Supplementary Material
Learning the Depths of Moving People by Watching Frozen People

Zhengqi Li[1] Tali Dekel[2] Forrester Cole[2] Richard Tucker[2]
Noah Snavely[1,2] Ce Liu[2] William T. Freeman[2]
[1] Cornell Tech, Cornell University [2] Google Research

This document includes the following:

1. Mathematical derivation of depth from motion parallax (described in Section 4 of the main manuscript).
2. Mathematical derivation of error metrics (described in Section 5 in the main manuscript).
3. Qualitative comparison to parametric human model fitting.

1. Derivations of depth from motion parallax

Here we provide detailed derivations of depth from motion parallax using the Plane+Parallax representation (Section 4.1).

Recall in the main manuscript, we define the relative camera pose as $\mathbf{R} \in SO(3), \mathbf{t} \in \mathbb{R}^3$ from source image I^s to reference image I^r with common intrinsics matrix $\mathbf{K}$. We denote the forward flow from I^r to I^s as $\mathbf{f}_{\text{fwd}}$, and the backward flow from I^s to I^r as $\mathbf{f}_{\text{bwd}}$. Let Π denote a real or virtual planar surface, and let d'_Π denote the distance between the camera center of source image I^s and the plane Π, and h the distance between the 3D scene point corresponding to 2D pixel $\mathbf{p}$ and Π. It can be shown (See Appendix of [2] for full intermediate derivations) that

$$\mathbf{p} = \mathbf{p}_w + \frac{h}{D_{\text{pp}}(\mathbf{p})}\frac{t_z}{d'_\Pi}\mathbf{p}_w - \frac{h}{D_{\text{pp}}(\mathbf{p})d'_\Pi}\mathbf{K}\mathbf{t} \tag{1}$$

$$= \mathbf{p}_w + \frac{h}{D_{\text{pp}}(\mathbf{p})d'_\Pi}(t_z\mathbf{p}_w - \mathbf{K}\mathbf{t}) \tag{2}$$

where $D_{\text{pp}}(\mathbf{p})$ is the estimated depth at $\mathbf{p}$ in the reference image I^r, t_z is the third component of translation vector $\mathbf{t}$, and $\mathbf{p}_w$ is the 2D image point in I^r that results from warping the corresponding 2D pixel (by optical flow $\mathbf{f}_{\text{fwd}}$) in I^s by a homography $\mathbf{A}$:

$$\mathbf{p}_w = \frac{\mathbf{A}\mathbf{p}'}{\mathbf{a}_3^T\mathbf{p}'} \tag{3}$$

$$\text{where } \mathbf{A} = \mathbf{K}\left(\mathbf{R} + \mathbf{t}\frac{\mathbf{n}'^T}{d'_\Pi}\right)\mathbf{K}^{-1}$$

where $\mathbf{p}' = \mathbf{p} + \mathbf{f}_{\text{fwd}}(\mathbf{p})$, $\mathbf{a}_3^T$ is the third row of $\mathbf{A}$, and $\mathbf{n}'$ is normal of plane Π with respect to the camera of source image I^s. Note that the original paper [2] divides the P+P representation into two cases depending on whether $t_z = 0$, but we combine these two cases into one equation shown in Equation 2 by algebraic manipulations.

Now, if we set plane Π at infinity, using L'Hôpital's rule, we can cancel out H and d'_Π and obtain the following equations:

$$\mathbf{p} = \mathbf{p}_w + \frac{t_z\mathbf{p}_w - \mathbf{K}\mathbf{t}}{D_{\text{pp}}(\mathbf{p})} \tag{4}$$

$$D_{\text{pp}}(\mathbf{p}) = \frac{||t_z\mathbf{p}_w - \mathbf{K}\mathbf{t}||_2}{||\mathbf{p} - \mathbf{p}_w||_2},$$

$$\text{where } \mathbf{p}_w = \frac{\mathbf{A}'\mathbf{p}'}{\mathbf{a}'^T_3\mathbf{p}'} \text{ and } \mathbf{A}' = \mathbf{K}\mathbf{R}\mathbf{K}^{-1}.$$

2. Derivation of error metrics

Recall that in Section 5 of our main manuscript we define five different deth error metrics based the on scale-invariant RMSE (si-RMSE). Here we provide definitions of each error metric. Note that we can use similar algebraic manipulations to those proposed in [3] to evaluate all terms in time linear in the number of pixels.

As in the main paper, we denote with $\hat{D}$ the predicted depth, and denote with D_{gt} the ground truth depth. We define $R(\mathbf{p}) = \log \hat{D}(\mathbf{p}) - \log D_{\text{gt}}(\mathbf{p})$, i.e., the difference between computed and ground truth log-depth. We also denote human regions as $\mathcal{H}$ (with N_h valid pixels), non-human (environment) regions as $\mathcal{E}$ (with N_e valid pixels), and the full image region as $I = \mathcal{H} \cup \mathcal{E}$ (with $N = N_e + N_h$ valid pixels).

Our error metrics are defined as follows:

si-full measures the si-RMSE between all pairs of pixels, giving the overall accuracy across the entire image:

$$\textbf{si-full} = \frac{1}{N^2} \sum_{\mathbf{p} \in I} \sum_{\mathbf{q} \in I} \left(R(\mathbf{p}) - R(\mathbf{q})\right)^2 \tag{5}$$

$$= \frac{1}{N^2} \sum_{\mathbf{p} \in I} \sum_{\mathbf{q} \in I} R(\mathbf{p})^2 + R(\mathbf{q})^2 - 2R(\mathbf{p})R(\mathbf{q}) \tag{6}$$

$$= \frac{2}{N^2} \left(N \sum_{\mathbf{p} \in I} R(\mathbf{p})^2 - \sum_{\mathbf{p} \in I} R(\mathbf{p}) \sum_{\mathbf{q} \in I} R(\mathbf{q}) \right) \tag{7}$$

$$= \frac{2}{N} \sum_{\mathbf{p} \in I} R(\mathbf{p})^2 - \frac{2}{N^2} \left(\sum_{\mathbf{p} \in I} R(\mathbf{p}) \right)^2 \tag{8}$$

si-env measures pairs of pixels in non-human regions $\mathcal{E}$ thus computing the accuracy of the depth in the environment:

$$\textbf{si-env} = \frac{1}{N_e^2} \sum_{\mathbf{p} \in \mathcal{E}} \sum_{\mathbf{q} \in \mathcal{E}} \left(R(\mathbf{p}) - R(\mathbf{q})\right)^2 \tag{9}$$

$$= \frac{2}{N_e^2} \left(N_e \sum_{\mathbf{p} \in \mathcal{E}} R(\mathbf{p})^2 - \sum_{\mathbf{p} \in \mathcal{E}} R(\mathbf{p}) \sum_{\mathbf{q} \in \mathcal{E}} R(\mathbf{q}) \right) \tag{10}$$

si-hum measures pairs where one pixel lies in the human region $\mathcal{H}$ and one lies anywhere in the image, thus computing overall depth accuracy for the people in the scene:

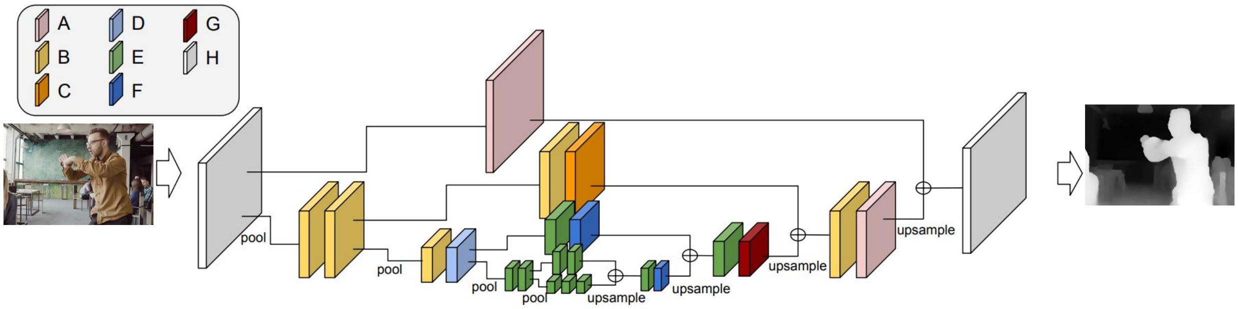
$$\textbf{si-hum} = \frac{1}{N N_h} \sum_{\mathbf{p} \in \mathcal{H}} \sum_{\mathbf{q} \in I} R(\mathbf{p})^2 + R(\mathbf{q})^2 - 2R(\mathbf{p})R(\mathbf{q}) \tag{11}$$

$$= \frac{1}{N N_h} \left(N \sum_{\mathbf{p} \in \mathcal{H}} R(\mathbf{p})^2 + N_h \sum_{\mathbf{q} \in I} R(\mathbf{q})^2 - 2 \sum_{\mathbf{p} \in \mathcal{H}} R(\mathbf{p}) \sum_{\mathbf{q} \in I} R(\mathbf{q}) \right) \tag{12}$$

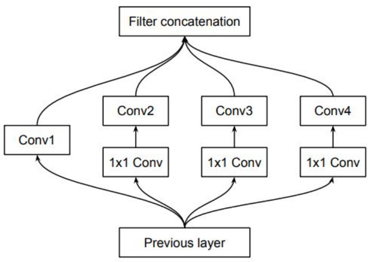
si-hum can further be divided into the sum of two error measures: **si-intra** measures si-RMSE within $\mathcal{H}$, or human accuracy independent of the environment:

$$\textbf{si-intra} = \frac{1}{N_h^2} \sum_{\mathbf{p} \in \mathcal{H}} \sum_{\mathbf{q} \in \mathcal{H}} \left(R(\mathbf{p}) - R(\mathbf{q})\right)^2 \tag{13}$$

$$= \frac{2}{N_h^2} \left(N_h \sum_{\mathbf{p} \in \mathcal{H}} R(\mathbf{p})^2 - \sum_{\mathbf{p} \in \mathcal{H}} R(\mathbf{p}) \sum_{\mathbf{q} \in \mathcal{H}} R(\mathbf{q}) \right) \tag{14}$$



(a)



(b)

Block Id	A	B	C	D	E	F	G
#In/#Out	128/64	128/128	128/128	128/256	256/256	256/256	256/128
Inter Dim	64	32	64	32	32	64	32
Conv1	1x1	1x1	1x1	1x1	1x1	1x1	1x1
Conv2	3x3	3x3	3x3	3x3	3x3	3x3	3x3
Conv3	7x7	5x5	7x7	5x5	5x5	7x7	5x5
Conv4	11x11	7x7	11x11	7x7	7x7	11x11	7x7

(c)

Figure 1: **Network Architecture**. Each block with a different color (id) in (a) indicates a convolutional layer. The block labeled H indicates a 3×3 convolutional layer and all other blocks are implemented as a variant of an Inception module [4], as shown in (b). Parameters for each type of layer are shown in (c). We use bilinear interpolation to upsample features in the network. Figures modified from Chen *et al*. [1].

si-inter measures si-RMSE between pixels in $\mathcal{H}$ and in $\mathcal{E}$, or human accuracy w.r.t. the environment:

$$\textbf{si-inter} = \frac{1}{N_e N_h} \sum_{\mathbf{p}\in\mathcal{H}} \sum_{\mathbf{q}\in\mathcal{E}} R(\mathbf{p})^2 + R(\mathbf{q})^2 - 2R(\mathbf{p})R(\mathbf{q}) \tag{15}$$

$$= \frac{1}{N_e N_h} \left(N_e \sum_{\mathbf{p}\in\mathcal{H}} R(\mathbf{p})^2 + N_h \sum_{\mathbf{q}\in\mathcal{E}} R(\mathbf{q})^2 - 2 \sum_{\mathbf{p}\in\mathcal{H}} R(\mathbf{p}) \sum_{\mathbf{q}\in\mathcal{E}} R(\mathbf{q}) \right). \tag{16}$$

3. Network Architecture

Our network architecture is a variant of the hourglass network proposed by Chen*et al*. [1], and is shown in Figure 1. Specifically, our network has a standard encoder and decoder U-Net structure, with matching input and output resolution, consisting of approximately 5M parameters. In addition, an Inception module [4] is used in each convolutional layer of the network. We replaced the nearest-neighbor upsampling layers by bilinear upsampling layers, which we found produced sharper depth maps while slightly improving overall accuracy.

4. Instructions on running SfM/MVS on MC dataset.

To aid in reproducing our results, we refer readers to the following URL for detailed instructions for running SfM and MVS on our MannequinChallenge dataset:
`https://docs.google.com/document/d/1lWOcbLIeGGVVpjkGiMaq0zRVZvaBJnewRUPN2mdAD_A/edit?usp=sharing`

References

[1] W. Chen, Z. Fu, D. Yang, and J. Deng. Single-image depth perception in the wild. In *Neural Information Processing Systems*, pages 730–738, 2016.

[2] M. Irani and P. Anandan. Parallax geometry of pairs of points for 3D scene analysis. In *Proc. European Conf. on Computer Vision (ECCV)*, pages 17–30. Springer, 1996.

[3] Z. Li and N. Snavely. Learning Intrinsic Image Decomposition from Watching the World. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, 2018.

[4] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, 2015.