ActiveStereoNet: End-to-End Self-Supervised Learning for Active Stereo Systems (Supplementary Materials)

Yinda Zhang[1,2], Sameh Khamis[1], Christoph Rhemann[1], Julien Valentin[1], Adarsh Kowdle[1], Vladimir Tankovich[1], Michael Schoenberg[1], Shahram Izadi[1], Thomas Funkhouser[1,2], Sean Fanello[1]

[1]Google Inc., [2]Princeton University

In this supplementary material, we provide additional information regarding our implementation details and more ablation studies on important components of the proposed framework. We conclude the supplementary material evaluating our method on passive stereo matching, showing that the proposed approach can be applied also on RGB images.

1 Additional Implementation Details

To ensure reproducibility, in the following, we give all the details needed to re-implement the proposed architecture.

1.1 Architecture

Our model extends the architecture of [3]. Although the specific architecture is out of the scope of this paper, we introduced three different modifications to address specific problems of active stereo matching.

First, we adjust Siamese tower to maintain more high frequency signals from the input image. Unlike [3] that aggressively reduce resolution at the beginning of the tower, we run several residual blocks on full resolution and reduce resolution later using convolution with stride (Fig. 1 (a)).

Second, we use a two-stream refinement network (Fig. 1 (c)). Instead of stacking the image and the upsampled low resolution disparity, we feed these two into different pathways, where each pathway consists of 1 layer of convolution with 16 channels and 3 resnet blocks that maintain the number channels. The output of two pathways are concatenated together into a 32-dim feature map, which is further fed into 3 resnet blocks. We found this architecture produces results with way less dot artifacts.

Lastly, we have an additional invalidation network which predicts the confidence of the estimated disparity (Fig. 1 (b)). One pursuing only very accurate depth could use this output to remove unwanted depth. The invalidation network takes the concatenation of the left and right tower features, and feed them into 5 resnet and 1 convolution in the end. The output is an invalidation map with the same resolution of the tower feature, i.e. 1/8 of the input resolution. This low resolution invalidation map is upsampled to full resolution by bilinear

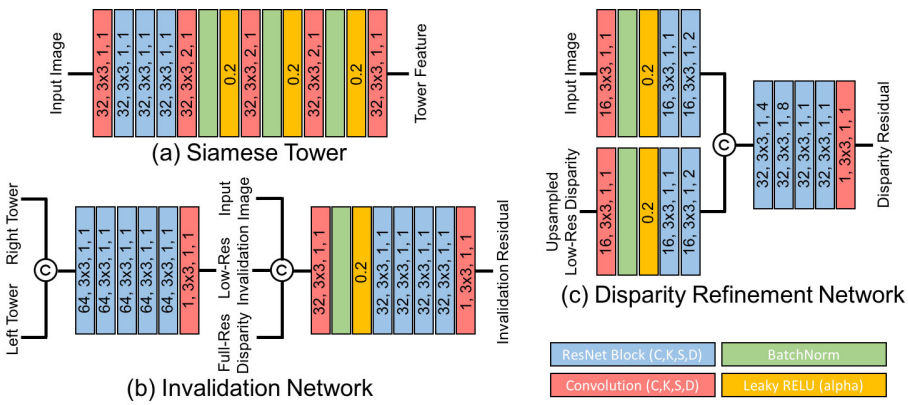

Fig. 1. Detailed Network Architecture. (a) Siamese Tower, (b) Invalidation Network, (c) Disparity Refinement Network. In resnet block and convolution, the numbers are number of channels, kernel size, stride, and dilated rate. In leaky RELU, the number is the slope for $x < 0$. © means feature map concatenation.

interpolation. In order to refine it, we concatenate this upsampled invalidation map with the full resolution refined disparity and the input image, and feed them into 1 conv + 4 resnet block + 1 conv. The output is a one dimensional score map with higher value representing invalidate area.

During the training, we formulate the invalidation mask learning as a classification task. Note that the problem is completely unbalanced, i.e. most of the pixels are valid, therefore a trivial solution for the network consists of assigning all the pixels a positive label. To avoid this, we reweigh the output space by a factor of 10 (we notice that invalid pixels are roughly 10% of the valid ones). For valid pixels we assign the value $y^+ = 1$, whereas for invalid pixels we use $y^- = -10$. The final loss is defined as an ℓ-1 loss between the estimation and the ground truth mask[1].

1.2 Training

Resolution Train/test on full resolution is important for active stereo system to get accurate depth (Sec. 2.2). However, during the training, the model cannot fit in a 12GB memory for a full resolution of 1280×720. To still enable the training in full resolution, we randomly pick a region with 1024×256 pixels, and crop from this same area from left and right images. This does not change the disparity, and thus models trained on small disparity can directly work on full resolution during the test.

[1] We also tried other classification losses, like the logistic loss, and they all led to very similar results.

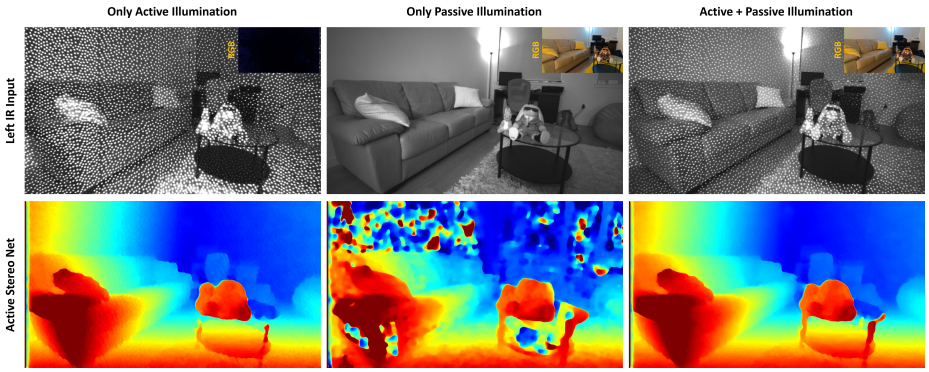

Fig. 2. Active vs Passive Illumination. Notice the importance of both: only active illumination exhibits higher level or noise, passive illumination struggles in textureless regions. When both the illumination are present we predict high quality disparity maps.

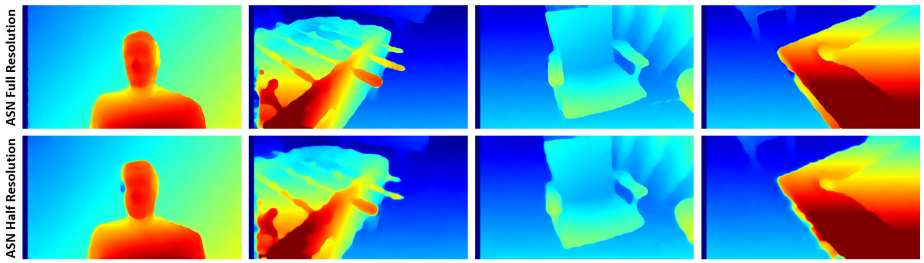

Fig. 3. Qualitative Evaluation Resolution. Notice how full resolution is needed to produce thin structures and crisp edges.

Invalidation Network Our invalidation network is trained fully self-supervised. The "ground-truth" is generated on the fly by using a hard left-right consistency check with a disparity threshold equal to 1. However, we found this function to be very unstable at the beginning of the training, since the network has not converged yet. To regularize its behavior, we first disable the invalidation network and only train the disparity network for 20000 iterations. After that we enable also the invalidation network. In practice we found this strategy is helpful for the disparity network to converge and prevent invalidation network from affecting the disparity.

2 Additional Evaluations

In this section we perform more ablation studies on different components of the network, we then show additional qualitative results and comparisons with other methods.

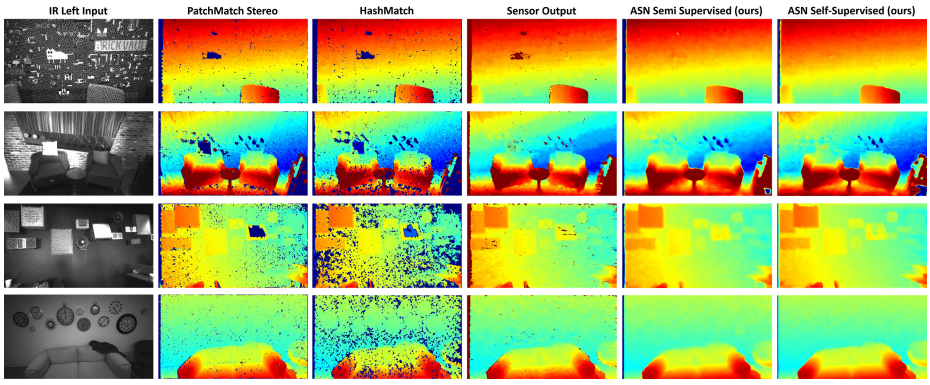

Fig. 4. We provide more qualitative examples. Notice in these challenging scenes how we provide a more complete output, crisper edges and less oversmoothed disparities.

2.1 Active + Passive Illumination

We first highlight the importance of having both active and passive illumination in our system. To show the impact of the two components, we run ASN on a typical living room scene with active only (i.e. all the lights in the room were off), passive only (laser from the sensor is off), and a combination of active and passive. The results are shown in Fig. 2. The result with only active illumination is noisier (see the presence of dot spike) and loses object edges (e.g. the coffee table) due to the strong presence of the dot pattern. The result from traditional passive stereo matching fails in texture-less regions, such as the wall. In contrast, the combination of the two gives a smooth disparity with sharp edges.

2.2 Image Resolution

We here evaluate the importance of image resolution for active stereo. We hypothesize that high-resolution input images are needed to generate high-quality disparity maps because any downsizing would lose considerable high frequency signals from the active pattern. To test this, we compare the performance of ASN trained with full and half resolution inputs in Fig. 3. Although deep architectures are usually good at super-resolution tasks, in this case there is a substantial degradation of the overall results. The thin structure and boundary details are missing in the ASN trained with half resolution. Please note that this result drives our decision to design a compact network architecture to avoid excessive memory usage.

2.3 Performance w.r.t. dataset size

Here we show the performance of the algorithm with respect to the dataset size. In Fig. 6, we show how our method is able to produce reasonable depthmaps on

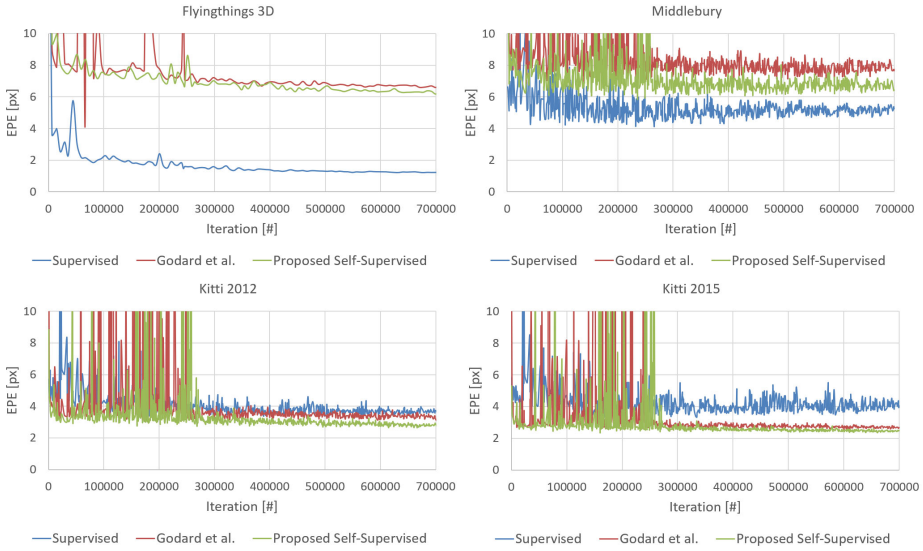

Fig. 5. Quantitative Evaluation - Passive Stereo. Our method outperforms the state of the art unsupervised method by Godard et al. [2] and it generalizes better than the supervised loss on the Kitti 2012 and Kitti 2015 datasets.

unseen images even when only a small dataset of 100 images is used. Increasing the training data size, leads to sharper results. We did not find any additional improvement when more than 10000 images are employed.

2.4 More Qualitative Results

Here we show additional challenging scenes and compare our method with local stereo matching algorithms as well as the sensor output. Results are depicted in Fig. 4. Notice how the proposed methods not only estimate more complete disparity maps, but suffer less from edge fattening and oversmoothing. Fig. 7 shows more results of the predicted disparity and invalidation mask on more diverse scenes.

3 Self-Supervised Passive Stereo

We conclude our study with an evaluation of the proposed self-supervised training method on RGB images. Although this is out of the scope of this paper, we show how the proposed loss generalizes well also for passive stereo matching problems. We consider the same architecture described in Sec. 1.1 and we train three networks from scratch on the SceneFlow dataset [5] using three different losses: supervised, Godard et al. [2] and our method. While training on SceneFlow, we evaluate the accuracy for the current iteration on the validation set

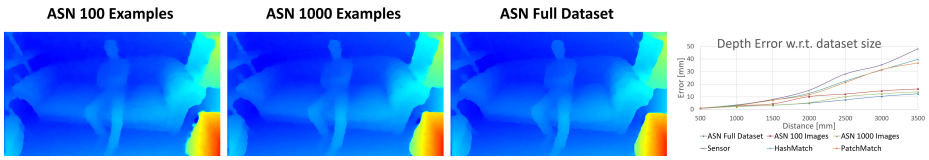

Fig. 6. Analysis of performance w.r.t data size. We found the model to be effective even with a small dataset. 100 images contain millions of individual pixels which are used as training examples. Increasing the dataset size leads to more accurate (sharper) depth maps, although the overall error does not improve that much. We did not observe any significant improvement beyond 10000 examples.

as well as on unseen datasets such as Middlebury [6], Kitti 2012 and Kitti 2015 [1, 4]. This is equivalent to testing our model on different datasets without fine-tuning. In Fig. 5 we show the EPE error of all the trained methods. Please note that the model trained using our loss achieves the lowest error on all datasets, which indicates that our method generalizes well. In particular on Kitti 2012 and Kitti 2015, we even outperform the supervised method trained on SceneFlow, which suggests the supervised loss is prone to overfitting.

Finally, we conclude this work showing qualitative results on passive RGB images. Fig. 8 we compare the three networks on the validation set used in SceneFlow (top two rows) as well as on unseen datasets (bottom 6 rows). Notice how our method always outperforms Godard et al. [2] in terms of edge fattening and frequent outliers on every dataset. Moreover, on the Kitti images, we do not suffer from gross errors such as the supervised model, which clearly overfit the original dataset.

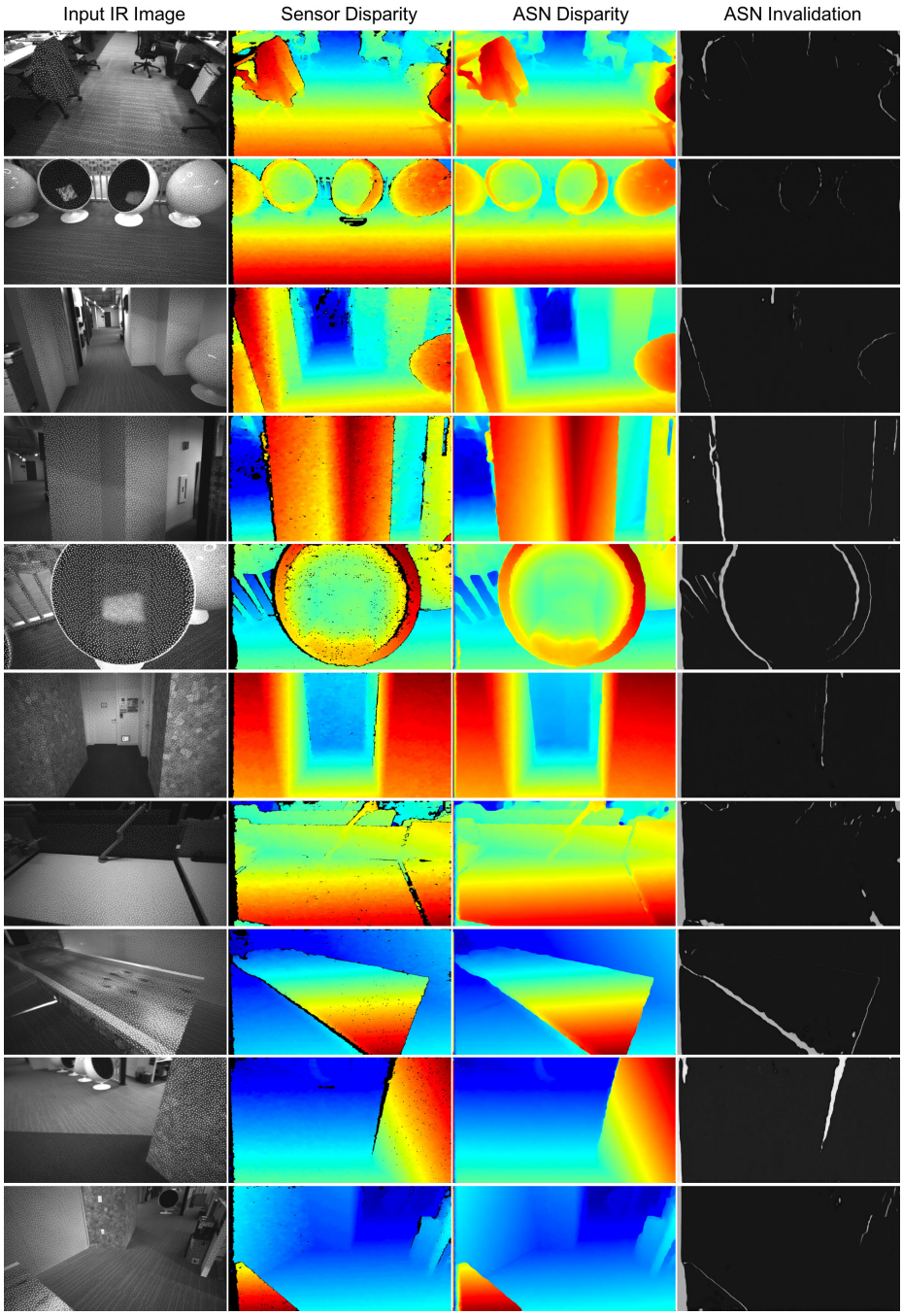

Fig. 7. Results of estimated disparity and invalidation mask on more diverse scenes.

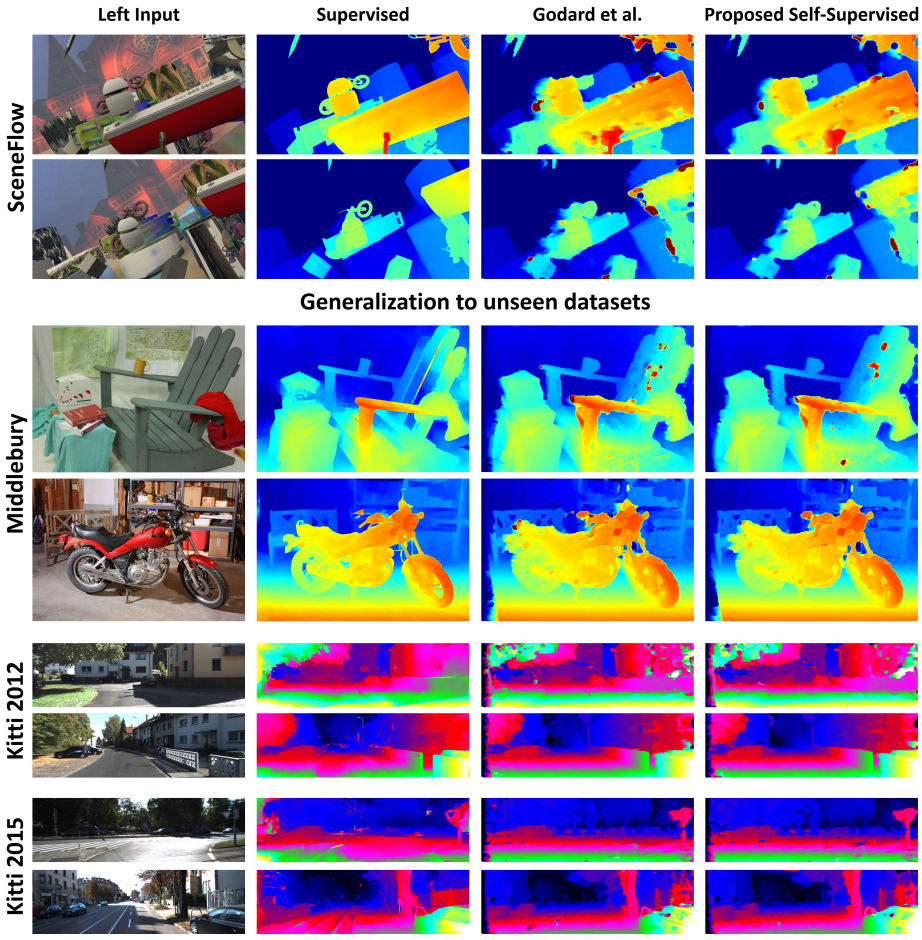

Fig. 8. Examples of qualitative results on RGB data using our method. We trained on SceneFlow dataset and tested on the others without any finetuning. Notice that for Kitti dataset we show better results compared to the supervised methods which suffer from gross errors due to overfitting.

References

1. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
2. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: CVPR. vol. 2, p. 7 (2017)
3. Khamis, S., Fanello, S., Rhemann, C., Valentin, J., Kowdle, A., Izadi, S.: Stereonet: Guided hierarchical refinement for edge-aware depth prediction. In: ECCV (2018)
4. Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
5. N.Mayer, E.Ilg, P.Häusser, P.Fischer, D.Cremers, A.Dosovitskiy, T.Brox: A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) (2016), http://lmb.informatik.uni-freiburg.de/Publications/2016/MIFDB16, arXiv:1512.02134
6. Scharstein, D., Hirschmuller, H., Kitajima, Y., Krathwohl, G., Nesic, N., Wang, X., Westling, P.: High-resolution stereo datasets with subpixel-accurate ground truth. In: GCPR (2014)