we employ a Gaussian kernel function to compute the weight matrix:

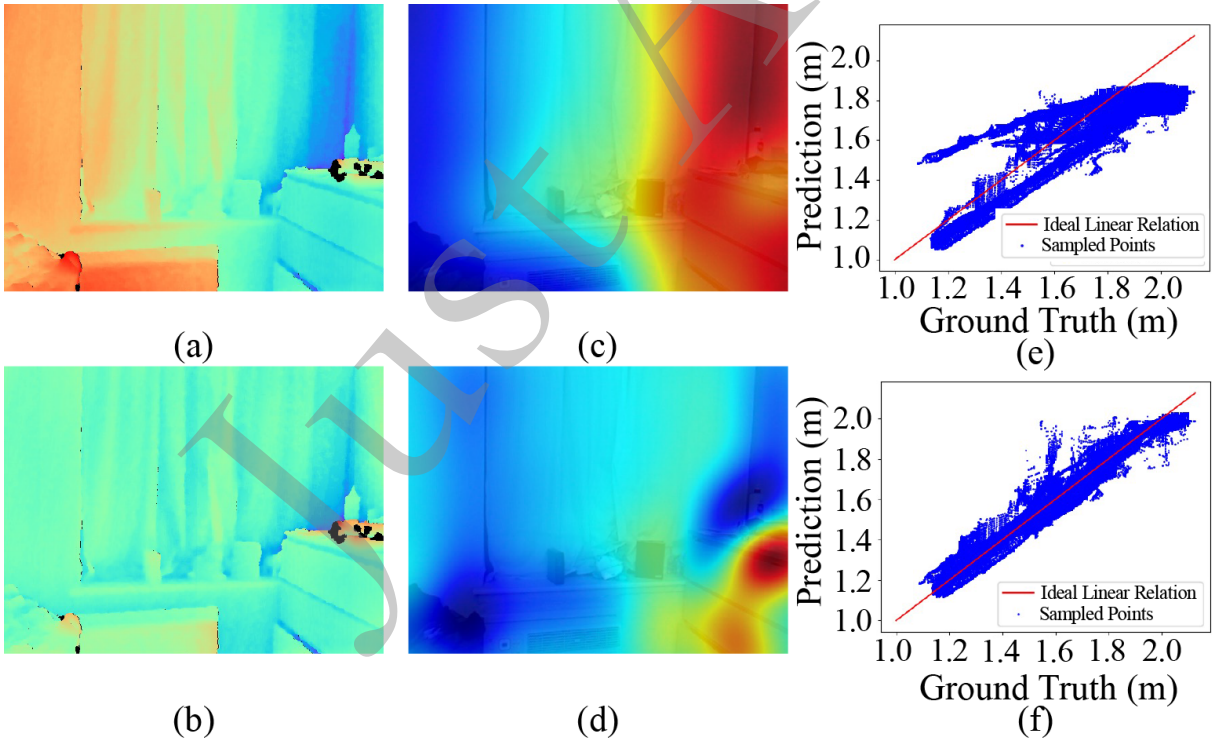


Fig. 2. The per-pixel error maps of ground-truth depth and predicted depth aligned through (a) global recovery and (b) local recovery, respectively. (c) The scale map and (d) the shift map of local recovery. Distribution of prediction-GT pairs obtained via (e) global recovery and (f) local recovery individually.

DOI: 10.52396/JUSTC-2023-0061
JUSTC, **2023**, 53(X):

$$w_i = \frac{1}{\sqrt{2\pi}} \exp(-\frac{\text{dist}_i^2}{2b^2}) \tag{4}$$

where b is the bandwidth of Gaussian kernel, and dist_i is the Euclidean distance between the guided point (u_i, v_i) and target point (u, v).

The scale map and shift map obtained this way can yield much more accurate metric depth than the global methods. However, some scale maps can be fitted to negative due to the shift-invariant characteristic of monocular depth and flexibility of weighted linear regression, which inverses the distribution of depth prediction and lacks reasonableness. Since the bias is not centered and the solution space is not bounded, results are widely distributed with no physical meanings and far from the real scale and shift. Therefore, we conduct the global recovery strategy to monocular depth first, and restrict the solution to be simple by adding an ℓ_2 regularization on the shift:

$$\begin{aligned}
&\min_{\boldsymbol{\beta}_{u,v}} \ (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_{u,v})^{\mathrm{T}} \boldsymbol{W}_{u,v} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_{u,v}) + \lambda \theta_{u,v}^2 \\
&\boldsymbol{X} = [\boldsymbol{d}^{\mathrm{T}}, \boldsymbol{1}] \ \in \mathcal{R}^{m\times 2}, \boldsymbol{W}_{u,v} = diag(w_1, w_2, ..., w_m) \\
&\boldsymbol{\beta}_{u,v} = [s_{u,v}, \theta_{u,v}]^{\mathrm{T}} \in \mathcal{R}^{2\times 1}, \ \ \boldsymbol{d}, \boldsymbol{y} \in \mathcal{R}^{m\times 1} \\
&\hat{\boldsymbol{\beta}}_{u,v} = (\boldsymbol{X}^{\mathrm{T}} \boldsymbol{W}_{u,v} \boldsymbol{X} + \boldsymbol{A})^{-1} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{W}_{u,v} \boldsymbol{y}, \boldsymbol{A} = \begin{bmatrix} \lambda & 0 \\ 0 & 0 \end{bmatrix} \\
&\hat{\boldsymbol{D}} = \boldsymbol{S} \odot \boldsymbol{D} + \boldsymbol{\Theta}, \ \text{with} \ \ \hat{\boldsymbol{D}}, \boldsymbol{S}, \boldsymbol{D}, \boldsymbol{\Theta} \in \mathcal{R}^{H\times W}
\end{aligned} \tag{5}$$

where $\boldsymbol{X}$ is the homogeneous representation of the globally recovered depth $\boldsymbol{d}$. With the regularization of shift, the location-related scale map is encouraged to be positive. Please see supplemental material for the visualization of the scale value distribution.

With our proposed local recovery strategy, we only need very sparse ground-truth depth (around 25-100 points) to recover the metric depth map by fitting a location-related scale map and a shift map. Figure 2 compares the global least-squares fitting and our proposed weighted linear regression results. Thanks to the optimized pixel-wise scale map (Figure 2(c)) and shift map (Figure 2(d)), the overall loss is reduced considerably (Figure 2(b)). Note that the predicted affine-invariant depths are the same for the two methods. Importantly, the recovered metric depth with our method is more linearly correlated to the ground truth (see Figure 2(e) and Figure 2(f)). Please see supplemental material for more examples.

3 Results and Discussion

The components of training datasets, the implementation details, and the evaluation details can be found in the supplemental material.

3.1 Dense 3D Scene Reconstruction

With our model trained on 6.3 million data and the local scale and shift recovery method, we can achieve high-quality 3D scene reconstruction through per-frame prediction and TSDF fusion[35]. To evaluate the consistency and accuracy, we collect 5 NYU videos and compare with the single image 3D reconstruction method (LeReS[1]), the state-of-the-art depth completion method (NLSPN[25]), the robust consistent video depth estimation method (RCVD[24]), the unsupervised video depth estimation method SC-DepthV2[20], and the learning-based MVS method DPSNet[38]. Note that NLSPN and SC-DepthV2 are trained on NYU, and only NLSPN can predict metric depth. Our method uses the same sparse ground truth (100 points) as NLSPN. For LeReS and RCVD, we align their predictions with metric depth globally. For SC-DepthV2 and DPSNet, only global scale values are recovered by ensuring the same medians as the ground truth. Besides leveraging sparse ground truth, we also sample points from SfM methods, *e.g.*, COLMAP[37], to reconstruct a 3D scene with an RGB video, ground-truth intrinsic and poses. The Rel, δ_1, the Chamfer l_1 distance and the F-score with the threshold of 5 cm are employed for evaluation.

Quantitative comparisons are shown in Table 1. First, we compare with the depth completion method NLSPN[25], which also uses the sparse guided points to obtain metric information. The main difference is that their model should be trained on the testing set and lacks generalization in the wild. By contrast, we can achieve better performance and generalize well to zero-shot datasets due to the robust depth prior. RCVD[24] and SC-DepthV2[20] aim to solve the visual consistency problem for video depth prediction. LeReS[1] reconstructs 3D scene shape from a single image and performs well in the wild. DPSNet[38] leverages CNNs to extract features and match between frames automatically. Before evaluating, the NLSPN and SC-depthV2 have been trained on the NYU data-

Table 1. Quantitative comparison of monocular depth estimation and 3D scene reconstruction with diverse related methods[1, 20, 24, 25, 38] on five NYU scenarios.

Method	Sparse Points	Basement_0001a				Bedroom_0015				Dining_room_0004				Kitchen_0008				Classroom_0004			
		Rel↓	δ_1↑	C-l_1↓	F-score↑	Rel↓	δ_1↑	C-l_1↓	F-score↑	Rel↓	δ_1↑	C-l_1↓	F-score↑	Rel↓	δ_1↑	C-l_1↓	F-score↑	Rel↓	δ_1↑	C-l_1↓	F-score↑
RCVD[24]	All	13.7	87.6	0.364	0.276	7.6	95.3	0.074	0.582	16.7	77.4	0.462	0.251	9.0	95.4	0.053	0.620	11.3	88.1	0.187	0.327
SC-DepthV2[20]	All	19.8	75.0	0.254	0.275	7.5	99.4	0.064	0.547	9.8	90.2	0.749	0.229	5.4	99.6	0.049	0.624	12.3	88.0	0.167	0.267
DPSNet[38]	All	19.1	62.3	0.243	0.299	16.2	78.2	0.195	0.276	21.1	65.7	0.995	0.186	18.6	76.4	0.269	0.203	18.9	62.9	0.296	0.195
NLSPN[25]	100	6.6	93.7	0.065	0.605	3.3	97.3	0.028	0.879	5.7	92.6	**0.073**	0.571	2.7	98.6	0.027	0.901	6.0	93.1	0.052	0.670
LeReS[1]	All	7.1	95.3	0.081	0.555	5.8	97.8	0.064	0.616	6.5	95.4	0.120	0.448	5.5	99.0	0.035	0.776	6.9	96.0	0.058	0.600
Ours (global)	All	8.9	92.5	0.085	0.525	5.9	97.6	0.053	0.629	6.2	96.0	0.111	0.474	2.7	**99.8**	0.033	0.798	6.4	96.0	0.061	0.570
Ours-SfM (local)	25	7.5	94.8	0.085	0.548	6.5	93.0	0.092	0.627	7.0	94.0	0.096	0.496	4.2	99.2	0.111	0.629	6.4	95.7	0.129	0.462
Ours (local)	100	**5.2**	**97.1**	**0.061**	**0.645**	**3.1**	**98.8**	**0.025**	**0.886**	**3.7**	**98.1**	**0.073**	**0.587**	**1.6**	**99.8**	**0.018**	**0.958**	**4.3**	**97.5**	**0.049**	**0.674**

DOI: 10.52396/JUSTC-2023-0061
JUSTC, **2023**, 53(X):

set. The "global" and "local" represent global and our proposed local metric depth recovery strategies separately. Ours-SfM (local) means performing the local recovery strategy with points sampled from SfM[31, 37] depth. As a result, our pipeline of 3D scene reconstruction from video achieves state-of-the-art performance on all five scenes.

Experiments of qualitative comparison are shown in the supplemental material. Depth completion method NLSPN performs well but misses some high-quality details due to the lack of geometry supervision during training. The RCVD focuses on visual depth consistency but fails to recover the shift of depth maps, leading to the distortion of the 3D structure. The SC-depthV2 achieves visually consistent video depth through an unsupervised paradigm, but the weak supervision brings some distortion during reconstruction. The LeReS achieves excellent detail prediction but lacks consistency between frames for misalignment caused by the global recovery strategy. The DPSNet improves the quality of extracted features with the help of CNNs, but without training on the NYU dataset, it lacks robustness to generalize to some unseen scenarios. With our local recovery strategy, our method can reconstruct better 3D point clouds than others. For Ours-SfM (local), we obtain SfM depth first, then fit the monocular depth with SfM depth and filter out the errors bigger than the 99th-percentile iteratively, before performing the local recovery strategy. Note that even with slightly inaccurate sparse SfM points, Ours-SfM (local) can still achieve comparable results with Ours(global), which requires ground-truth depth acquired from sensors.

3.2 Monocular Depth Estimation

Comparison with State-of-the-Art Methods. In this experiment, we compare with state-of-the-art robust monocular depth estimation methods[1–7, 39, 40] on five zero-shot datasets, whose scale and shift are recovered with a globally least-squares fitting method. During evaluation, the latest released model weights are adopted uniformly. As shown in Table 2, our ResNet50[29] model outperforms other ResNet50 and ResNeXt101[41] models on four testing datasets, and our Swin-L[30] model achieves comparable results with the ViT-large[42] model of DPT[2]. Through recovering scale and shift with the proposed locally weighted linear regression method, our method with ResNet50 and Swin-L backbones (*i.e.*, "Ours-R50 (local)", "Ours-swin (local)") can outperform all previous methods and our global predictions by a large margin over all zero-shot testing datasets. The qualitative comparison can be found in the supplemental material.

Effectiveness of Locally Weighted Linear Regression. To demonstrate our proposed locally weighted linear regression can boost various monocular depth estimation methods, we enforce it on several different methods: 1) learning affine-invariant depth methods, *e.g.*, LeReS[1], MiDaS[3], and DPT[2]; 2) learning metric depth on a specific dataset (VNL[7]); 3) learning scale-invariant depth with unsupervised methods (MonoDepth2[21]); 4) depth completion method (NLSPN[25]). Results are shown in Table 3. We uniformly sample 100 guided points to perform the local recovery, and all their performances can be boosted significantly (see the "w" columns). Critically, even though the NLSPN method has input such 100 sampled points for completion, our method can still further boost its performance. Note that the latest released weights and code are used for this experiment, NLSPN (KITTI) and Monodepth2 are trained on the KITTI dataset, and NLSPN (NYU) and VNL are trained on the NYU dataset.

Decoupling of Monocular Depth Error. Besides improving performance, the local recovery strategy is also performed to decouple monocular depth error between ground truth and globally aligned prediction into coarse misalignment error and detail-missing error. Compared with the error of global recovery, the alleviated error brought by local recovery represents the coarse misalignment error, and the remaining one stands for detail-missing error. As shown in Table 3, the percentage of coarse misalignment error ("%" columns)

Table 2. Quantitative comparison of monocular depth estimation with state-of-the-art methods on five unseen datasets. The numbers in brackets represent the reduced absolute relative error brought by our local recovery method.

Method	Backbone	KITTI		NYU		ScanNet		ETH3D		DIODE		Rank
		Rel↓	δ_1↑	Rel↓	δ_1↑	Rel↓	δ_1↑	Rel↓	δ_1↑	Rel↓	δ_1↑	
OASIS[9]	ResNet50[29]	31.7	43.7	21.9	66.8	19.8	69.7	29.2	59.5	48.4	53.4	12.7
MegaDepth[5]	ResNet50[29]	20.1	63.3	19.4	71.4	19.0	71.2	26.0	64.3	39.1	61.5	10.7
Xian *et al.* [4]	ResNet50[29]	27.0	52.9	16.6	77.2	17.4	75.9	27.3	63.0	42.5	61.8	10.7
WSVD[39]	ResNet50[29]	24.4	60.2	22.6	65.0	18.9	71.4	26.1	61.9	35.8	63.8	10.8
Chen *et al.* [40]	ResNet50[29]	32.7	51.2	16.6	77.3	16.5	76.7	23.7	67.2	37.9	66.0	9.7
DiverseDepth[7]	ResNet50[29]	19.0	70.4	11.7	87.5	10.8	88.2	22.8	69.4	37.6	63.1	8.3
LeReS[1]	ResNext101[41]	14.8	78.6	8.6	92.1	9.5	91.2	9.7	90.3	21.6	80.8	5.7
MiDaS-large[3]	ResNext101[41]	13.4	81.9	10.2	90.0	9.8	90.9	10.1	90.5	19.0	76.1	6.3
DPT-large[2]	ViT-Large[42]	10.0	90.1	9.8	90.3	7.8	93.8	7.8	94.6	18.2	75.8	4.2
Ours-R50 (global)	ResNet50[29]	10.9	88.5	8.2	92.6	8.9	92.0	8.4	92.1	22.0	80.1	4.6
Ours-swin (global)	Swin-L[30]	11.2	88.3	7.2	94.1	7.2	94.5	8.1	93.9	22.4	79.5	4.2
Ours-R50 (local)	ResNet50[29]	**5.7** (−5.2)	**95.4**	4.7 (−3.5)	96.9	4.3 (−4.6)	97.4	5.0 (−3.4)	96.6	**16.5** (−5.5)	**85.2**	1.6
Ours-swin (local)	Swin-L[30]	5.8 (−5.4)	95.0	**4.5** (−2.7)	**97.1**	**3.8** (−3.4)	**97.9**	**4.7** (−3.4)	**96.7**	**16.5** (−5.9)	85.1	**1.3**

DOI: 10.52396/JUSTC-2023-0061
JUSTC, **2023**, 53(X):

Table 3. Boosting various monocular depth estimations with our local recovery method. We compare the accuracy without ("w/o") and with ("w") our recovery methods, and show the reduced errors and percentages ("w" and "%").

Method	Type	KITTI			NYU			ScanNet			ETH3D			DIODE		
		Rel↓														
		w/o	w	%	w/o	w	%	w/o	w	%	w/o	w	%	w/o	w	%
LeReS[1]	Affine-invariant	14.8	8.0 (−6.8)	−46%	8.6	4.8 (−3.8)	−44%	9.5	4.4 (−5.1)	−54%	9.7	5.6 (−4.1)	−42%	21.6	16.2 (−5.4)	−25%
MiDaS-large[3]		13.4	6.2 (−7.2)	−54%	10.2	5.1 (−5.1)	−50%	9.8	4.3 (−5.5)	−56%	10.1	4.9 (−5.2)	−52%	19.0	14.5 (−4.5)	−24%
DPT-large[2]		10.0	5.2 (−4.8)	−48%	9.8	5.0 (−4.8)	−49%	7.8	3.8 (−4.0)	−51%	7.8	4.2 (−3.6)	−54%	18.2	14.3 (−3.9)	−21%
NLSPN[25]	Depth Completion	4.3	3.7 (−0.6)	−14%	3.7	3.3 (−0.4)	−11%	–	–	–	–	–	–	–	–	–
VNL[7]	Metric	–	–	–	10.6	4.0 (−6.6)	−62%	–	–	–	–	–	–	–	–	–
Monodepth2[21]	Scale-invariant	8.1	6.1 (−2.0)	−25%	–	–	–	–	–	–	–	–	–	–	–	–

nearly remains consistent between affine-invariant depth estimation, which may reveal the drawback of coarse misalignment of monocular depth estimation. The bottleneck of NLSPN[25] and MonoDepth2[21] is detail-missing for the alleviated error in percentage remains small.

3.3 Ablation Study

Ablation Study for Training Data. In this experiment, we aim to study the relations between data volume and performance improvement. We gradually aggregate more data for training, and evaluate the performance on 5 zero-shot datasets. Note that 3 different quality data sources are increased in balance, and the results are illustrated in Table 4. We can observe that when the data size increases from 42K to 900K (around by 20 times), the performance is boosted significantly. However, when further increased by 7 times, the accuracy can only be improved slightly. We conjecture that such large-scale data has fully exploited the capacity of the model (ResNet50 backbone).

Furthermore, we also conduct local recovery here to decouple the error into coarse misalignment error and detail-missing error. As shown in Table 4, the percentage of coarse misalignment error remains nearly constant, which shows the model can study the detail information and the global structure simultaneously.

Ablation Study for Locally Weighted Linear Regression Method. The performance of our proposed local recovery strategy may be affected by the amount of sparse points, the sparsity distribution, random noises from sparse points, and the bandwidth b. Their effects on the depth accuracy are explored and shown in Table 5. Here, "Amount", "Distribution", and "Noise" correspond to the number, distribution, and maximum perturbation percentage of the sampled ground truth. Parameter b represents the bandwidth of the Gaussian kernel function. "Grid" means sampling points from the vertexes of the evenly divided image plane, and "Uniform" means sampling randomly. The whole image and half image stand for only sampling ground truth from the original whole or half image. All experiments are conducted on the

Table 4. Ablation study for training data of our robust depth estimation module. With the increase of data, the performance of depth estimation improves gradually.

Training Data	KITTI	NYU	ScanNet	ETH3D	DIODE
	Rel↓				
Global Recovery					
42K	14.1	11.0	11.8	9.5	23.2
352K	15.0	9.2	10.0	9.4	22.1
900K	11.5	8.6	9.7	9.0	21.6
3.8M	11.1	8.2	8.9	8.4	21.9
6.3M	10.9	8.2	8.9	8.4	22.0
Local Recovery					
42K	6.9 (−51%)	5.8 (−47%)	5.2 (−56%)	5.9 (−38%)	16.9 (−27%)
352K	7.3 (−51%)	5.2 (−43%)	4.7 (−53%)	6.2 (−34%)	16.6 (−25%)
900K	5.8 (−50%)	4.9 (−43%)	4.6 (−53%)	5.4 (−40%)	16.3 (−25%)
3.8M	5.8 (−48%)	4.7 (−43%)	4.3 (−52%)	5.2 (−38%)	16.5 (−25%)
6.3M	5.7 (−48%)	4.7 (−43%)	4.3 (−52%)	5.0 (−40%)	16.5 (−25%)

DOI: 10.52396/JUSTC-2023-0061
JUSTC, **2023**, 53(X):

Table 5. Ablation study for parameters of our proposed local recovery strategy.

Amount	Distribution	Noise	b	NYU	
				Rel ↓	δ_1 ↑
10×10	Grid	0%	50	4.8	96.4
5×5	Grid	0%	50	5.8	94.7
20×20	Grid	0%	50	4.6	96.7
10×10	Grid	0%	25	4.7	96.0
10×10	Grid	0%	100	5.8	95.5
10×10	Grid	10%	50	5.5	96.2
10×10	Grid	20%	50	6.8	95.5
100	Uniform (whole image)	0%	50	4.3	97.0
100	Uniform (half image)	0%	50	11.3	87.6

NYU dataset.

According to the experiments, simple 5×5 ground-truth depth can be leveraged to recover metric depth and improve accuracy. More ground truth leads to more performance boost, please see Table 6 for more detailed analysis. The "Global" and "Local" represent the global recovery and local recovery strategies separately. "Grid" and "Uniform" stand for sampling from grid and sampling uniformly. Our strategy performs robust to the amount, distribution and noise of sparse points, but the overly concentrated sampling strategy should be avoided in practice.

Table 6. Analysis for the amount of ground-truth points during recovering monocular metric depth. The Rel decreases faster with our proposed local recovery strategy with the increase of ground-truth points.

Ground-Truth Points	Local-Uniform	Global-Uniform	Local-Grid	Global-Grid
	Rel↓			
1	73.84	73.84	74.43	74.43
2	27.1	27.1	–	–
3	14.73	14.78	–	–
4	12.77	13.15	9.33	10.23
5	10.73	11.23	–	–
6	9.56	10.2	–	–
7	9.13	10.07	–	–
8	8.85	9.85	–	–
9	8.49	9.74	7.1	9.18
16	7.09	9.11	5.9	8.76
25	6.17	8.68	5.38	8.62
36	5.74	8.56	5.12	8.52
49	5.4	8.36	5.06	8.58
64	5.2	8.33	4.92	8.49
81	5.06	8.4	4.79	8.48
100	4.9	8.41	4.68	8.46
144	4.78	8.34	4.55	8.43
196	4.62	8.26	4.47	8.42
256	4.57	8.28	4.4	8.33
324	4.49	8.24	4.35	8.35
400	4.42	8.21	4.3	8.34
900	4.23	8.24	4.13	8.27

As for bandwidth b, it represents the effect of distance on the weight matrix. Experimentally, we suggest simply setting parameter bandwidth to the value $\frac{l}{n}$, where l is the width of the RGB image and n is the number of sampled ground-truth points of one side. More precisely, if we sample 10×10 ground-truth points for a 500×500 image, the parameter bandwidth can be set to 50.

4 Conclusions

In this paper, we have leveraged the robust data-driven monocular depth estimation model, local recovery strategy and an RGB-D fusion module to implement a complete dense 3D scene reconstruction pipeline. Compared to existing 3D reconstruction methods, our pipeline achieves improved robustness, accuracy and consistency along consecutive RGB frames. Extensive experiments show that our method demonstrates a significantly better generalization ability to monocular depth estimation and 3D scene reconstruction.

The proposed local recovery strategy can not only improve the accuracy and consistency of depth estimation significantly with robustness to both the amount and randomly-generated noises of the ground truth, but also can be an analytical tool to expose the shortcomings of existing depth estimation methods.

Supplemental Information

The supplemental information includes three sections, eight figures, and two tables. We first elaborate on some preliminary information of existing 3D scene reconstruction methods and affine-invariant depth estimation. Then, we introduce some related works. Finally, the experimental details and more visualization are supplied.

Conflict of Interest

The authors declare that they have no conflict of interest, and promise the preprint version on arXiv has not been published anywhere else.

References

[1] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Long Mai, Simon Chen, and Chunhua Shen. Learning to recover 3d scene shape from a single image. In: *IEEE Conf. Comput. Vis. Pattern Recogn.*, 204–213, **2021**.

[2] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In: *IEEE Conf. Comput. Vis. Pattern Recogn.*, 12179–12188, **2021**.

[3] 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 Trans. Pattern Anal. Mach. Intell.*, **2022**, *44* (3): 1623–1637.

[4] Ke Xian, Jianming Zhang, Oliver Wang, Long Mai, Zhe Lin, and Zhiguo Cao. Structure-Guided Ranking Loss for Single Image Depth Prediction. In: *IEEE Conf. Comput. Vis. Pattern Recogn.*, 611–620,

DOI: 10.52396/JUSTC-2023-0061
JUSTC, **2023**, 53(X):

2020.

[5] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In: *IEEE Conf. Comput. Vis. Pattern Recogn.*, 2041–2050, **2018**.

[6] Weifeng Chen, Shengyi Qian, David Fan, Noriyuki Kojima, Max Hamilton, and Jia Deng. OASIS: A large-scale dataset for single image 3d in the wild. In: *IEEE Conf. Comput. Vis. Pattern Recogn.*, 679–688, **2020**.

[7] Wei Yin, Yifan Liu, and Chunhua Shen. Virtual normal: Enforcing geometric constraints for accurate and robust depth prediction. *TPAMI,* **2021**, *44* (10): 7282–7295.

[8] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. In: *Adv. Neural Inform. Process. Syst.*, 730–738, **2016**.

[9] Jia-Wang Bian, Huangying Zhan, Naiyan Wang, Zhichao Li, Le Zhang, Chunhua Shen, Ming-Ming Cheng, and Ian Reid. Unsupervised scale-consistent depth learning from video. *Int. J. Comput. Vis.,* **2021**, *129* (9): 2548–2564.

[10] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In: *IEEE Conf. Comput. Vis. Pattern Recogn.*, 8445–8453, **2019**.

[11] Wanli Peng, Hao Pan, He Liu, and Yi Sun. Ida-3d: Instance-depth-aware 3d object detection from stereo vision for autonomous driving. In: *IEEE Conf. Comput. Vis. Pattern Recogn.*, 13015–13024, **2020**.

[12] Jingwei Huang, Zhili Chen, Duygu Ceylan, and Hailin Jin. 6-DOF VR videos with a single 360-camera. In: *IEEE Vir. Real.*, 37–44, **2017**.

[13] Julien Valentin, Adarsh Kowdle, Jonathan T Barron, Neal Wadhwa, Max Dzitsiuk, Michael Schoenberg, Vivek Verma, Ambrus Csaszar, Eric Turner, Ivan Dryanovski, et al. Depth from motion for smartphone AR. *ACM Transh. Graph.,* **2018**, *37* (6): 1–19.

[14] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In: *IEEE Conf. Comput. Vis. Pattern Recogn.*, 1746–1754, **2017**.

[15] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In: *Eur. Conf. Comput. Vis.*, 746–760. Springer, **2012**.

[16] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In: *IEEE Conf. Comput. Vis. Pattern Recogn.*, 567–576, **2015**.

[17] Xingbin Yang, Liyang Zhou, Hanqing Jiang, Zhongliang Tang, Yuanbo Wang, Hujun Bao, and Guofeng Zhang. Mobile3DRecon: real-time monocular 3D reconstruction on a mobile phone. *IEEE Transh. Vish. and Comph. Graphh.,* **2020**, *26* (12): 3446–3456.

[18] Dejan Azinović, Ricardo Martin-Brualla, Dan B Goldman, Matthias Nießner, and Justus Thies. Neural RGB-D surface reconstruction. In: *IEEE Conf. Comput. Vis. Pattern Recogn.*, 6290–6301, **2022**.

[19] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In: *Int. Conf. Comput. Vis.*, 10786–10796, **2021**.

[20] Jia-Wang Bian, Huangying Zhan, Naiyan Wang, Tat-Jun Chin, Chunhua Shen, and Ian Reid. Auto-Rectify Network for Unsupervised Indoor Depth Estimation. *TPAMI,* **2021**: 1–1.

[21] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In: *Int. Conf. Comput. Vis.*, 3828–3838, **2019**.

[22] Jamie Watson, Oisin Mac Aodha, Victor Prisacariu, Gabriel Brostow, and Michael Firman. The temporal opportunist: Self-supervised multi-frame monocular depth. In: *IEEE Conf. Comput. Vis. Pattern Recogn.*, 1164–1174, **2021**.

[23] Xuan Luo, Jia-Bin Huang, Richard Szeliski, Kevin Matzen, and Johannes Kopf. Consistent video depth estimation. *ACM Trans. Graph.,* **2020**, *39* (4): 71–1.

[24] Johannes Kopf, Xuejian Rong, and Jia-Bin Huang. Robust consistent video depth estimation. In: *IEEE Conf. Comput. Vis. Pattern Recogn.*, 1611–1621, **2021**.

[25] Jinsun Park, Kyungdon Joo, Zhe Hu, Chi-Kuei Liu, and In So Kweon. Non-local spatial propagation network for depth completion. In: *Eur. Conf. Comput. Vis.*, 120–136. Springer, **2020**.

[26] Alex Wong and Stefano Soatto. Unsupervised depth completion with calibrated backprojection layers. In: *Int. Conf. Comput. Vis.*, 12747–12756, **2021**.

[27] Alex Wong, Safa Cicek, and Stefano Soatto. Learning topology from synthetic data for unsupervised depth completion. *IEEE Rob. and Auto. Lett.,* **2021**, *6* (2): 1495–1502.

[28] Yanchao Yang, Alex Wong, and Stefano Soatto. Dense depth posterior (ddp) from single image and sparse range. In: *IEEE Conf. Comput. Vis. Pattern Recogn.*, 3353–3362, **2019**.

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In: *IEEE Conf. Comput. Vis. Pattern Recogn.*, 770–778, **2016**.

[30] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In: *Int. Conf. Comput. Vis.*, 10012–10022, **2021**.

[31] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In: *IEEE Conf. Comput. Vis. Pattern Recogn.*, 4104–4113, **2016**.

[32] Zachary Teed and Jia Deng. DeepV2D: Video to Depth with Differentiable Structure from Motion. In: *Int. Conf. Learn. Represent.*, **2020**.

[33] Chao Liu, Jinwei Gu, Kihwan Kim, Srinivasa G. Narasimhan, and Jan Kautz. Neural RGB → D sensing: Depth and uncertainty from a video camera. In: *IEEE Conf. Comput. Vis. Pattern Recogn.*, 10986–10995, **2019**.

[34] Arda Duzceker, Silvano Galliani, Christoph Vogel, Pablo Speciale, Mihai Dusmanu, and Marc Pollefeys. DeepVideoMVS: Multi-view stereo on video with recurrent spatio-temporal fusion. In: *IEEE Conf. Comput. Vis. Pattern Recogn.*, 15324–15333, **2021**.

[35] Andy Zeng, Shuran Song, Matthias Nießner, Matthew Fisher, Jianxiong Xiao, and Thomas Funkhouser. 3dmatch: Learning local geometric descriptors from rgb-d reconstructions. In: *IEEE Conf. Comput. Vis. Pattern Recogn.*, 1802–1811, **2017**.

[36] Raúl Mur-Artal, J. M. M. Montiel, and Juan D. Tardós. ORB-SLAM: a Versatile and Accurate Monocular SLAM System. *IEEE Transactions on Robotics,* **2015**, *31* (5): 1147–1163.

[37] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In: *Eur. Conf. Comput. Vis.*, 501–518. Springer, **2016**.

[38] Sunghoon Im, Hae-Gon Jeon, Stephen Lin, and In So Kweon. DPSNet: Endto-end Deep Plane Sweep Stereo. In: *Int. Conf. Learn. Represent.*, **2019**.

[39] Chaoyang Wang, Simon Lucey, Federico Perazzi, and Oliver Wang. Web stereo video supervision for depth prediction from dynamic scenes. In: *Int. Conf. 3D Vis.*, 348–357, **2019**.

[40] Weifeng Chen, Shengyi Qian, and Jia Deng. Learning single-image depth from videos using quality assessment networks. In: *IEEE Conf. Comput. Vis. Pattern Recogn.*, 5604–5613, **2019**.

[41] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Ag-gregated residual transformations for deep neural networks. In: *IEEE Conf. Comput. Vis. Pattern Recogn.*, 1492–1500, **2017**.

[42] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: *Int. Conf. Learn. Represent.*, **2021**.

DOI: 10.52396/JUSTC-2023-0061
JUSTC, **2023**, 53(X):