A. Importance of Accurate Shift and Scale for 3D Geometry Recovering

In this section, we show that accurate shift is crucial for recovering 3D geometry while knowing scale value is not necessary.

For data obtained from stereo videos [8], or stereo images [12], ground-truth disparity can be extracted only up to unknown scale and shift coefficients. Without knowing the correct disparity shift value, 3D geometry can not be recovered.

Hereinafter, we assume that an RGB camera can be described in terms of a pinhole model.

Let us consider the 3D line l parametrized with coefficients a, b, c. For projection point x, y on the camera plane, its depth can be calculated as:

$$d = ax + by + c, \tag{1}$$

So, we can assign a depth value for each point along this 3D line. Suppose that inverse depth (disparity) values for the points along the line are known up to shift and scale:

$$\frac{1}{\tilde{d}} = \frac{C_1}{ax + by + c} + C_2, \tag{2}$$

or, equivalently,

$$\tilde{d} = \frac{ax + by + c}{C_1 + C_2(ax + by + c)}. \tag{3}$$

The expression above defines a 3D line if and only if $C_2 = 0$. Therefore, to obtain predictions from which 3D geometry can be recovered, a neural network should explicitly estimate the C_2 coefficient.

As shown, C_2 coefficient has a large impact on the 3D geometry. At the same time, C_1 affects only the scale of the scene. To illustrate that, we can consider mapping from a camera plane point (x, y) having a depth d to a 3D point:

$$\begin{pmatrix} x \\ y \\ d \end{pmatrix} \mapsto \begin{pmatrix} \frac{(x-c_x)d}{f_x} \\ \frac{(y-c_y)d}{f_y} \\ d \end{pmatrix}. \tag{4}$$

Suppose that the original depth map gets scaled by a factor C_1. According to 4, the coordinates of 3D points get then multiplied by C_1 as well. We can interpret this as the entire scene getting scaled by C_1 without affecting the geometry correctness (e.g., all angles and curvatures remain unchanged).

B. LRN-based Neural Network

Fig. 1 depicts the architecture of our LRN-based SVDE models. Following [8], we use a RefineNet architecture to address the depth estimation problem. For the sake of efficiency, we use Light-Weight RefineNet (LRN) [6].

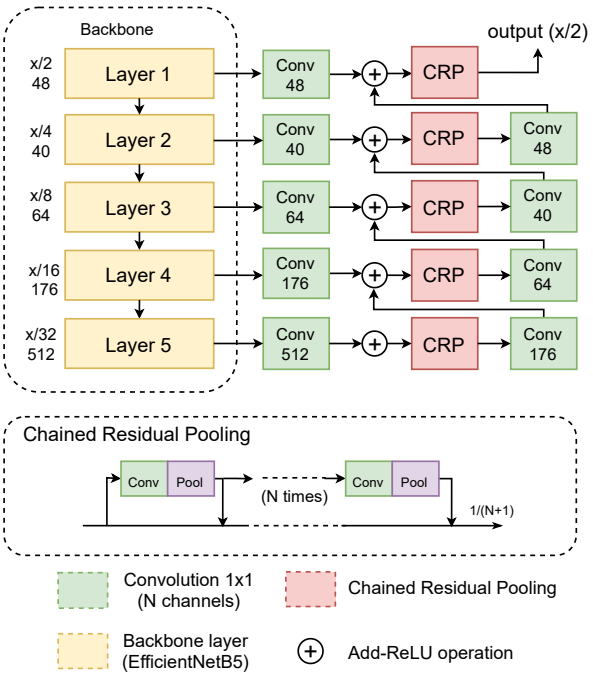

Figure 1. Architecture of our LRN-based SVDE models.

The encoders are based on architectures from the EfficientNet family [10]: EfficientNet-Lite0 and EfficientNet-B5, both pre-trained on *ImageNet*.

In the original LRN model, each encoder output is processed with a 1×1 convolution with 256 output channels. This parameter is hard-coded for both light-weight and powerful models, so we can neither choose smaller values if building a light-weight decoder or use more channels for more powerful decoder architectures. We claim this non-adaptive approach to be suboptimal and propose a more flexible alternative. Unlike the original LRN, we set the number of output channels in the fusion convolutional layer equal to the number of channels in the corresponding backbone level. Then, each encoder output gets fused with the features from a deeper layer (see Fig. 1), and the fused output has the same number of channels as the encoder output.

C. Data

C.1. MiDaS Dataset Mixture

Stereo Movies The original StereoMovies dataset collected in MiDaS [8] consists of 23 stereo videos and features video frames from various non-static environments. We follow the similar data acquisition and processing protocol, but, we use 26 additional stereo movies, totalling 49 movies (listed in Tab.1 and Tab. 2). The obtained video

Name	Frames
3-D Sex and Zen: Extreme Ecstasy (2011)	3080
Battle of the Year (2013)	4074
Cirque du Soleil: Journey of Man (2000)	897
Creature from the Black Lagoon (1954)	680
Dark Country (2009)	324
Drive Angry (2011)	2437
Exodus: Gods and Kings (2014)	5650
Final Destination 5 (2011)	2212
Flying Swords of Dragon Gate (2011)	2618
Galapagos: The Enchanted Voyage (1999)	230
Ghosts of the Abyss (2003)	946
Hugo (2011)	3338
Into the Deep (1994)	22
Jack the Giant Slayer (2013)	5174
Journey 2: The Mysterious Island (2012)	3184
Journey to the Center of the Earth (2008)	1416
Life of Pi (2012)	5160
My Bloody Valentine (2009)	1627
Oz the Great and Powerful (2013)	4559
Pina (2011)	1827
Piranha 3DD (2012)	1766
Pirates of the Caribbean: On Stranger Tides (2011)	5015
Pompeii (2014)	3644
Prometheus (2012)	4188
Sanctum (2011)	1976
Saw 3D: The Final Chapter (2010)	2757

Table 1. Stereo movies used in our experiments, part 1

Name	Frames
Sea Rex 3D (2010)	1015
Silent Hill: Revelation 3D (2012)	1747
Sin City: A Dame to Kill For (2014)	3585
Space Station 3D (2002)	362
Stalingrad (2013)	6453
Step Up 3D (2010)	3209
Step Up Revolution (2012)	3542
Texas Chainsaw 3D (2013)	3089
The Amazing Spider-Man (2012)	5378
The Child's Eye (2010)	1232
The Darkest Hour (2011)	3640
The Final Destination (2009)	1998
The Great Gatsby (2013)	4788
The Hobbit: An Unexpected Journey (2012)	4128
The Hobbit: The Desolation of Smaug (2013)	7266
The Hobbit: The Battle of the Five Armies (2014)	6568
The Hole (2010)	1685
The Martian (2015)	4893
The Three Musketeers (2011)	5284
The Ultimate Wave Tahiti (2010)	638
Ultimate G's (2000)	366
Underworld: Awakening (2012)	3093
X-Men: Days of Future Past (2014)	3482
Overall	**146242**

Table 2. Stereo movies used in our experiments, part 2

frames are highly diverse, containing landscapes, architecture, humans in action, and other various scenes.

We sample one frame per second from the collected videos. We omit the first and the last 10% of frames as they usually contain opening and closing credits. We consider valid only the pixels where the discrepancy between left to right and right to left disparities is less than 8 pixels. Accordingly, we use only images where the disparity is valid for more than 80% of pixels and the difference between maximal and minimal disparities exceeds 8 pixels.

WSVD. We do not use the WSVD dataset [11] used in the original MiDaS mixture since it contains data in the form of web links referring to the sources that have already been partially deleted.

C.2. LeReS Dataset Mixture

DIML Outdoor. DIML Outdoor contains calibrated and rectified stereo images so that disparities can be extracted via stereo matching. LeReS [14] uses GANet [15], while we perform stereo matching with AANet [13].

Holopix. To obtain disparities from stereo data in Holopix [1], we opt for a more accurate PWCNet [9] instead of FlowNet2 [2] used in LeReS.

The other datasets in MiDaS and LeReS dataset mixtures are publicly available for downloading and contain ground truth data, so we use these datasets 'as is'.

C.3. Data Used in Ablation Studies.

In ablation studies, we use the NYUv2 raw [5] dataset. We select approximately 150k images from the training subset, evaluating on the original test subset of 654 images.

D. Visualizations

In this section, we provide additional visualizations of point clouds reconstructed from depth maps estimated with different SVDE methods. To create a point cloud from predictions of our SVDE models, we apply the pre-trained focal recovery module from LeReS [14].

Qualitative comparison of our best GP^2-trained B5-LRN model with other existing SVDE methods is presented in Fig. 2. Mannequin [3] is trained on UTS data obtained from

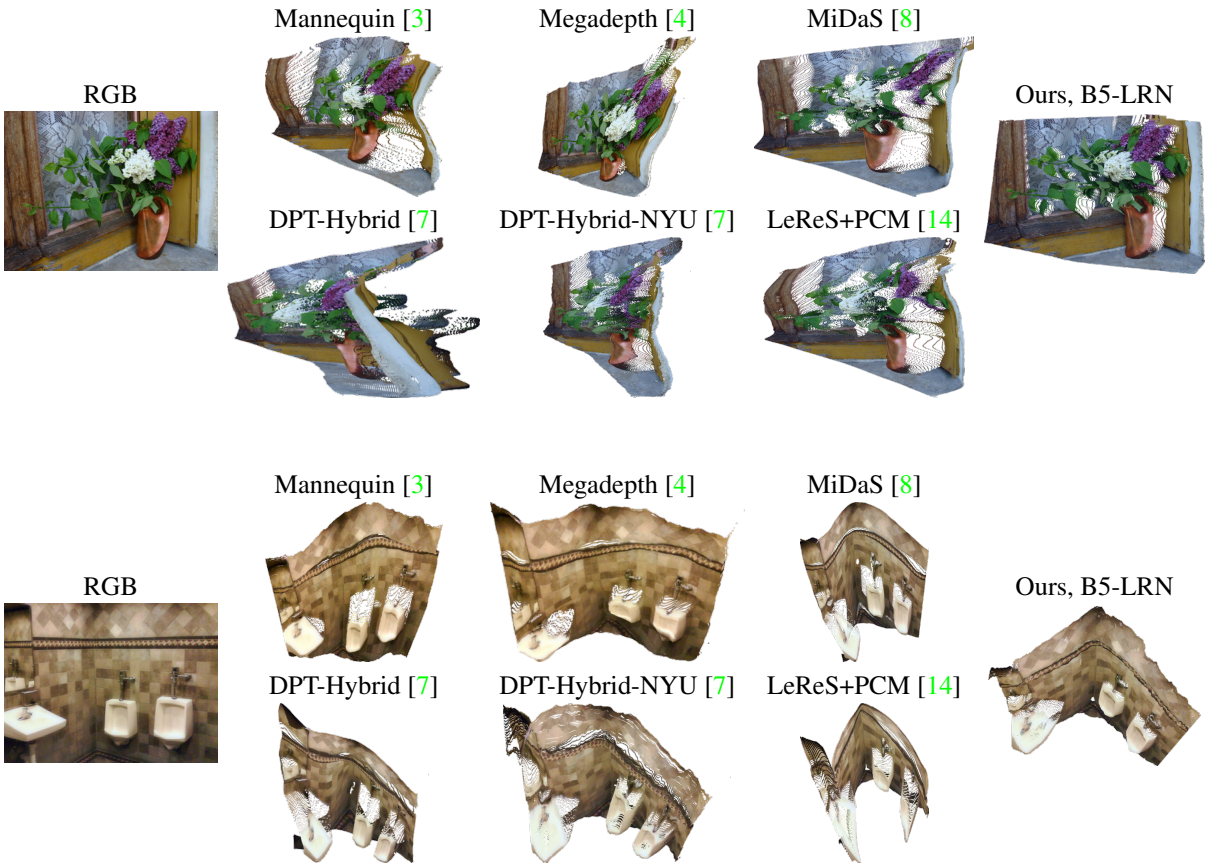

Figure 2. Point clouds reconstructed from depth estimates obtained with existing SVDE methods, including our GP^2-trained B5-LRN SVDE model.

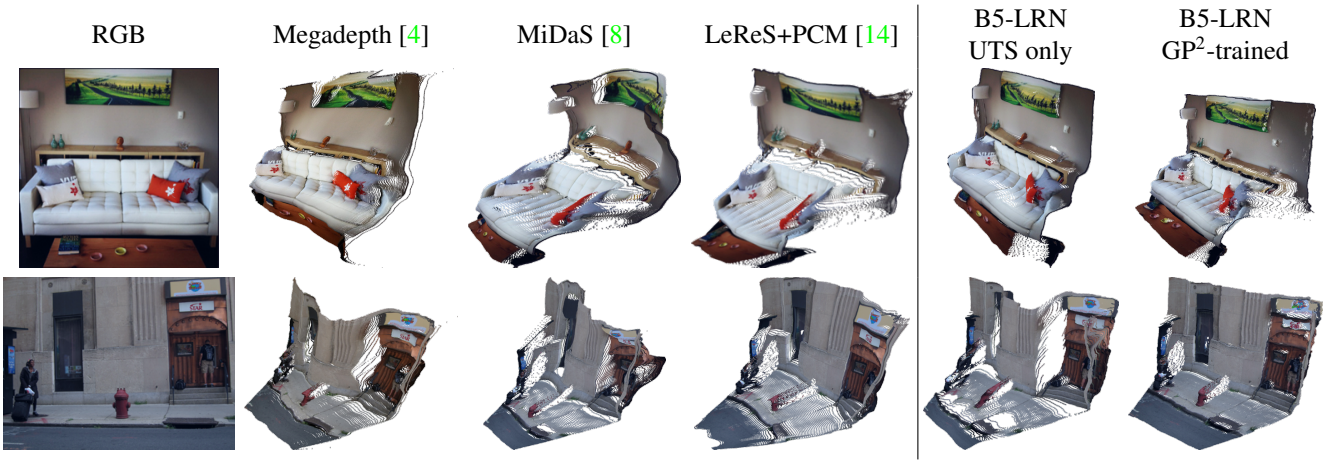

Figure 3. Point clouds reconstructed from depth estimates obtained with existing SVDE methods, including our UTS-trained and GP^2-trained B5-LRN SVDE model.

3D reconstruction, so it predicts UTS depth. MiDaS [8] predicts depth with incorrect shifts, resulting in severely distorted point clouds. DPT fine-tuned on NYUv2 estimates absolute depth but also fails to restore the actual scene geometry.

In Fig. 3, we show the benefits of using UTSS data for training a geometry-preserving SVDE model. For this purpose, we train our B5-LRN model either on UTS data only

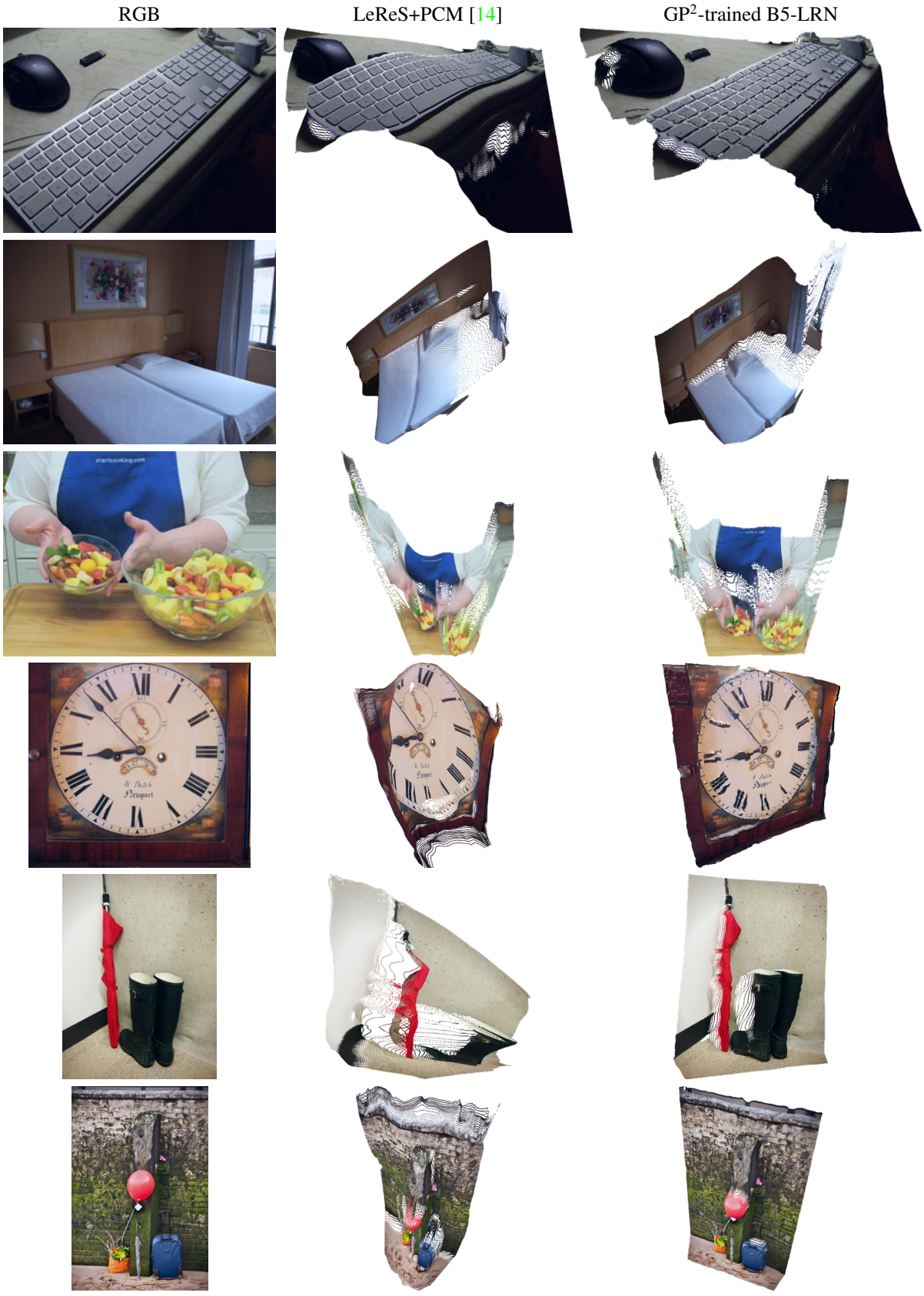

Figure 4. Point clouds reconstructed from depth estimates obtained with LeReS+PCM and GP^2-trained B5-LRN.

Figure 5. Point clouds obtained from depth estimates of our B5-LRN model. Paintings are a new data domain unseen during training, however, our method successfully handles these images, estimating depth adequately.

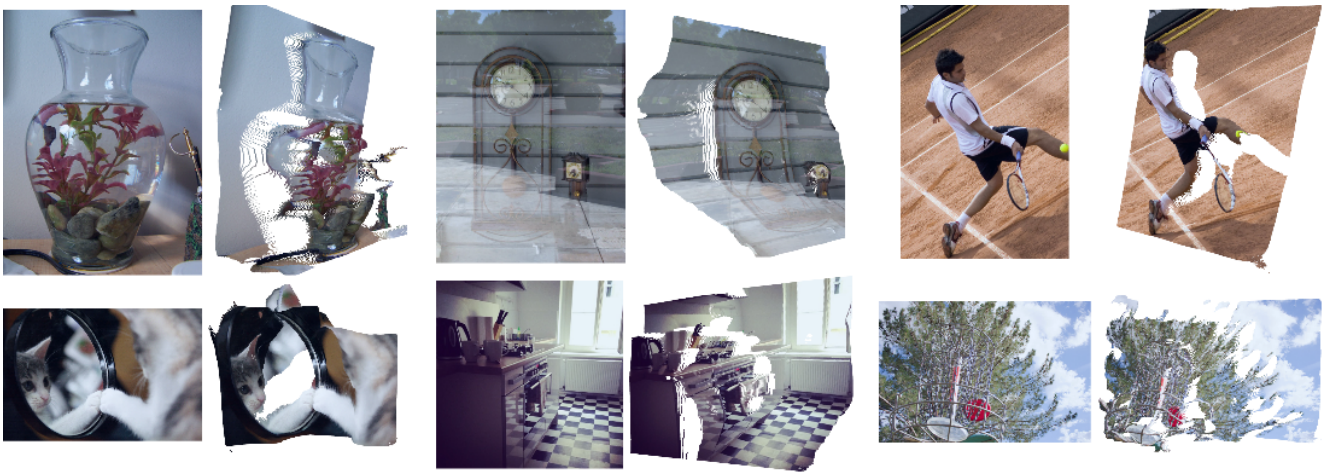

Figure 6. Failure cases of our B5-LRN model: as one might see, reflective and glassy surfaces, mirrors, thin objects are difficult for our model.

or on a mixture of UTS and UTSS data using GP^2. According to the visualized point clouds, adding UTSS data to the training mixture improves the quality of reconstructions. We also compare our GP^2-trained B5-LRN model with geometry-preserving MegaDepth trained on only UTS data, non-geometry preserving MiDaS trained on a mixture of UTS and UTSS data, and geometry-preserving LeReS trained on a mixture of UTS and UTSS data. As one might see, our GP^2-trained B5-LRN allows recovering more accurate point clouds, which testifies in favor of the proposed training scheme.

To demonstrate the effectiveness of our training scheme for general-purpose geometry-preserving SVDE, we compare directly to LeReS+PCM being the only existing general-purpose and geometry-preserving SVDE method trained on a mixture of UTS and UTSS data. In Fig. 4, we visualize point clouds reconstructed using depth maps predicted by our GP^2-trained B5-LRN SVDE model and LeReS+PCM. Since both our model and LeReS use the same module for focal length estimation, the observed quality gap should be attributed to the better depth estimates obtained with our model.

Fig. 5 depicts the point clouds reconstructed from paintings by the GP^2-trained B5-LRN model. Paintings are a new data domain unseen during training; however, our model generalizes well even on non-photorealistic images.

Finally, we visualize some failures of GP^2-trained B5-LRN in Fig. 6 to give a complete picture. Expectedly, our model fails on highly reflective and glassy surfaces, mirrors, and thin objects.

References

[1] Yiwen Hua, Puneet Kohli, Pritish Uplavikar, Anand Ravi, Saravana Gunaseelan, Jason Orozco, and Edward Li. Holopix50k: A large-scale in-the-wild stereo image dataset. In *CVPR Workshop on Computer Vision for Augmented and Virtual Reality, Seattle, WA, 2020.*, June 2020. 2

[2] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017. 2

[3] Zhengqi Li, Tali Dekel, Forrester Cole, Richard Tucker, Noah Snavely, Ce Liu, and William T. Freeman. Learning the depths of moving people by watching frozen people. *CoRR*, abs/1904.11111, 2019. 2, 3

[4] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[5] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 2

[6] Vladimir Nekrasov, Chunhua Shen, and Ian Reid. Light-weight refinenet for real-time semantic segmentation, 2018. 1

[7] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12179–12188, October 2021. 3

[8] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer, 2019. 1, 3

[9] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943, 2018. 2

[10] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 2019. 1

[11] Chaoyang Wang, Simon Lucey, Federico Perazzi, and Oliver Wang. Web stereo video supervision for depth prediction from dynamic scenes, 2019. 2

[12] Ke Xian, Chunhua Shen, Zhiguo Cao, Hao Lu, Yang Xiao, Ruibo Li, and Zhenbo Luo. Monocular relative depth perception with web stereo data supervision. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1

[13] Haofei Xu and Juyong Zhang. Aanet: Adaptive aggregation network for efficient stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1959–1968, 2020. 2

[14] 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 *Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (CVPR)*, 2021. 2, 3, 4

[15] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip HS Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 185–194, 2019. 2