FrozenRecon: Pose-free 3D Scene Reconstruction with Frozen Depth Models Supplementary Material

Guangkai Xu[1]*, Wei Yin[2]*, Hao Chen[3], Chunhua Shen[3,4], Kai Cheng[1], Feng Zhao[1]†

[1] University of Science and Technology of China [2] DJI Technology [3] Zhejiang University, China [4] Ant Group

1. Details about the LWLR Module

Given a globally aligned depth map $\boldsymbol{D}^g$ and sparse guided points $\boldsymbol{y}$, the LWLR module [9] recovers a location-aware scale-shift map. Concretely, for each 2D coordinate (u, v), the sampled globally aligned 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 $\boldsymbol{W}_{u,v}$.

$$
\begin{gathered}
\min_{\boldsymbol{\beta}_{u,v}} \ (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}_{u,v})^{\top}\boldsymbol{W}_{u,v}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}_{u,v})+\lambda\theta_{u,v}^2, \\
\boldsymbol{W}_{u,v}=diag(w_1,w_2,...,w_m),\ w_i=\frac{1}{\sqrt{2\pi}}\exp(-\frac{\mathrm{dist}_i^2}{2b^2}), \\
\hat{\boldsymbol{\beta}}_{u,v}=(\boldsymbol{X}^{\top}\boldsymbol{W}_{u,v}\boldsymbol{X}+\boldsymbol{A})^{-1}\boldsymbol{X}^{\top}\boldsymbol{W}_{u,v}\boldsymbol{y}, \\
\boldsymbol{A}=\begin{bmatrix}\lambda & 0\\ 0 & 0\end{bmatrix},\ \boldsymbol{D}=\boldsymbol{S}\odot\boldsymbol{D}^g+\boldsymbol{\Theta}, \\
\boldsymbol{X}=[\boldsymbol{d}^{\top},\mathbf{1}]\ \in\mathcal{R}^{m\times 2},\ \boldsymbol{\beta}_{u,v}=[s_{u,v},\theta_{u,v}]^{\top}\in\mathcal{R}^{2\times 1}, \\
\boldsymbol{D},\boldsymbol{S},\boldsymbol{D}^g,\boldsymbol{\Theta}\in\mathcal{R}^{H\times W},\ \ \boldsymbol{d},\boldsymbol{y}\in\mathcal{R}^{m\times 1},
\end{gathered}
\tag{1}
$$

where $\boldsymbol{y}$ is the sampled sparse ground-truth metric depth, $\boldsymbol{d}$ is the sampled globally aligned depth, whose 2D coordinates are the same with those of sparse guided points $\boldsymbol{y}$. $\boldsymbol{X}$ is the homogeneous representation of $\boldsymbol{d}$, m stands for the number of sampled points. b is the bandwidth of Gaussian kernel, and dist_i is the Euclidean distance between the the coordinate (u_i, v_i) of i-th guided point and target point (u, v). λ is a l_2 regularization hyperparameter used for restricting the solution to be simple. By iterating the target point (u, v) 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 $\boldsymbol{S}$. The operation above can be summarized as below.

$$
\begin{aligned}
\boldsymbol{S},\boldsymbol{\Theta}&=f_{\mathrm{LWLR}}(\boldsymbol{D}^g,\boldsymbol{y}),\\
\boldsymbol{D}&=\boldsymbol{S}\odot\boldsymbol{D}^g+\boldsymbol{\Theta}.
\end{aligned}
\tag{2}
$$

Rather than relying on sparse ground-truth metric depth $\boldsymbol{y}$, we replace it with $\{\omega_{i,t}\cdot d_i^g(\mathbf{p}_t)\}_{t=1}^M$, which is related to parameters $\{\omega_{i,t}\}_{t=1}^M$ and sampled sparse global depth $\{d_i^g(\mathbf{p}_t)\}_{t=1}^M$. By ensuring multi-frame consistency, we can retrieve scale-consistent depth maps.

$$
\begin{aligned}
\mathbf{A}_i,\mathbf{B}_i&=f_{\mathrm{LWLR}}(\mathbf{D}_i^g,\{\omega_{i,t}\cdot d_i^g(\mathbf{p}_t)\}_{t=1}^M),\\
\mathbf{D}_i&=\mathbf{A}_i\odot\mathbf{D}_i^g+\mathbf{B}_i.
\end{aligned}
\tag{3}
$$

Table S1: **Efficiency of the photometric constraint.** We simply replace the photometric constraint with spatial constraint ("w/ flow"), and supervises the coordinates consistency warped by optical flow [7] and optimized parameters. Although comparable performance can be achieved, predicting dense optical flows with a deep network between every two frames is computationally expensive. Our algorithm can be more efficient while remaining comparable performance.

Method	Time	Depth		Pose			Reconstruction	
	Complexity	AbsRel↓	δ_1 ↑	ATE↓	RPE-T↓	RPE-R↓	C-l_1 ↓	F-score↑
Ours	$O(1)$	0.092	0.923	0.096	0.144	0.053	0.099	0.622
w/ flow	$O(N^2)$	0.102	0.907	0.095	0.155	0.052	0.085	0.627

2. Efficiency of Photometric Constraint

Furthermore, we also explore the efficiency of the photometric constraint on the NYU [5] dataset. We supervise the consistency of coordinates warped by the optical flow and the optimized parameters, as a replacement for photometric constraint. As shown in Table S1, although comparable performance can the flow-guided constraint achieve, it relies on predicting dense optical flow between every two frames with a robust model RAFT [7], which can be time-consuming due to the recurrent refinement model and the

*Both authors contributed equally. † Corresponding author. Email: fzhao956@ustc.edu.cn.

Table S2: **Keyframe sampling strategy.** We sample keyframes according to the valid regions of the estimated optical flow. As a result, our algorithm not only consumes less time but also achieves better performance.

Method	Time Complexity	Depth		Pose			Reconstruction	
		AbsRel↓	δ_1 ↑	ATE↓	RPE-T↓	RPE-R↓	C-l_1 ↓	F-score↑
Ours	$O(1)$	0.092	0.923	0.096	0.144	0.053	0.099	0.622
w/ flow keyframe	$O(N^2)$	0.103	0.897	0.174	0.262	0.113	0.155	0.565

$O(N^2)$ time complexity. In contrast, the photometric constraint does not require offline-computed optical flow and can be more efficient, especially on long-form videos.

3. Keyframe Sampling Strategy

We sample the keyframes with the optical flow computed with RAFT, and compare it with ours on the NYU dataset. The flow-guided keyframes will be selected if the valid regions are larger than 30 percent after checking the forward-backward consistency. As shown in Table S2, our algorithm is more efficient and even outperforms the flow-guided keyframe sampling strategy due to the pose-based long-range keyframe sampling.

4. Analysis of Optimization Objectives

Our optimization objectives are composed of photometric constraint, geometric constraint, and regularization constraint. The photometric constraint L_{pc} ensures the color consistency between the reference frame and the warped source frame. If we directly optimize the per-frame pixel-wise depth map with the photometric constraint, the supervision signal can be too weak to achieve satisfactory performance, especially on some low-texture regions. Here weak means supervising the color consistency instead of precise coordinate correspondences. However, the weak supervision becomes an advantage when it is employed with the geometric constraint and the robust affine-invariant depth prior. The affine-invariant depth maps can offer reliable inherent geometry information and narrow the solution space together with the geometric constraint. Concretely, the photometric constraint offers accurate guidance on the rich texture regions. For some low-texture regions, the photometric constraint will be small, and the optimization is mainly guided by the supervision of geometric consistency, which is also reliable due to the geometric accuracy of affine-invariant depth.

For the geometric constraint L_{gc}, it can ensure the multi-view geometric consistency and will not bring any incorrect supervision, but the weight should not be too large to prevent from encouraging the whole depth map to be infinitely large. The weakly normalization supervision on the sparse guided points L_{regu} is utilized to avoid extreme point cloud distortion and stabilize the optimization.

Table S3: **Upper bound analysis.** With known GT poses, our algorithm can achieve better performance. The GT intrinsics alone does not improve the reconstruction performance, but can achieve slight performance improvement together with GT poses.

GT intrinsics	GT poses	Depth		Pose			Reconstruction	
		AbsRel↓	δ_1 ↑	ATE↓	RPE-T↓	RPE-R↓	C-l_1 ↓	F-score↑
		0.092	0.923	0.096	0.144	0.053	0.099	0.622
✓		0.092	0.920	0.095	0.147	0.050	0.103	0.622
	✓	0.085	0.933	-	-	-	0.070	0.662
✓	✓	0.081	0.938	-	-	-	0.064	0.674

5. Upper Bound Analysis

Our optimization can work better if accurate ground-truth (GT) poses and intrinsics are given. As shown in Table S3, the performance on the NYU dataset remains nearly the same with known GT intrinsics. With known GT poses, the quality of depth, pose, and reconstruction can be improved compared to video-only optimization. With both GT poses and intrinsics, the performance can achieve slightly better results.

6. Discussion of Imperfect Cases

Unlike other scenes, the outdoor KITTI sequences involve mostly straight-line movement with small differences between frames. During optimization, the photometric and geometric losses remain small even without accurate depths and poses. Although inexact, the depths and poses are consistent to enable acceptable reconstruction.

Despite the imperfect estimation of camera poses, we can still yield satisfactory reconstruction results, because it depends on the accuracy and the consistency of depth maps and poses. The advantage of our optimization pipeline lies in enabling the practical use of affine-invariant depth, and ensuring the aforementioned consistency. Also, the robustness of affine-invariant depth is transferred to pose estimation, leading to fewer failure cases such as 'scene0707_00'. Our pipeline also allows users to input offline-obtained poses, such as SfM poses, and jointly optimize for further improvement.

After optimization, the reconstructed point cloud and trajectory are consistent but still up to an unknown scale w.r.t. the real world. It is an intrinsic limitation of purely monocular reconstruction methods. The unknown scale can be recovered by providing the GT poses, or measuring the length of an object and aligning the reconstructed object's size.

Table S4: **Evaluation sequences of five zero-shot testing datasets.** Note that we evaluate the first sequences of 7-Scenes [4].

Dataset	Scenes
NYU [5]	basement_0001a, bedroom_0015, bedroom_0036, bedroom_0059, classroom_0004, computer_lab_0002, dining_room_0004, dining_room_0033, home_office_0004, kitchen_0008, kitchen_0059, living_room_0058, office_0006, office_0024, playroom_0002
ScanNet [2]	scene0707_00, scene0708_00, scene0709_00, scene0710_00, scene0711_00, scene0712_00, scene0713_00, scene0714_00, scene0715_00, scene0716_00, scene0717_00, scene0718_00, scene0719_00, scene0720_00
7-Scenes [4]	chess, fire, heads, office, pumpkin, redkitchen, stairs
TUM [6]	360, desk, desk2, floor, plant, room, rpy, teddy, xyz
KITTI [3]	2011_09_26_0001_sync, 2011_09_26_0009_sync, 2011_09_26_0091_sync, 2011_09_28_0001_sync, 2011_09_29_0004_sync, 2011_09_29_0071_sync

7. Evaluation Details

For 3D scene reconstruction, we evaluate the Chamfer l_1 distance (C-l_1) and F-score with a threshold of 5cm on the point cloud. Because of the unknown scale of estimated point clouds, we propose first to align the scale of depth maps and poses with ground truth through a global sharing scale factor, which is the ratio of the median depth value of all frames between optimized depth maps and ground-truth depth maps. Then, we match the estimated poses with ground truth through a 4×4 transformation matrix. The matrix is computed by employing Open3D's iterative closest point (ICP) [1] algorithm between the optimized and the ground-truth point clouds. To reduce the negative effect of outliers for ICP matching, we remove some noisy points whose AbsRel errors are greater than 20%.

When evaluating depth, absolute relative error (AbsRel= $\frac{|d_{pred}-d_{gt}|}{d_{gt}}$) and the percentage of accurate depth pixels with $\delta_1 = \max(\frac{d_{pred}}{d_{gt}}, \frac{d_{gt}}{d_{pred}}) < 1.25$ are employed. To compare the consistency of depths along the video, we align all frames' depths with a global sharing factor. Similar to the 3D reconstruction evaluation, the scale factor is obtained by the ratio of the median depth value of all frames' depths between predictions and ground truths.

For pose estimation, we follow [6] to evaluate the absolute trajectory error (ATE), relative pose error of rotation (RPE-R) and translation (RPE-T). Before evaluation, the predicted poses are globally aligned with the ground truth.

For camera intrinsics, we evaluate the accuracy with the "FOV AbsRel", which is defined as the absolute relative error of the field of view (FOV AbsRel= $\frac{|\text{FOV}_{pred}-\text{FOV}_{gt}|}{\text{FOV}_{gt}}$).

8. Optimization Details

Frames downsampling. We propose a two-stage frame downsampling strategy to reduce the optimization time complexity. First, all frames $\{\mathbf{I}_i\}_{i=1}^N$ are fed to the LeReS-ResNeXt-101 network and get the backbone's last layer feature (the last layer of $1/32$ stage features) as their embeddings $\{\mathbf{e}_i\}_{i=1}^S \in \mathbb{R}^{N \times C \times H/32 \times W/32}$. The first frame $\mathbf{I}_0$ is selected. We compute the similarity between $\mathbf{e}_0$ and its next neighboring 20 frames $\{\mathbf{e}_i\}_{i=1}^{20}$, and each similarity between two frames is computed by constructing a 4D $H/32 \times W/32 \times H/32 \times W/32$ cosine similarity volume of all pairs of two feature maps (similar to RAFT [7]) and take the maximum value as the image similarity.

If a frame's similarity is just lower than a threshold value σ, then it is selected. We iteratively perform this process to sample several frames coarsely. In the second stage, we will evenly sample 3 frames between the first-stage adjacent samples. All sampled frames $\{\mathbf{I}_i\}_{i=1}^P$ are employed for next keyframes sampling and optimization. We set σ to 0.85.

Hyperparameters. In the local stage, we use PyTorch's AdamW to optimize all learnable parameters. We iterate 2000 steps in total. In each step, we random sample 50 reference frames, and their paired keyframes are sampled based on p_l for optimization. λ_{pc}, λ_{gc}, λ_{regu} are set to 2, 0.5, and 0.01 for indoor scenes and 2, 0.001, and 0.01 for outdoor scenes respectively.

In the global stage, We iterate 4000 steps in total for the global stage. In each step, we randomly sample 50 reference frames and the paired keyframes based on p_g. ϕ is set to $\pi/4$. For indoor scenes, the λ_{pc}, λ_{gc}, λ_{regu} are set to $2, 1, 0.1$ in first 2000 iters and $2, 0.1, 0.1$ in the last 2000 iters. The λ_{gc} is set to 0.001 for outdoor scenes.

Besides, we also filter out the sky regions for outdoor scenes by predicting semantic segmentation with SegFormer-B3 [8] during optimization.

9. Testing Datasets

In our experiments, we perform the evaluation on five zero-shot datasets: NYU [5], ScanNet [2], 7-Scenes [4], TUM [6], and KITTI [3]. The evaluation sequences of five zero-shot datasets are shown in Table S4. Note that we only evaluate the first sequences of 7 scenes on 7-Scenes, and 15, 14, 9, and 6 scenes on NYU, ScanNet, TUM, and KITTI individually.

10. More Qualitative Comparisons

More qualitative comparisons with seven representative algorithms are shown in Fig. S1. Our method can reconstruct accurate and robust 3D scene shapes on diverse scenes.

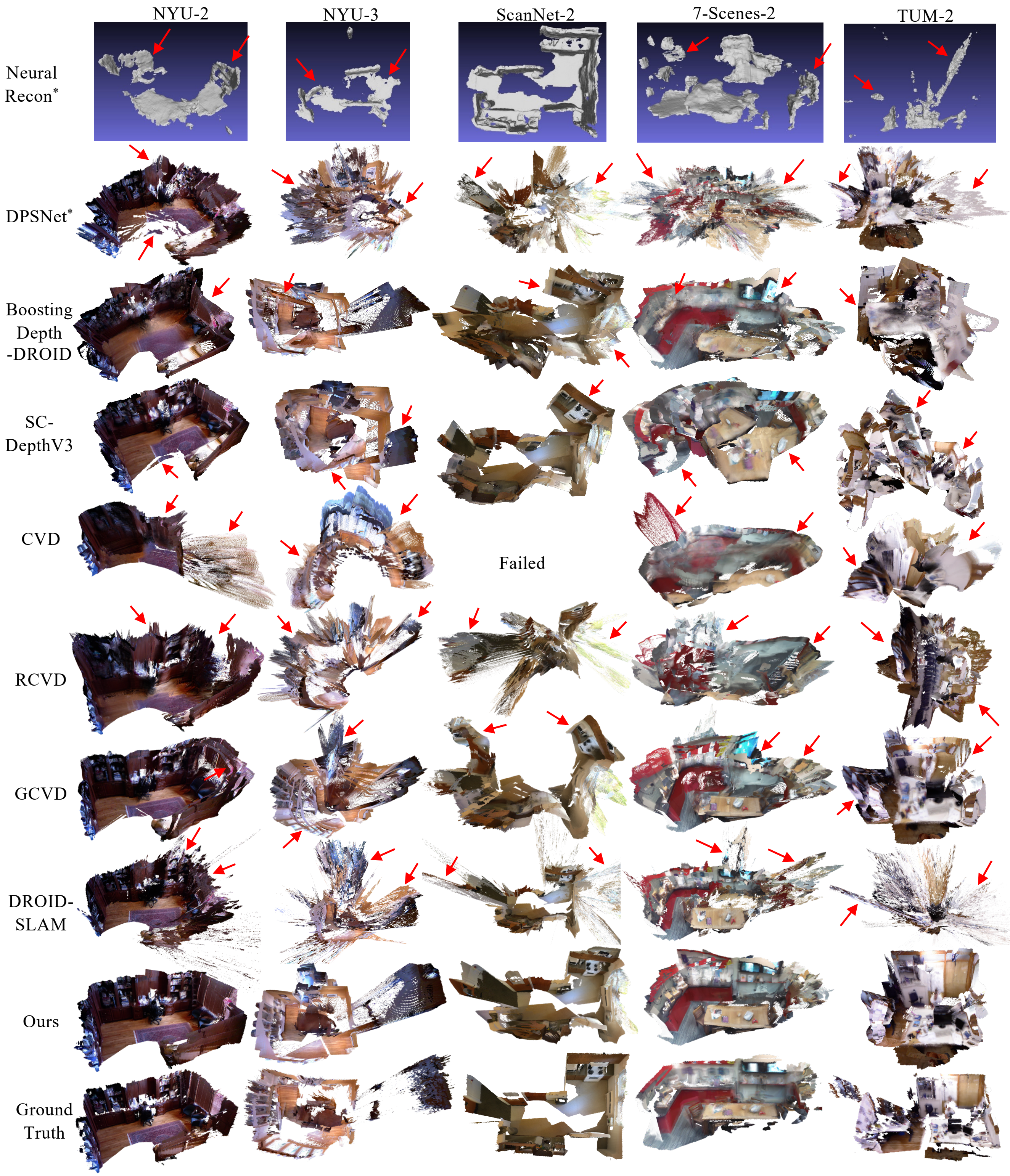

Figure S1: **More qualitative comparisons of zero-shot 3D scene reconstruction.** Note that NeuralRecon is trained on ScanNet [2] and can only output uncolored mesh, and * represents the employment of ground-truth camera poses during reconstruction.

References

[1] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, volume 1611, pages 586–606. Spie, 1992. 3

[2] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 5828–5839, 2017. 3, 4

[3] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *Int. J. Robotics Research*, 32(11):1231–1237, 2013. 3

[4] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in RGB-D images. In *IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 2930–2937, 2013. 3

[5] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Eur. Conf. Comput. Vis.*, pages 746–760. Springer, 2012. 1, 3

[6] Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of RGB-D slam systems. In *IEEE/RSJ Int. Conf. Intelligent Robots and Systems*, pages 573–580, 2012. 3

[7] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Eur. Conf. Comput. Vis.*, pages 402–419. Springer, 2020. 1, 3

[8] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 3

[9] Guangkai Xu, Wei Yin, Hao Chen, Chunhua Shen, Kai Cheng, Feng Wu, and Feng Zhao. Towards 3d scene reconstruction from locally scale-aligned monocular video depth. *arXiv preprint arXiv:2202.01470*, 2022. 1