Stereo Conversion with Disparity-Aware Warping, Compositing and Inpainting – Supplementary Material –

1. Technical Details

In the following, we describe additional technical details for the dataset creation, train augmentation, inference, interactive disparity mapping and mitigation strategies.

Dataset When creating our dataset, we crop floating windows [2], black boundaries in the image data. To this end, we computing the mean over the channel and height dimension and removing the left-most and right-most contiguous number of pixels below a threshold of 5e-3.

Augmentation During training, we perform data augmentation as follows: We randomly flip the training image pair horizontally or vertically (0.5 / 0.1 probability) and add an offset randomly sampled from $[-10, 10]$ to the second image. We also randomly crop the images to a size of 512x256, perform photometric augmentation [5], and randomly blur the reference disparity with 0.2 probability.

Since we omitted the optical flow warping, there is also no flow disocclusion mask $\mathcal{D}_{\text{OF}}$. We thus make use of the disparity warping disocclusion mask $\mathcal{D}$, but augment it by applying a random horizontal displacement sampled from a normal distribution ($\mu = 0$, $\sigma = 10$).

Inference At inference time, we make use of the recent MiDaS [3, 4] single image depth prediction, version 3.1, `dpt_beit_large_512`. For our automatic disparity mapping, we select a as 30,20,15 for close-up, medium shot and long shot, respectively. We choose b such that a only a smaller percentage of the scene is positioned in front of the screen [1], with values of 25, 15 and 10. For the hole filling step, we slightly erode the disocclusion mask with a size of 5 prior to performing the single-sided dilation in order to prevent edge artifacts.

Interactive Disparity Mapping In order to perform disparity mapping with user-provided sparse scribbles, we proceed as follows: In a first step, we compute the optical flow between the reference frame (with provided scribbles) and each other frame. Then we forward-warp the reference frame scribbles into all other frames using the optical flow. Finally, for each frame, we estimate the parameters a and b using a least-squares fit between the inverse depth and the sparse scribbles, at non-empty locations. With a and b given, the disparity can be obtained, which is then used for the stereo conversion.

Mitigation Strategies As introduced in the main paper, we implement two optional mitigation strategies: background stretch and foreground enlargement. For these strategies, we restrict ourselves to scenes with a single foreground object with simple foreground-background separation and only small motion, which in practice still covers significant portions of a movie. Then, we can perform a simple foreground-background segmentation by thresholding the disparity with its mean. Then, for the *background stretch*, users choose a horizontal center and the maximum displacement. Afterwards the left and right side are considered separately. For each side, a quadratic function for horizontal displacement dx is determined that is zero and the image boundary and the horizontal center point and has the maximum displacement as its maximum value. These parabolas are depicted in the main paper and yield positive values for the left side and negative values for the right side. The *foreground enlargement* is realized by a zoom relative to the center point of the foreground mask. Both strategies are integrated into the main model through an additional optical flow that is added to the disparity before warping. For our quantitative evaluation, we select 10 sequences from the test split of our dataset and compare the size of the disocclusion masks before and after applying the strategies.

References

[1] Takashi Kawai, Masahiro Hirahara, Yuya Tomiyama, Daiki Atsuta, and Jukka Häkkinen. Disparity analysis of 3d movies and emotional representations. In *Stereoscopic Displays and Applications XXIV*, volume 8648, pages 293–301. SPIE, 2013. 1

[2] Bernard Mendiburu. *3D Movie Making: Stereoscopic Digital Cinema from Script to Screen*. Focal Press, 2009. 1

[3] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12179–12188, 2021. 1

[4] 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 (TPAMI)*, 44(3):1623–1637, 2022. 1

[5] Zachary Teed and Jia Deng. RAFT: recurrent all-pairs field transforms for optical flow. In *Proc. European Conference on Computer Vision (ECCV)*, pages 402–419. Springer LNCS 12347, 2020. 1