Supplementary Material - Semi-synthesis: A fast way to produce effective datasets for stereo matching

A. Results on Middlebury test sets

We submit CasStereo [1] only trained on Semi-Synthetic-M datasets (namely **SSCasStereo**) to Middlebury website to evaluate the performance of it on official test sets. All pixels are evaluated.

Table 1 summarizes the results of CasStereo and SSCasStereo on Middlebury test sets. Our SSCasStereo improves by a large margin on all metrics.

Table 2 shows the average error on each image of Middlebury test sets. The suffix P stands for perfect rectification (imperfect by default), E stands for exposure changes between views, and L stands for lighting changes between views. Without training on any Middlebury images, our SSCasStereo gets top performance on several images and even ranked **NO.1** on Hoops and Stairs.

For more detailed results of our method, please go check the official evaluation website Middlebury. Our submission name on it is **SSCasStereo**.

B. Fine-tuning on Middlebury

From Table 2 we can see that the performance of our SSCasStereo drops greatly when meeting changes of several factors such as exposures and lighting. We argue that this is because without fine-tuning our model hasn't met such special situations and we do not generate such scenes in our semi-synthetic datasets. We can solve this by either adding data augmentation strategies introduced in HSM [2] or fine-tuning on Middlebury train sets. Here we show the visualization results of our fine-tuned SSCasStereo (namely **SSCasFine**) on Middlebury test sets in the following pages. We can not show the quantitative performance of it because Middlebury only allows one submission per method.

C. Visualization results on Middlebury test sets

Figure 1 gives the visualization results of CasStereo, SSCasStereo, SSCasFine on Middlebury test sets. We download the visualization results of CasStereo and SSCasStereo from the evaluation website and run the visualization for SSCasFine by ourselves. Images from top to bottom are: Australia, AustraliaP, Bicycle2, Classroom2, Classroom2E, Computer, Crusade, CrusadeP, Djembe, DjembeL, Hoops, Livingroom, Newkuba, Plants, Staircase.

We observe that depth maps produced by SSCasStereo outperform CasStereo a lot while SSCasFine further yields more accurate results and becomes more robust to the changes of factors such as exposure, lighting.

D. Visualization results on ETH3D validation sets

Figure 2 shows the visualization results of CasStereo models trained on different datasets on ETH3D validation sets.

Note that models trained on Semi-Synthetic-E give sharper results and at the same time avoid errors in continuous regions which demonstrates the effectiveness of our semi-synthetic datasets once again.

E. Detailed training procedure of CasStereo

Follow the design by the authors, our CasStereo contains three-stage cascade cost volumes. The spatial resolution of feature maps gradually increases and is set to $1/4$, $1/2$ and 1 of the original input image size. Due to the variety of maximum disparity among different scenes, we set the number of depth hypothesis and depth interval to be different for each dataset. From the first to the third stage, the number of depth hypothesis is set to 48, 24, 12 for KITTI and Middlebury, 16, 8, 4 for ETH3D. The depth interval is set to 4, 2, 1 for KITTI and ETH3D, 8, 4, 1 for Middlebury.

We adopt Adam as the optimizer with a base learning rate set to be 0.001. At fine-tuning stage, we decay the learning rate with a ratio of $\frac{1}{3}$ every 10 epochs.

[1] Cascade cost volume for high-resolution multi-view stereo and stereo matching

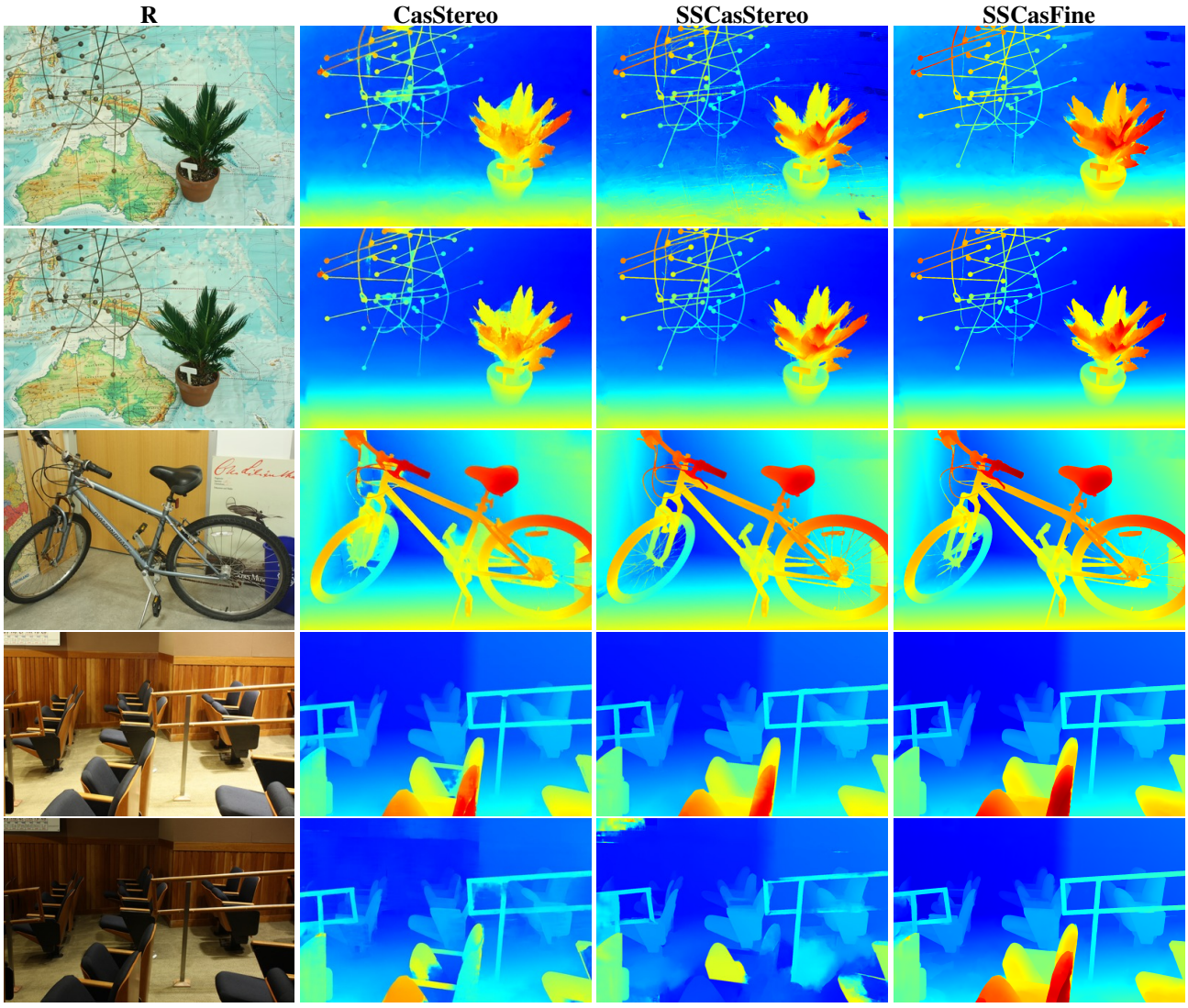
[2] Hierarchical deep stereo matching on high-resolution images

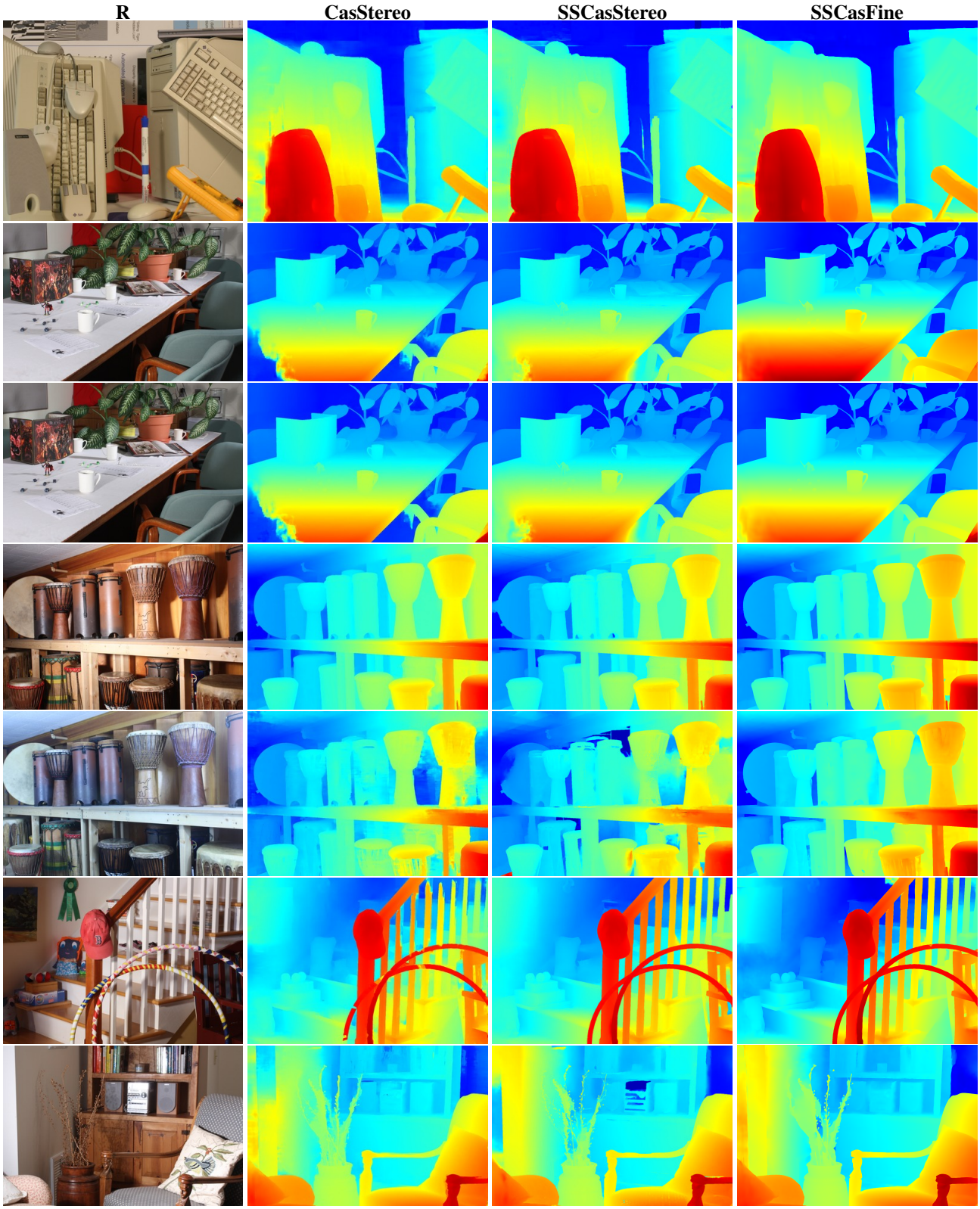
Table 1: Results on Middlebury-v3 official test images where all pixels are evaluated. CasStereo stands for the official implementation of the paper. SSCasStereo refers to CasStereo only trained on Semi-Synthetic-M datasets.

Method	avgerr	rms	bad-1.0	bad-2.0	bad-4.0	A90	A95	A99
CasStereo	8.98	30.4	38.7	26.0	18.5	16.8	45.6	162
SSCasStereo	**6.38**	**21.3**	**34.7**	**21.7**	**14.0**	**11.8**	**36.6**	**107**

Table 2: Average error on each image of Middlebury test sets where all pixels are evaluated. The subscript number shows the absolute rank among the benchmark. Results ranked Top15 are underlined, and results ranked **Top5** are bolded.

Method	avgerr	Austr	AustrP	Bicyc2	Class	ClassE	Compu	Crusa
SSCasStereo	6.38_{31}	5.28_{46}	3.12_{9}	$\mathbf{1.57}_{2}$	$\mathbf{3.12}_{5}$	41.7_{110}	$\mathbf{3.10}_{7}$	8.64_{51}
Method	CrusaP	Djemb	DjembL	Hoops	Livgrm	Nkuba	Plants	Stairs
SSCasStereo	9.24_{61}	1.12_{11}	7.52_{55}	$\mathbf{4.38}_{1}$	4.35_{33}	$\mathbf{4.74}_{13}$	10.3_{26}	$\mathbf{1.88}_{1}$




R
CasStereo
SSCasStereo
SSCasFine

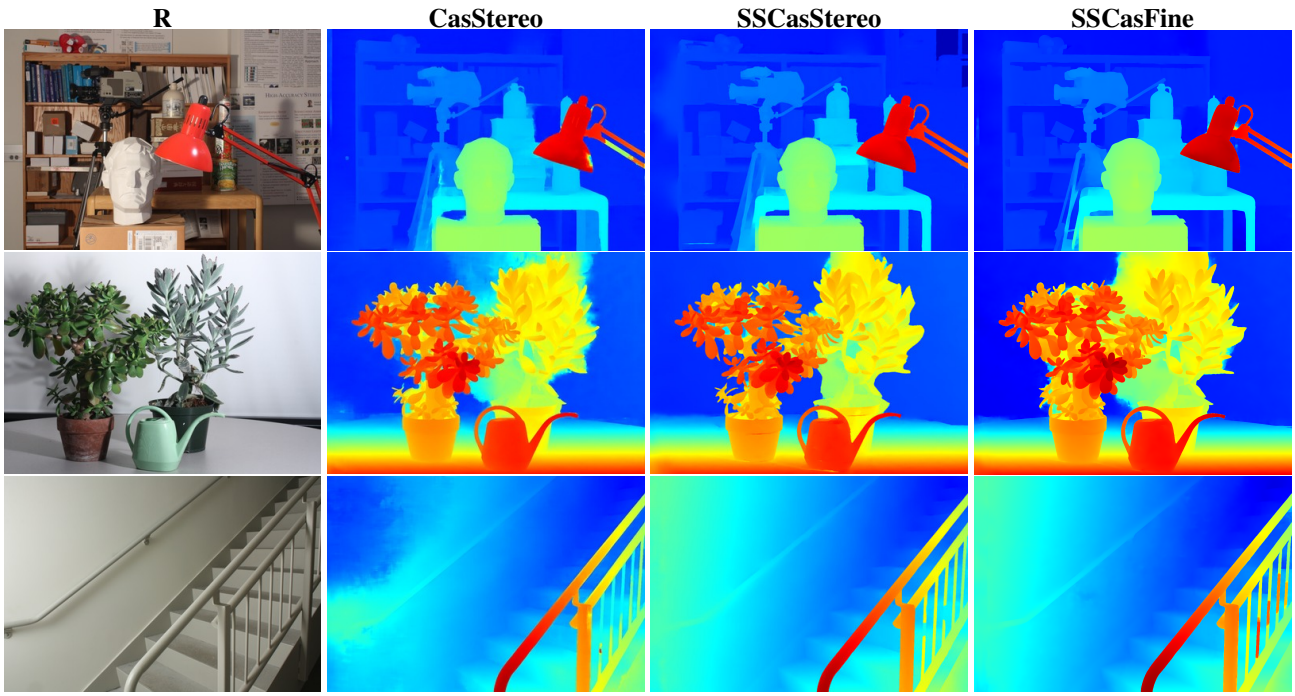


Figure 1: Visualization results on a subset of Middlebury test sets. Each row stands for one image. First column: right view of the image pairs. Second column: CasStereo results. Third column: SSCasStereo results. Fourth column: SSCasFine results.

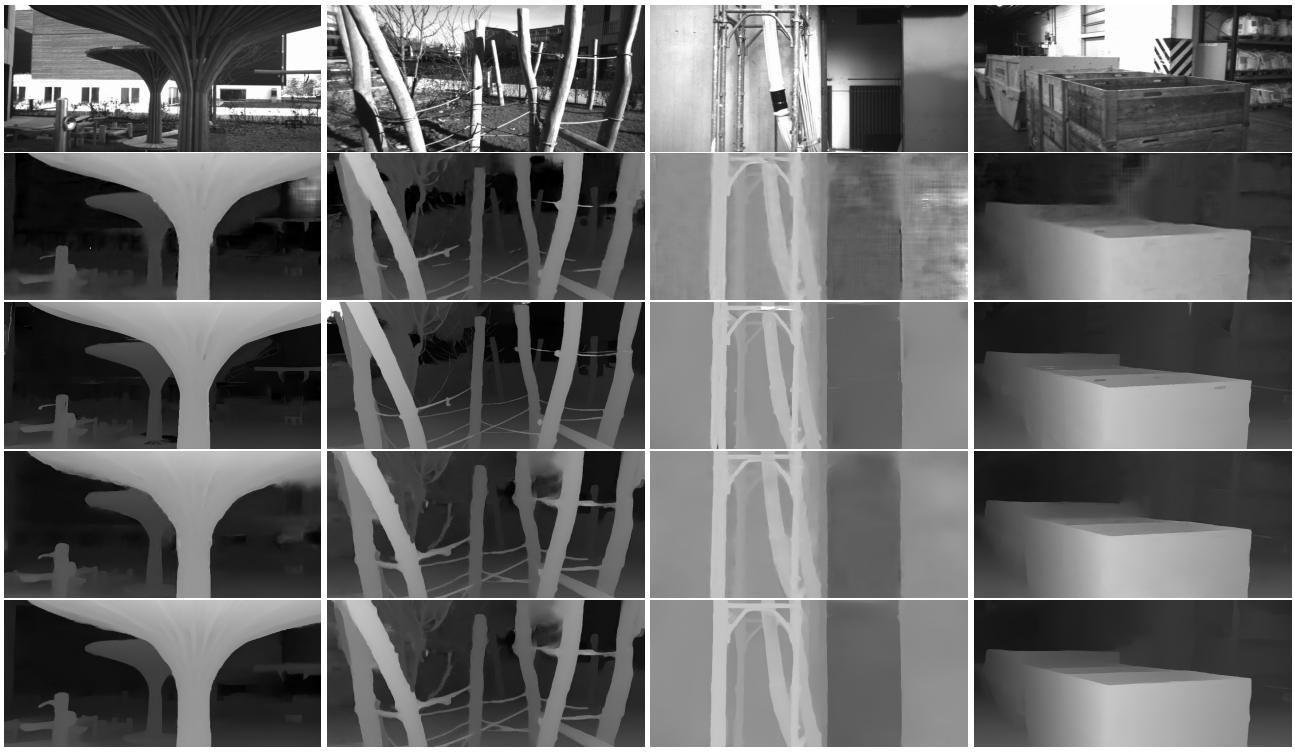


Figure 2: Visualization results on a subset of ETH3D validation sets. Each column stands for one image. First row: left view of the image pairs. Second row: CasStereo trained on SceneFlow. Third row: CasStereo trained on Semi-Synthetic-E. Fourth column: CasStereo trained on SceneFlow and ETH3D. Fifth column: CasStereo trained on Semi-Synthetic-E and ETH3D.