

MoGe: Unlocking Accurate Monocular Geometry Estimation for Open-Domain Images with Optimal Training Supervision

Supplementary Material

A. Algorithm Details

A.1. Recovering Shift and Camera Focal

We assume a simple pinhole camera model with isotropic focal length and centered principal point. The 2D image plane is parameterized with the center as $(0, 0)$. The image plane coordinate of pixel i is denoted as (u_i, v_i) , corresponding to its predicted 3D point $\mathbf{p}_i = (x_i, y_i, z_i)$. The focal length and shift is obtained by minimizing the projection error,

$$\min_{f, t_z} \sum_{i \in \mathcal{M}} \left(\frac{f x_i}{z_i + t'_z} - u_i \right)^2 + \left(\frac{f y_i}{z_i + t'_z} - v_i \right)^2, \quad (1)$$

which can be further reduced to have a single variable t'_z by substituting f with its close-form solution with respect to t'_z ,

$$f = \frac{\sum_{i \in \mathcal{M}} \left(\frac{x_i}{z_i + t'_z} \right) u_i + \sum_{i \in \mathcal{M}} \left(\frac{y_i}{z_i + t'_z} \right) v_i}{\sum_{i \in \mathcal{M}} \left(\frac{x_i}{z_i + t'_z} \right)^2 + \sum_{i \in \mathcal{M}} \left(\frac{y_i}{z_i + t'_z} \right)^2}. \quad (2)$$

We use a numerical solver for this least squares problem with Levenberg-Marquardt algorithm [19] implemented by SciPy [32] package. For efficiency, the point map is resized to low resolution (64×64) for running this algorithm. In our practice, it typically converges within 10 iterations in around 3ms.

A.2. ROE Alignment

We will first introduce an algorithm to a simpler subproblem then derive the solution to either with or without the constraint of $t_x = t_y = 0$ (1D-shift case or 3D-shift case, respectively).

Subproblem (w/o truncation). Consider the optimization objective with respect to scale s only, denoted as $l_0(s)$. We omit the mask \mathcal{M} for simplicity and denote N as the number of valid points:

$$\min_s l_0(s) = \min_s \sum_{i=1}^N w_i |s \hat{x}_i - x_i|, \quad (3)$$

where $w_i > 0$ and $\hat{x}_i > 0$ without loss of generality. The objective, as a summation of convex functions, is also convex obviously. The minimum occurs where its left-hand $l_0^-(s)$ derivative and right-hand derivative $l_0^+(s)$ have op-

Algorithm 1 ROE alignment subproblem w/o truncation

input: arrays $\hat{X}[1..n]$, $X[1..n]$, $W[1..n]$
output: optimal scale s^* and objective value l^* to Eq. 3

function SOLVESUBPROBLEM(\hat{X} , X , W)
 sort arrays \hat{X} , X , W by $X[i]/\hat{X}[i]$
 $Q[1..n] \leftarrow$ accumulated sum of $W * \hat{X}$
 $D[0..n] \leftarrow \{-Q[n]\} \cup \{2 \cdot Q[i] - Q[n]\}_{i=1}^n$
 $i^* \leftarrow$ the first i s.t. $D[i-1] \leq 0 \leq D[i]$
 $s^* \leftarrow X[i^*]/\hat{X}[i^*]$
 $l^* \leftarrow$ objective function value at s^* .
return s^* , l^* .
end function

posite signs or one of them is zero,

$$l_0^-(s) = \sum_{\frac{x_i}{\hat{x}_i} < s} w_i \hat{x}_i - \sum_{s \leq \frac{x_i}{\hat{x}_i}} w_i \hat{x}_i = 2 \sum_{\frac{x_i}{\hat{x}_i} < s} w_i \hat{x}_i - \sum_{i=1}^N w_i \hat{x}_i,$$

$$l_0^+(s) = \sum_{\frac{x_i}{\hat{x}_i} \leq s} w_i \hat{x}_i - \sum_{s < \frac{x_i}{\hat{x}_i}} w_i \hat{x}_i = 2 \sum_{\frac{x_i}{\hat{x}_i} \leq s} w_i \hat{x}_i - \sum_{i=1}^N w_i \hat{x}_i. \quad (4)$$

$l_0^-(s)$ and $l_0^+(s)$ differ at $\{\frac{x_i}{\hat{x}_i}\}$. First, we sort $\{\frac{x_i}{\hat{x}_i}\}_{i=1}^N$ and compute the prefix summations of $\{w_i \hat{x}_i\}_{i=1}^N$. This allows us to evaluate the derivatives in $O(1)$ time for each point in $\{\frac{x_i}{\hat{x}_i}\}_{i=1}^N$. Finally, $\frac{\hat{x}_i}{x_i}$ such that $l_0^-(\frac{x_i}{\hat{x}_i}) \leq 0 \leq l_0^+(\frac{x_i}{\hat{x}_i})$ is the minimum point. The solution is outlined in Algorithm 1.

Subproblem (w/ truncation). We truncate each residual term to suppress outliers. The truncated objective is

$$\min_s l_1(s) = \min_s \sum_{i=1}^N \min(\tau, w_i |s \hat{x}_i - x_i|), \quad (5)$$

where τ is set to 1 in all our experiments. For each item $l_{1,i}(s) = \min(\tau, w_i |s \hat{x}_i - x_i|)$ in the equation, the one-sided derivatives are

$$l_{1,i}^-(s) = \begin{cases} -w_i \hat{x}_i & \frac{w_i x_i - \tau_i}{w_i \hat{x}_i} < s \leq \frac{x_i}{\hat{x}_i} \\ w_i \hat{x}_i & \frac{x_i}{\hat{x}_i} < s \leq \frac{w_i x_i + \tau_i}{w_i \hat{x}_i} \\ 0 & \text{otherwise} \end{cases}, \quad (6)$$

$$l_{1,i}^+(s) = \begin{cases} -w_i \hat{x}_i & \frac{w_i x_i - \tau_i}{w_i \hat{x}_i} \leq s < \frac{x_i}{\hat{x}_i} \\ w_i \hat{x}_i & \frac{x_i}{\hat{x}_i} \leq s < \frac{w_i x_i + \tau_i}{w_i \hat{x}_i} \\ 0 & \text{otherwise} \end{cases}.$$

Therefore, the one-sided derivatives of $l_0(s)$ are □

$$\begin{aligned}
l_1^-(s) &= \sum_{i=1}^N l_{1,i}^-(s) = \sum_{\frac{x_i}{\hat{x}_i} < s \leq \frac{w_i x_i + \tau}{w_i \hat{x}_i}} w_i \hat{x}_i - \sum_{\frac{w_i x_i - \tau}{w_i \hat{x}_i} < s \leq \frac{x_i}{\hat{x}_i}} w_i \hat{x}_i \\
&= 2 \sum_{\frac{x_i}{\hat{x}_i} < s} w_i \hat{x}_i - \sum_{\frac{w_i x_i - \tau}{w_i \hat{x}_i} < s} w_i \hat{x}_i - \sum_{\frac{w_i x_i + \tau}{w_i \hat{x}_i} < s} w_i \hat{x}_i, \\
&\quad (7) \\
l_1^+(s) &= \sum_{i=1}^N l_{1,i}^+(s) = \sum_{\frac{x_i}{\hat{x}_i} < s < \frac{w_i x_i + \tau}{w_i \hat{x}_i}} w_i \hat{x}_i - \sum_{\frac{w_i x_i - \tau}{w_i \hat{x}_i} \leq s < \frac{x_i}{\hat{x}_i}} w_i \hat{x}_i \\
&= 2 \sum_{\frac{x_i}{\hat{x}_i} \leq s} w_i \hat{x}_i - \sum_{\frac{w_i x_i - \tau}{w_i \hat{x}_i} \leq s} w_i \hat{x}_i - \sum_{\frac{w_i x_i + \tau}{w_i \hat{x}_i} \leq s} w_i \hat{x}_i. \\
&\quad (8)
\end{aligned}$$

Lemma 1 shows that the minimum of the objective function in Eq. 5 is still achieved at one of the points in the set $\{\frac{x_i}{\hat{x}_i}\}_{i=1}^N$, despite the function is non-convex and may contain local minima.

Solving the subproblem requires two steps, as outlined in Algorithm 2. The first step is to identify all extrema in $\{\frac{x_i}{\hat{x}_i}\}_{i=1}^N$ that satisfy $l_0'(\frac{x_i}{\hat{x}_i}) < 0 \leq l_0'(\frac{x_i}{\hat{x}_i})$ by evaluating the derivative values. This can be done efficiently through first binary searching on the sorted arrays $\{\frac{x_i}{\hat{x}_i}\}_{i=1}^N$, $\{\frac{w_i x_i - \tau}{w_i \hat{x}_i}\}_{i=1}^N$ and $\{\frac{w_i x_i + \tau}{w_i \hat{x}_i}\}_{i=1}^N$ and then indexing the prefix summations of $\{w_i \hat{x}_i\}$ in the associated orders. This step has a complexity of $O(N \log N)$. The second step involves computing the objective values at these extrema and determining the minimum, which takes $O(N n_e)$ time, where n_e is the number of extrema. As n_e approximates the number of outliers, it is typically a small constant in practice.

Lemma 1. *There exists at least one pair of (k^*, s^*) such that $s^* \hat{x}_{k^*} - x_{k^*} = 0$ and s^* minimizes Eq. 5.*

Proof. The minimum of $l_1(s)$ must exist, because $l_1(s)$ is continuous, piece-wisely linear and bounded in $[0, N\tau]$.

We first prove that there must exist s^* such that $l_1(s^*) = \min l_1(s)$ and $l_1^+(s^*) > l_1^-(s^*)$. Otherwise, for all s^* such that $l_0(s^*) = \min l_1(s)$, there will be $l_1^+(s^*) = l_1^-(s^*) = 0$, hence the value of $l_1(s)$ in the linear interval where the minimum locates is constant. As a consequence, all neighboring intervals will be constant until the boundary where $\min l_1(s) = l_1(-\infty) = N\tau$, which contradicts the obvious fact that $\min l_1(s) \leq l_1(x_1/\hat{x}_1) < N\tau$.

Given $l_0^+(s^*) > l_0^-(s^*)$, there exists an index k^* such that $s^* = \hat{x}_{k^*}/x_{k^*}$, because

$$\begin{aligned}
0 &< l_1^+(s^*) - l_1^-(s^*) \\
&= 2 \sum_{\frac{x_i}{\hat{x}_i} = s^*} w_i \hat{x}_i - \sum_{\frac{w_i x_i - \tau}{w_i \hat{x}_i} = s^*} w_i \hat{x}_i - \sum_{\frac{w_i x_i + \tau}{w_i \hat{x}_i} = s^*} w_i \hat{x}_i \\
&\leq 2 \sum_{\frac{x_i}{\hat{x}_i} = s^*} w_i \hat{x}_i.
\end{aligned} \quad (9)$$

Algorithm 2 ROE alignment subproblem w/ truncation

input: arrays $\hat{X}[1..n]$, $X[1..n]$, $W[1..n]$, float τ

output: the optimal scale s^* , objective value l^* to Eq. 5

function SOLVESUBPROBLEM(\hat{X} , X , W , τ)

$A[1..n] \leftarrow X/\hat{X}$

$B[1..n] \leftarrow (W * X - \tau)/(W * \hat{X})$

$C[1..n] \leftarrow (W * X + \tau)/(W * \hat{X})$

for each array \mathcal{A} in $\{A, B, C\}$ **do**

 sort \mathcal{A} and obtain sorted indices $I_{\mathcal{A}}[1..n]$

$Q_{\mathcal{A}}[1..n] \leftarrow$ accumulated sum of $\{W \hat{X}[I_{\mathcal{A}}[i]]\}_{i=1}^n$

end for

Initialize I_E as empty set

for $i = 1$ to n **do**

\triangleright parallel computation

for each array \mathcal{A} in $\{A, B, C\}$ **do**

$j_{\mathcal{A}}^- \leftarrow$ the last j s.t. $\mathcal{A}[j] < X[i]/\hat{X}[i]$

$j_{\mathcal{A}}^+ \leftarrow$ the last j s.t. $\mathcal{A}[j] \leq X[i]/\hat{X}[i]$

end for

$d^- \leftarrow 2 \cdot Q_A[j_{\mathcal{A}}^-] - Q_B[j_{\mathcal{A}}^-] - Q_C[j_{\mathcal{A}}^-]$

$d^+ \leftarrow 2 \cdot Q_A[j_{\mathcal{A}}^+] - Q_B[j_{\mathcal{A}}^+] - Q_C[j_{\mathcal{A}}^+]$

if $d^- < 0 \leq d^+$ **then** append i to I_E

end for

Initialize $l[1..N]$ with ∞

for i in I_E **do**

\triangleright parallel computation

$s \leftarrow X[i]/\hat{X}[i]$

$l[i] \leftarrow$ objective function value at s

end for

$i^* \leftarrow$ index of the minimum in $l[i], i \in I_E$

$s^* \leftarrow X[i^*]/\hat{X}[i^*], l^* \leftarrow l[i^*]$

return s^*, l^*

end function

Alignment with 1D shift. Recall the alignment objective and let w_i be $1/z_i$. We rewrite it as follows:

$$\min_{s, t_z} \sum_{i=1}^n [w_i |s \hat{x}_i - x_i| + w_i |s \hat{y}_i - y_i| + w_i |s \hat{z}_i + t_z - z_i|], \quad (10)$$

or apply truncation to each absolute residual term

$$\begin{aligned}
\min_{s, t_z} \sum_{i=1}^n &[\min(\tau, w_i |s \hat{x}_i - x_i|) + \min(\tau, w_i |s \hat{y}_i - y_i|) \\
&+ \min(\tau, w_i |s \hat{z}_i + t_z - z_i|)].
\end{aligned} \quad (11)$$

The proposed solution is outlined in Algorithm 3, with proof as follows. The corresponding subproblem solver is selected based on whether truncation is applied.

Lemma 2. *There exists at least one triplet of (k^*, s^*, t_z^*) such that $s^* \hat{z}_{k^*} + t_z^* - z_{k^*} = 0$ and (s^*, t_z^*) minimizes the objective of Equation 11.*

Proof. Denote the objective as $l_2(s, t_z)$,

$$l_2(s, t_z) = \sum_{i=1}^n [\min(\tau, w_i |s\hat{x}_i - x_i|) + \min(\tau, w_i |s\hat{y}_i - y_i|)] \\ + \sum_{i=1}^N \min(\tau, w_i |t_z - (z_i - s\hat{z}_i)|)]. \quad (12)$$

Given arbitrary s , using Lemma 1, there exists at least one pair (t_z, k) such that $t_z - (z_k - s\hat{z}_k) = 0$ and t_z minimizes $\sum_{i=1}^n \min(\tau, w_i |s\hat{z}_i + t_z - z_i|)$, hence minimizes $l_2(s, t_z)$ as the rest parts are constant with regard to t_z . Therefore, a solution s^* is always associated with corresponding (t_z^*, k^*) such that that $s^* \hat{z}_{k^*} + t_z^* - z_{k^*} = 0$. \square

Lemma 2 allows us to reduce Eq. 11 to the subproblem with respect to some index k . For each possible index k , the objective is formed as:

$$\min_s \sum_{i=1}^n \min(\tau, w_i |s\hat{x}_i - sx_i|) + \min(\tau, w_i |s\hat{y}_i - y_i|) \\ + \min(\tau, w_i |s(\hat{z}_i - \hat{z}_k) - (z_i - z_k)|), \quad (13)$$

which is solvable in $O(N \log N)$ complexity. We enumerate all possible indices for k and find the minimum. Therefore, the total time complexity is $O(N^2 \log N)$.

In our implementation, the point map is resized to low resolution (64×64) for alignment, with $N = 4096$ at most. The algorithm is further parallelized with tensor operations on GPUs.

Algorithm 3 ROE alignment w/ 1-D shift

input: point arrays $\hat{P}[1..N, 1..3]$, $P[1..N, 1..3]$,
weight array $W[1..N]$
output: the optimal scale s^* , shift t_z^* ,
objective value l^* to Eq. 10 or Eq. 11

$W[1..3N] \leftarrow$ repeat each element in W 3 times
Initialize arrays $s[1..N]$, $l[1..N]$, $t_z[1..N]$
for $k = 1$ to N **do** \triangleright parallel computation
 $\hat{X}[1..3N] \leftarrow \text{FLATTEN}(\hat{P}[1..N, 1..3] - \{0, 0, \hat{P}[k, 3]\})$
 $X[1..3N] \leftarrow \text{FLATTEN}(P[1..N, 1..3] - \{0, 0, P[k, 3]\})$
 $(s[k], l[k]) \leftarrow \text{SOLVESUBPROBLEM}(\hat{X}, X, W)$
 $t_z[k] \leftarrow P[k, 3] - s[k] \cdot \hat{P}[k, 3]$
end for
 $k^* \leftarrow$ index of the minimum in $l[1..N]$
 $s^* \leftarrow s[k^*]$, $l^* \leftarrow l[k^*]$, $t_z^* \leftarrow t_z[k^*]$
return s^* , t_z^* , l^*

Alignment with 3D shift. We apply truncation and rewrite the objective as follows:

$$\min_{s, t_z} \sum_{i=1}^n \min(\tau, w_i |s\hat{x}_i + t_x - x_i|) \\ + \min(\tau, w_i |s\hat{y}_i + t_y - y_i|) \\ + \min(\tau, w_i |s\hat{z}_i + t_z - z_i|). \quad (14)$$

Similarly to the proof of Lemma 2, there exists at least one group $(k_1^*, k_2^*, k_3^*, s^*, t^*)$ such that $s^* \hat{x}_{k_1^*} + t_x^* - x_{k_1^*} = 0$, $s^* \hat{y}_{k_2^*} + t_y^* - y_{k_2^*} = 0$, $s^* \hat{z}_{k_3^*} + t_z^* - z_{k_3^*} = 0$, and (s^*, t^*) minimizes the objective. However, the $O(N^4 \log N)$ time complexity of a brute-force search is prohibitive. Motivated by the strong locality of surface points within a 3D sphere, we introduce a reasonable assumption, $k_1 = k_2 = k_3$, to obtain an approximately optimal solution with $O(N^2 \log N)$ complexity. This assumption posits that the predicted and ground truth patches can be well aligned under the condition that one corresponding pair of points coincides. The effectiveness of the approximated solution has been empirically validated.

Algorithm 4 ROE alignment w/ 3-D shift

input: point arrays $\hat{P}[1..N, 1..3]$, $P[1..N, 1..3]$,
weight array $W[1..N]$
output: the optimal scale s^* , shift t^* ,
objective value l^* to Eq. 14

$W[1..3N] \leftarrow$ repeat each element in W 3 times
Initialize arrays $s[1..N]$, $l[1..N]$, $t[1..N, 3]$
for $k = 1$ to N **do** \triangleright parallel computation
 $\hat{X}[1..3N] \leftarrow \text{FLATTEN}(\hat{P}[1..N, 1..3] - \hat{P}[k, 1..3])$
 $X[1..3N] \leftarrow \text{FLATTEN}(P[1..N, 1..3] - P[k, 1..3])$
 $(s[k], l[k]) \leftarrow \text{SOLVESUBPROBLEM}(\hat{X}, X, W)$
 $t[k] \leftarrow P[k] - s[k] \cdot \hat{P}[k]$
end for
 $k^* \leftarrow$ index of the minimum in $l[1..N]$
 $s^* \leftarrow s[k^*]$, $l^* \leftarrow l[k^*]$, $t^* \leftarrow t[k^*]$
return s^* , t^* , l^*

B. Experiment Details

B.1. Training Data

The datasets used in our training are listed in Table 1. The number of frames may slightly differ from that of the original data because some invalid frames are dropped.

To assign balanced weights to the datasets for training, we compute the retrieval probability of each dataset relative to OpenImagesV7 [15], a large and diverse natural image dataset. Specifically, we leverage DINOv2 [22] to extract feature vectors and calculate the probability that the nearest neighbor of a randomly selected image from OpenImagesV7 is found in each respective training dataset.

| Name | Domain | #Frames | Type | Weight |
|------------------|---------------------|---------|------|--------|
| A2D2[9] | Outdoor/Driving | 196K | C | 0.8% |
| Argoverse2[37] | Outdoor/Driving | 1.1M | C | 7.4% |
| ARKitScenes[2] | Indoor | 449K | B | 8.6% |
| DIML-indoor[5] | Indoor | 894K | D | 4.8% |
| BlendedMVS[39] | In-the-wild | 115K | B | 12.0% |
| MegaDepth[17] | Outdoor/In-the-wild | 92K | B | 5.6% |
| Taskonomy[41] | Indoor | 3.6M | B | 14.1% |
| Waymo[28] | Outdoor/Driving | 788K | C | 6.4% |
| GTA-SfM[33] | Outdoor/In-the-wild | 19K | A | 2.8% |
| Hypersim[25] | Indoor | 75K | A | 5.0% |
| IRS[34] | Indoor | 101K | A | 5.6% |
| KenBurns[21] | In-the-wild | 76K | A | 1.6% |
| MatrixCity[16] | Outdoor/Driving | 390K | A | 1.3% |
| MidAir[8] | Outdoor/In-the-wild | 423K | A | 4.0% |
| MVS-Synth[13] | Outdoor/Driving | 12K | A | 1.2% |
| Spring[18] | In-the-wild | 5K | A | 0.7% |
| Structured3D[42] | Indoor | 77K | A | 4.8% |
| Synthia[26] | Outdoor/Driving | 96K | A | 1.2% |
| TartanAir[36] | In-the-wild | 306K | A | 5.0% |
| UrbanSyn[11] | Outdoor/Driving | 7K | A | 2.1% |
| ObjaverseV1[4] | Object | 167K | A | 4.8% |

| Type | Label quality | | | Applied losses | | | | | |
|-----------------|---------------|-------------|---------------|-----------------|---------------------|---------------------|---------------------|-----------------|-----------------|
| | Accuracy | Range | Density | \mathcal{L}_G | \mathcal{L}_{S_1} | \mathcal{L}_{S_2} | \mathcal{L}_{S_3} | \mathcal{L}_N | \mathcal{L}_M |
| A. Synthetic | Perfect | ∞ | Dense | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| B. SfM/MV Recon | High | ∞ | Dense&Partial | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| C. LiDAR/Laser | High | $\sim 100m$ | Sparse | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| D. Kinect | Medium | $\sim 10m$ | Dense | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

Table 1. Datasets used for training and tailored loss combination.

B.2. Evaluation Data

The raw evaluation datasets are processed accordingly for reliable evaluation and fair comparison. We report the details as follows.

- **NYUv2 [20]**. We use the official test split of 654 samples. Due to the inaccuracy of ground truth values captured by Kinect V1 near boundaries, we filter and remove boundary regions by a simple edge detection method. Specifically, we set a threshold for the difference between the minimum and maximum depth values within a local window. Depth values beyond 5 meters are excluded because they are unreliable due to the limited sensor range [29]. Additionally, we manually mask out areas with reflective and transparent surfaces, such as mirrors and glass, which cannot be accurately captured by the sensor.
- **KITTI [30]**. We utilize the test split of 652 images of Eigen *et al.* [7] following previous works. The original resolution of 1242×375 does not match our training aspect ratio (ranging from 1 : 2 to 2 : 1), so we apply center cropping to obtain a resolution of 750×375 from the raw images.
- **ETH3D [27]**. All 454 images are included. The images are undistorted with the official calibration data and downsized from the original resolution of 6202×4135 to 2048×1365 .
- **iBims-1 [14]**. All 100 images are included at an original

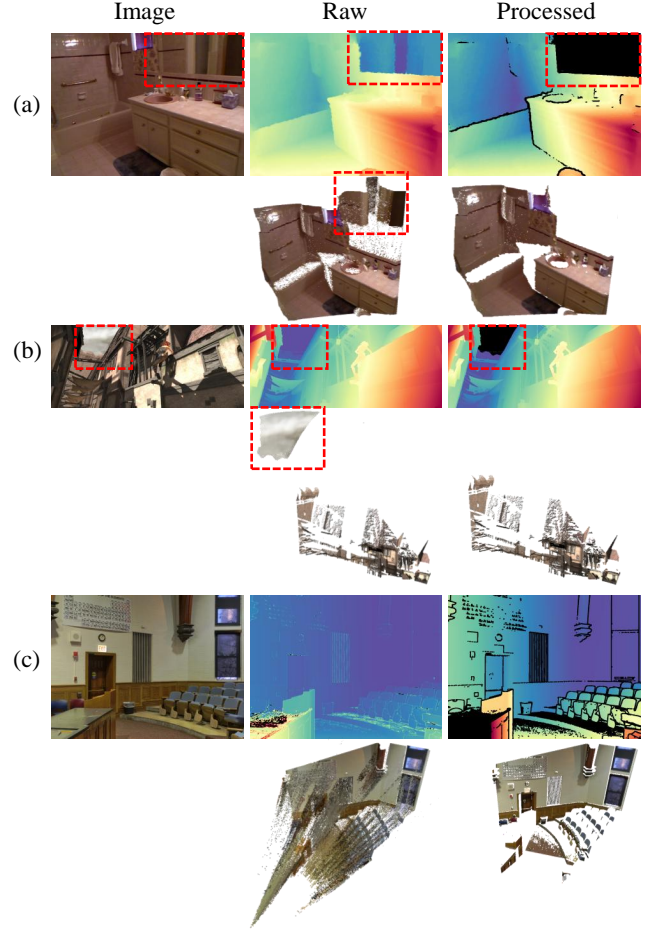


Figure 1. Examples of evaluation data preprocessing: (a) Removing mirror and boundary artifacts from the ground truth depth in NYUv2. (b) Excluding sky regions in Sintel. (c) Removing boundary artifacts from the ground truth depth in DIODE.

resolution of 640×480 .

- **GSO [6]**. The dataset contains 1,030 objects. For each object, we render a single view at 512×512 resolution. The view is randomly sampled with a FOV ranging from 30° to 60° . The object is centered in the image, and its bounding box occupies approximately 70% of the image's size.
- **Sintel [3]**. We use all 1,064 frames and center-crop the images to 872×436 from the original 1024×436 resolution to fit our aspect ratio range. The sky regions are manually masked out because evaluating models with sky depth included is not meaningful.
- **DDAD [10]**. We randomly select 1,000 samples from the validation set. The dataset was collected using multiple cameras and LiDAR sensors mounted on a moving vehicle. Some cameras inadvertently capture parts of the vehicle, causing discrepancies with the sensor's depth data. To address this issue, we crop the regions that are not obstructed by the vehicle itself.
- **DIODE [31]**. We utilize the official validation split,

which includes 325 indoor images and 446 outdoor images at an original resolution of 1024×768 . Due to artifacts in ground truth depth values near the boundaries in this dataset, we identify and remove these boundary regions using a similar approach as described above.

B.3. Evaluation Protocol

For all our models and baselines, predictions and ground truth are aligned in scale (and shift, if applicable) for each image before measuring errors. To clarify the notations in this section:

- $\hat{\mathbf{p}}_i$ and \mathbf{p}_i are the predicted and ground-truth points, respectively.
- \hat{z}_i and z_i are the predicted and ground-truth depths, which are the Z -coordinate of corresponding points.
- \mathcal{M} is the mask of valid ground-truth.
- a and b denote the scale and shift used to align predictions with the ground truth for evaluation, to avoid confusion with similar symbols used in the training objectives.
- **Scale-invariant point map.** The scale a^* to align prediction with ground truth is computed as:

$$a^* = \operatorname{argmin}_a \sum_{i \in \mathcal{M}} \frac{1}{z_i} \|a\hat{\mathbf{p}}_i - \mathbf{p}_i\|_1, \quad (15)$$

- **Affine-invariant point map.** The scale a^* and shift b^* are computed as:

$$(a^*, b^*) = \operatorname{argmin}_{a, b} \sum_{i \in \mathcal{M}} \frac{1}{z_i} \|a\hat{\mathbf{p}}_i + b - \mathbf{p}_i\|_1. \quad (16)$$

- **Scale-invariant depth map,** the scale a^* is computed as

$$a^* = \operatorname{argmin}_s \sum_{i \in \mathcal{M}} \frac{1}{z_i} |a\hat{z}_i - z_i|. \quad (17)$$

- **Affine-invariant depth map.** The scale a^* and shift b^* are computed as

$$(a^*, b^*) = \operatorname{argmin}_s \sum_{i \in \mathcal{M}} \frac{1}{z_i} |a\hat{z}_i + b - z_i|. \quad (18)$$

- **Affine-invariant disparity map.** We follow the established protocol for affine disparity alignment [24], using least-squares to align predictions in disparity space:

$$(a^*, b^*) = \operatorname{argmin}_s \sum_{i \in \mathcal{M}} (a\hat{d}_i + b - d_i)^2, \quad (19)$$

where \hat{d}_i is the predicted disparity and d_i is the ground truth, defined as $d_i = 1/z_i$. To prevent aligned disparities from taking excessively small or negative values, the aligned disparity is truncated by the inverted maximum depth $1/z_{\max}$ before inversion. The final aligned depth \hat{z}_i^* is computed as:

$$\hat{z}_i^* := \frac{1}{\max(a^*\hat{d}_i + b^*, 1/z_{\max})}. \quad (20)$$

C. More Results

Full table of depth estimation results In Table 2, we present detailed results for depth estimation where methods that predict metric or scale-invariant depth are also evaluated on affine-invariant depth and disparity for a fair comparison.

More qualitative comparisons Fig. 3 and Fig. 4 present additional visual comparisons on zero-shot evaluation datasets and in-the-wild images. Our method is compared with LeReS [40], UniDepth [23], DUS3R [35], Metric3D V2 [12] and Depth Anything V2 [38]. Since Metric3D V2 and Depth Anything V2 predict depth map and require ground truth camera focal to obtain 3D points cloud results, we visualize them using our estimated focal lengths.

In the supplementary videos, we present *extensive and uncurated comparisons* using the first 100 images from the DIV2K[1] dataset.

More visual results In Fig. 5 and Fig. 6, we demonstrate more reconstruction results of our method for more open-domain images.

D. Limitations and Future Work

While our model demonstrates strong performance, accurately capturing thin structures remains a significant challenge. This difficulty arises from the network’s limited capacity and the presence of noisy real-world training data. As illustrated in Fig. 2, our model may fail to recover these intricate structures.

Additionally, while monocular video reconstruction holds great promise as an application, achieving temporal consistency presents substantial challenges. Our model, designed for single-image input, cannot inherently maintain temporal coherence due to the ambiguity of the task. Addressing this issue would require non-trivial solutions, such as global optimization techniques. Given the rapid advancements in video depth estimation, we believe that an end-to-end model for monocular video reconstruction could significantly benefit from our proposed techniques. Exploring this direction is a compelling avenue for future work.

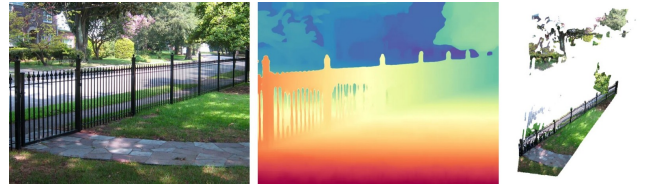


Figure 2. A failure case. Our model fails to capture the thin structure of the fence, leading to a flattened geometry.

| Method | NYUv2 | | KITTI | | ETH3D | | iBims-1 | | GSO | | Sintel | | DDAD | | DIODE | | Average | | |
|----------------------------|--------------------|----------------|-------------|----------------|--------------------|----------------|-------------|----------------|--------------------|----------------|--------------------|----------------|--------------------|----------------|--------------------|----------------|--------------------|----------------|-------------|
| | Rel ^d ↓ | δ_1^d ↑ | Rel.↓ | δ_1^d ↑ | Rel ^d ↓ | δ_1^d ↑ | Rel.↓ | δ_1^d ↑ | Rel ^d ↓ | δ_1^d ↑ | Rel ^d ↓ | δ_1^d ↑ | Rel ^d ↓ | δ_1^d ↑ | Rel ^d ↓ | δ_1^d ↑ | Rel ^d ↓ | δ_1^d ↑ | Rank↓ |
| Scale-invariant depth | | | | | | | | | | | | | | | | | | | |
| LeReS | 12.1 | 82.6 | 19.2 | 64.8 | 14.2 | 78.4 | 14.0 | 78.8 | 13.6 | 77.9 | 30.5 | 52.1 | 26.5 | 52.0 | 18.2 | 69.6 | 18.5 | 69.5 | 7.31 |
| ZoeDepth | <u>5.62</u> | <u>96.3</u> | <u>7.27</u> | <u>91.9</u> | 10.4 | 87.3 | 7.45 | 93.2 | 3.23 | <u>99.9</u> | 27.4 | 61.8 | 17.0 | 72.8 | 11.3 | 85.2 | <u>11.2</u> | <u>86.1</u> | 5.50 |
| DUST3R | 4.40 | 97.1 | 7.81 | 90.6 | 6.04 | 95.7 | 4.98 | 95.8 | 3.27 | <u>99.5</u> | 31.1 | 57.2 | 18.6 | 73.3 | 8.91 | 88.8 | 10.6 | 87.2 | 5.00 |
| Metric3D V2 | 4.69 | 97.4 | 4.00 | <u>98.5</u> | <u>3.84</u> | <u>98.5</u> | <u>4.23</u> | <u>97.7</u> | <u>2.46</u> | <u>99.9</u> | <u>20.7</u> | <u>69.8</u> | <u>7.41</u> | <u>94.6</u> | 3.29 | 98.4 | <u>6.33</u> | <u>94.3</u> | <u>2.07</u> |
| UniDepth | <u>3.86</u> | 98.4 | 3.73 | 98.6 | 5.67 | 97.0 | 4.79 | 97.4 | 4.18 | <u>99.7</u> | 28.3 | 58.8 | <u>10.1</u> | 90.5 | 6.83 | 92.8 | <u>8.43</u> | <u>91.6</u> | 3.00 |
| DA V1 | <u>4.77</u> | <u>97.5</u> | <u>5.61</u> | <u>95.6</u> | 9.41 | 88.9 | 5.53 | 95.8 | 5.49 | 99.3 | 28.3 | 56.7 | 13.2 | 81.5 | 10.3 | 87.5 | <u>10.3</u> | <u>87.9</u> | 5.67 |
| -metric indoor | <u>4.77</u> | <u>97.5</u> | 15.4 | 73.6 | 9.41 | 88.9 | 5.53 | 95.8 | 5.49 | 99.3 | 28.3 | 56.7 | 24.2 | 57.4 | 10.3 | 87.5 | 12.9 | 82.1 | - |
| -metric outdoor | 15.9 | 72.3 | <u>5.61</u> | <u>95.6</u> | 8.77 | 92.4 | 13.8 | 78.8 | 8.59 | 93.6 | 28.1 | 54.8 | 13.2 | 81.5 | 13.0 | 81.4 | <u>13.4</u> | <u>81.3</u> | - |
| DA V2 | 5.03 | 97.3 | 7.23 | 93.7 | 6.12 | 95.5 | 4.32 | 97.9 | 4.38 | 99.3 | 23.0 | 65.2 | 14.7 | 78.0 | 7.95 | 90.0 | 9.09 | 89.6 | 4.06 |
| -metric indoor | 5.03 | 97.3 | 7.61 | 90.9 | 6.12 | 95.5 | 4.32 | 97.9 | 4.38 | 99.3 | 23.0 | 65.2 | 16.6 | 73.4 | 7.95 | 90.0 | 9.38 | 88.7 | - |
| -metric outdoor | 15.3 | 72.3 | 7.23 | 93.7 | 9.30 | 89.6 | 10.6 | 84.9 | 9.62 | 92.5 | 28.6 | 57.3 | 14.7 | 78.0 | 12.2 | 83.2 | 13.4 | 81.4 | - |
| Ours | 3.44 | 98.4 | 4.25 | 97.8 | 3.36 | 98.9 | 3.46 | 97.0 | 1.47 | 100 | 19.3 | 73.4 | 9.17 | 90.5 | <u>4.89</u> | <u>94.7</u> | 6.17 | 93.8 | 1.62 |
| Affine-invariant depth | | | | | | | | | | | | | | | | | | | |
| LeReS | 6.21 | 95.4 | 8.28 | 90.3 | 8.95 | 90.8 | 6.68 | 94.5 | 4.03 | 99.4 | 24.0 | 64.8 | 16.2 | 75.8 | 9.99 | 88.1 | 10.5 | 87.4 | 8.81 |
| ZoeDepth | <u>4.76</u> | <u>97.3</u> | <u>5.59</u> | <u>95.1</u> | 7.27 | 94.2 | 5.85 | 95.7 | 2.54 | 99.9 | 21.8 | 69.2 | 14.2 | 80.1 | 7.80 | 90.9 | <u>8.73</u> | <u>90.3</u> | 7.33 |
| DUST3R | 3.73 | 97.8 | 7.30 | 91.6 | 4.96 | 96.4 | 3.94 | 96.6 | 2.55 | 99.6 | 25.4 | 64.2 | 16.9 | 76.2 | 6.68 | 92.6 | 8.93 | 89.4 | 6.62 |
| Metric3D V2 | 3.94 | 97.6 | 3.50 | <u>98.4</u> | <u>3.24</u> | <u>99.0</u> | <u>3.28</u> | 98.3 | 2.10 | 99.4 | 26.6 | 71.7 | <u>7.15</u> | <u>94.8</u> | 2.75 | 98.7 | <u>6.57</u> | <u>94.7</u> | 3.64 |
| UniDepth V1 | <u>3.40</u> | 98.6 | <u>3.55</u> | 98.7 | 4.92 | 97.5 | 3.76 | 98.2 | 2.48 | 99.9 | 24.9 | 64.1 | <u>9.46</u> | <u>90.8</u> | 4.90 | 96.2 | 7.17 | <u>93.0</u> | 3.62 |
| Marigold | 4.63 | 97.3 | 7.29 | 93.8 | 6.08 | 96.3 | 4.35 | 97.2 | 2.78 | 99.9 | 21.2 | 75.0 | 14.6 | 80.5 | 6.34 | 94.3 | 8.41 | 91.8 | 5.69 |
| Geowizard | 4.69 | 97.4 | 8.14 | 92.5 | 6.90 | 93.9 | 4.50 | 97.1 | 2.00 | 99.9 | 17.8 | 76.2 | 16.5 | 75.7 | 7.03 | 92.7 | 8.45 | 90.7 | 6.44 |
| DA V1 | <u>3.82</u> | <u>98.3</u> | <u>5.04</u> | <u>96.4</u> | 6.23 | 95.2 | 4.23 | 97.3 | 1.98 | 100 | 20.1 | 71.8 | 11.3 | 86.1 | 6.75 | 92.6 | <u>7.43</u> | <u>92.2</u> | 4.83 |
| -metric indoor | <u>3.82</u> | <u>98.3</u> | 9.95 | 86.5 | 6.23 | 95.2 | 4.23 | 97.3 | 1.98 | 100 | 20.1 | 71.8 | 17.0 | 74.0 | 6.75 | 92.6 | 8.76 | 89.5 | - |
| -metric outdoor | 7.68 | 93.8 | <u>5.04</u> | <u>96.4</u> | 6.21 | 96.6 | 7.00 | 94.2 | 2.77 | 99.8 | 20.6 | 70.0 | 11.3 | 86.1 | 7.03 | 93.2 | <u>8.45</u> | <u>91.3</u> | - |
| DA V2 | 4.16 | 97.9 | 6.77 | 94.3 | 4.63 | 97.2 | 3.44 | 98.3 | <u>1.44</u> | 100 | <u>17.1</u> | <u>76.6</u> | 13.4 | 81.8 | 5.41 | 94.6 | <u>7.04</u> | <u>92.6</u> | - |
| -metric indoor | 4.16 | 97.9 | 7.09 | 92.3 | 4.63 | 97.2 | 3.44 | 98.3 | 1.44 | 100 | 17.1 | 76.6 | 14.3 | 79.8 | 5.41 | 94.6 | 7.20 | 92.1 | - |
| -metric outdoor | 8.65 | 91.0 | 6.77 | 94.3 | 7.24 | 93.5 | 6.80 | 93.5 | 2.29 | 100 | 22.4 | 67.1 | 13.4 | 81.8 | 8.19 | 90.7 | 9.47 | 89.0 | 2.94 |
| Ours | 2.92 | 98.6 | 3.94 | 98.0 | 2.69 | 99.2 | 2.74 | 97.9 | 0.94 | 100 | 13.0 | 83.2 | 8.40 | 92.1 | <u>3.16</u> | <u>97.5</u> | 4.72 | 95.8 | 1.56 |
| Affine-invariant disparity | | | | | | | | | | | | | | | | | | | |
| LeReS | 7.31 | 95.5 | 12.2 | 87.1 | 10.2 | 90.1 | 8.44 | 92.9 | 4.33 | 99.7 | 28.9 | 59.6 | 23.4 | 73.0 | 10.7 | 88.3 | 13.2 | 85.8 | 8.25 |
| ZoeDepth | <u>5.21</u> | <u>97.7</u> | <u>5.84</u> | <u>95.6</u> | 8.07 | 94.0 | 6.19 | 96.1 | 2.60 | 99.9 | 26.9 | 66.3 | 14.1 | 81.7 | 8.17 | 92.0 | <u>9.63</u> | <u>90.4</u> | 6.75 |
| DUST3R | 4.24 | 98.1 | 7.72 | 92.1 | 5.60 | 96.2 | 4.49 | 96.6 | 2.63 | 99.8 | 40.0 | 56.7 | 17.4 | 76.2 | 7.10 | 92.8 | 11.1 | 88.6 | 6.75 |
| Metric3D V2 | 13.4 | 81.5 | <u>3.76</u> | <u>98.2</u> | <u>4.30</u> | <u>97.7</u> | 8.55 | 92.3 | 1.80 | 100 | 21.8 | 72.4 | <u>7.35</u> | <u>94.1</u> | 7.70 | 90.2 | <u>8.58</u> | <u>90.8</u> | 5.29 |
| UniDepth V1 | <u>3.78</u> | 98.7 | 3.64 | 98.7 | 5.34 | 97.2 | 4.06 | <u>98.1</u> | 2.56 | 99.9 | 28.6 | 60.7 | <u>9.94</u> | <u>89.1</u> | 5.95 | 95.5 | 7.98 | 92.2 | 3.62 |
| MiDaS V3.1 | 4.58 | 98.1 | 6.25 | 94.7 | 5.77 | 96.8 | 4.73 | 97.4 | 1.86 | 100 | 21.3 | 73.1 | 14.5 | 82.6 | 6.05 | 94.9 | 8.13 | 92.2 | 5.00 |
| DA V1 | 4.20 | 98.4 | 5.40 | 97.0 | 4.68 | <u>98.2</u> | 4.18 | 97.6 | 1.54 | 100 | <u>20.1</u> | <u>77.6</u> | 12.7 | 86.9 | 5.69 | 95.7 | <u>7.31</u> | <u>93.9</u> | <u>3.00</u> |
| DA V2 | 4.14 | 98.3 | 5.61 | 96.7 | 4.71 | 97.9 | <u>3.47</u> | 98.5 | <u>1.24</u> | 100 | 21.4 | 72.8 | 13.1 | 86.4 | <u>5.29</u> | <u>96.1</u> | 7.37 | 93.3 | 3.12 |
| Ours | 3.38 | <u>98.6</u> | 4.05 | 98.1 | 3.11 | 98.9 | 3.23 | 98.0 | 0.96 | 100 | 18.4 | 79.5 | 8.99 | 91.5 | 3.98 | 97.2 | 5.76 | 95.2 | 1.44 |

Table 2. Full table of comparison for depth map estimation. * methods have multiple model versions available for respective benchmarks, among which the best for each benchmark is chosen for ranking, followed by the detailed results in smaller text size for each version. Gray numbers denote models trained on respective benchmarks.



Figure 3. Additional qualitative comparisons from the *evaluation datasets*. *: for methods without camera intrinsics prediction, ground-truth camera intrinsics (and disparity shifts) were used to lift their results into 3D points. **Best viewed with zoom.**



Figure 4. Additional qualitative comparisons for *in-the-wild* images from the DIV2K dataset. *: for methods without camera intrinsics prediction, our camera intrinsics prediction were used to lift their results into 3D points. Supplementary videos contain more results. **Best viewed with zoom.**

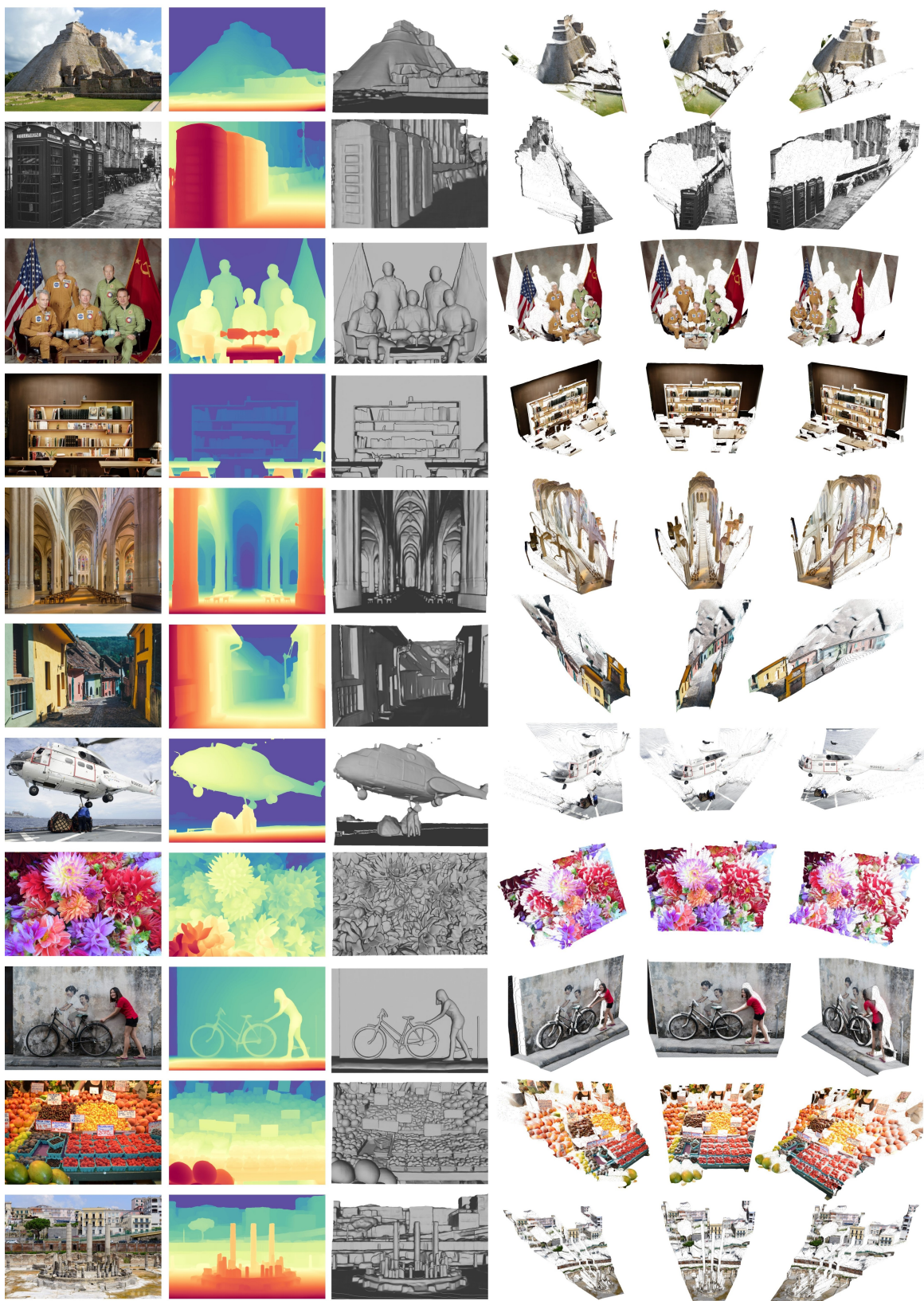


Figure 5. Additional visual results for open-domain images of our model (page 1 of 2). The columns from left to right are the input images, reconstructed disparity maps, reconstructed surface geometry viewed from the source view, and three novel-view images, respectively.

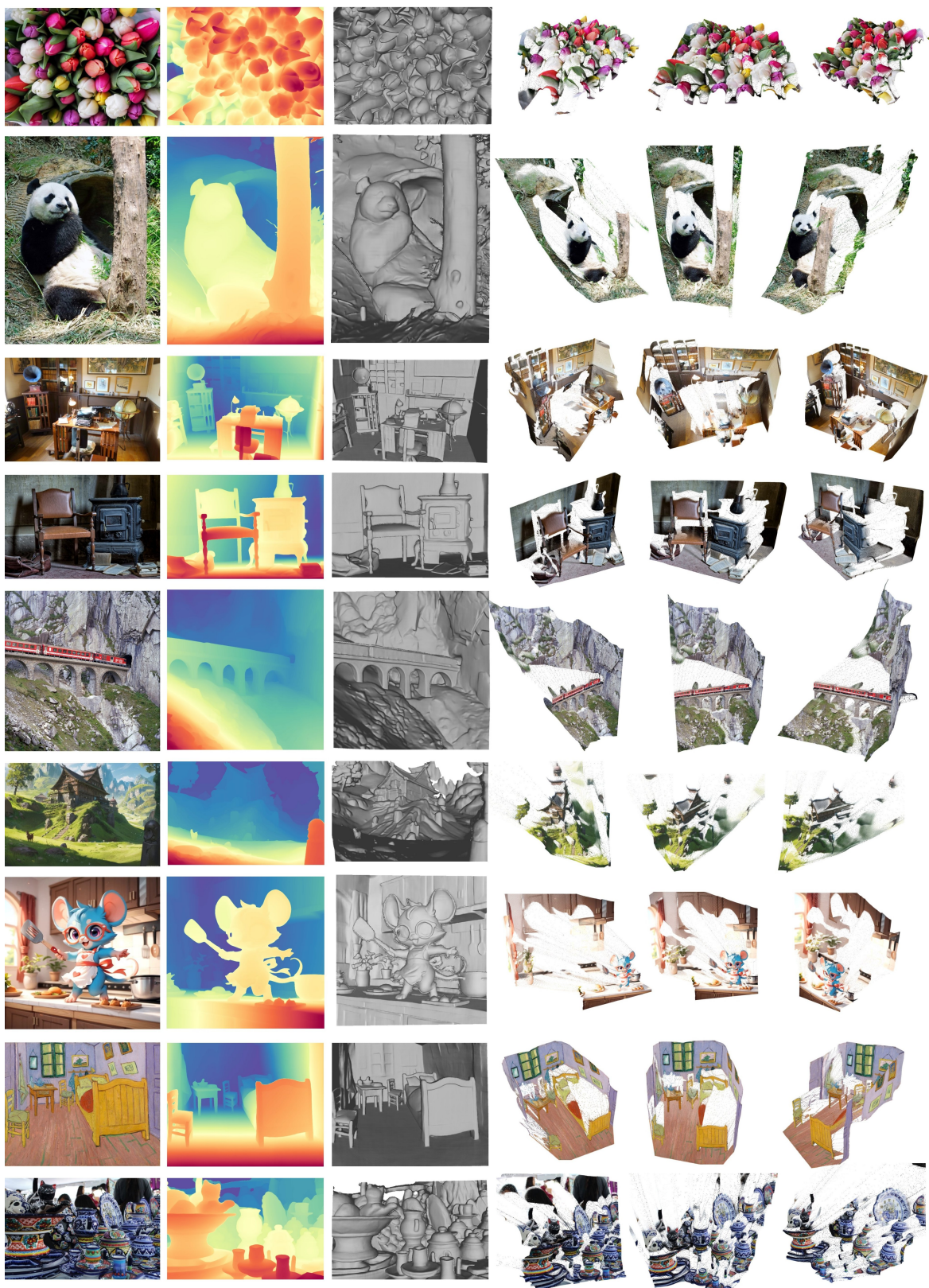


Figure 6. Additional visual results for open-domain images of our model (page 2 of 2). The columns from left to right are the input images, reconstructed disparity maps, reconstructed surface geometry viewed from the source view, and three novel-view images, respectively.

References

- [1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017. 5
- [2] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. ARK-itscenes - a diverse real-world dataset for 3d indoor scene understanding using mobile RGB-d data. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. 4
- [3] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *European Conf. on Computer Vision (ECCV)*, pages 611–625. Springer-Verlag, 2012. 4
- [4] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 4
- [5] Digital Image Media Laboratory (DIML) and Computer Vision Laboratory (CVL). Diml/cvl rgb-d dataset: 2m rgb-d images of natural indoor and outdoor scenes. https://dimlrgbd.github.io/downloads/technical_report.pdf. 4
- [6] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B. McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items, 2022. 4
- [7] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014. 4
- [8] Michael Fonder and Marc Van Droogenbroeck. Mid-air: A multi-modal dataset for extremely low altitude drone flights. In *Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2019. 4
- [9] Jakob Geyer, Yohannes Kassahun, Mentar Mahmudi, Xavier Ricou, Rupesh Durgesh, Andrew S. Chung, Lorenz Hauswald, Viet Hoang Pham, Maximilian Mühlegg, Sebastian Dorn, Tiffany Fernandez, Martin Jänicke, Sudesh Mirashi, Chiragkumar Savani, Martin Sturm, Oleksandr Vorobiov, Martin Oelker, Sebastian Garreis, and Peter Schuberth. A2D2: Audi Autonomous Driving Dataset. 2020. 4
- [10] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 4
- [11] Jose L. Gómez, Manuel Silva, Antonio Seoane, Agnès Borrás, Mario Noriega, Germán Ros, Jose A. Iglesias-Guitian, and Antonio M. López. All for one, and one for all: Urbansyn dataset, the third musketeer of synthetic driving scenes, 2023. 4
- [12] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *arXiv preprint arXiv:2404.15506*, 2024. 5
- [13] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 4
- [14] Tobias Koch, Lukas Liebel, Marco Körner, and Friedrich Fraundorfer. Comparison of monocular depth estimation methods using geometrically relevant metrics on the ibims-1 dataset. *Computer Vision and Image Understanding (CVIU)*, 191:102877, 2020. 4
- [15] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Mallocci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020. 3
- [16] Yixuan Li, Lihan Jiang, Linning Xu, Yuanbo Xiangli, Zhenzhi Wang, Dahua Lin, and Bo Dai. Matrixcity: A large-scale city dataset for city-scale neural rendering and beyond. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3205–3215, 2023. 4
- [17] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 4
- [18] Lukas Mehl, Jenny Schmalfuss, Azin Jahedi, Yaroslava Naliwayko, and Andrés Bruhn. Spring: A high-resolution high-detail dataset and benchmark for scene flow, optical flow and stereo. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 4
- [19] Jorge J Moré. The levenberg-marquardt algorithm: implementation and theory. In *Numerical analysis: proceedings of the biennial Conference held at Dundee, June 28–July 1, 1977*, pages 105–116. Springer, 2006. 1
- [20] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgb-d images. In *ECCV*, 2012. 4
- [21] Simon Niklaus, Long Mai, Jimei Yang, and Feng Liu. 3d ken burns effect from a single image. *ACM Transactions on Graphics*, 38(6):184:1–184:15, 2019. 4
- [22] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 3
- [23] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10106–10116, 2024. 5
- [24] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020. 5

- [25] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *International Conference on Computer Vision (ICCV) 2021*, 2021. 4
- [26] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M. Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 4
- [27] Thomas Schöps, Torsten Sattler, and Marc Pollefeys. BAD SLAM: Bundle adjusted direct RGB-D SLAM. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 4
- [28] Pei Sun, Henrik Kretschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 4
- [29] Michal Tölgýessy, Martin Dekan, L’uboš Chovanec, and Peter Hubinský. Evaluation of the azure kinect and its comparison to kinect v1 and kinect v2. *Sensors*, 21(2):413, 2021. 4
- [30] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *International Conference on 3D Vision (3DV)*, 2017. 4
- [31] Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z. Dai, Andrea F. Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R. Walter, and Gregory Shakhnarovich. DIODE: A Dense Indoor and Outdoor DEpth Dataset. *CoRR*, abs/1908.00463, 2019. 4
- [32] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. 1
- [33] Kaixuan Wang and Shaojie Shen. Flow-motion and depth network for monocular stereo and beyond. *CoRR*, abs/1909.05452, 2019. 4
- [34] Qiang Wang, Shizhen Zheng, Qingsong Yan, Fei Deng, Kaiyong Zhao, and Xiaowen Chu. IRS: A large synthetic indoor robotics stereo dataset for disparity and surface normal estimation. *CoRR*, abs/1912.09678, 2019. 4
- [35] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024. 5
- [36] Wenshan Wang, DeLong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. 2020. 4
- [37] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS Datasets and Benchmarks 2021)*, 2021. 4
- [38] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024. 5
- [39] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. *Computer Vision and Pattern Recognition (CVPR)*, 2020. 4
- [40] 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. *CoRR*, abs/2012.09365, 2020. 5
- [41] Amir R Zamir, Alexander Sax, , William B Shen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018. 4
- [42] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *Proceedings of The European Conference on Computer Vision (ECCV)*, 2020. 4