Supplemental material for Depth Reconstruction from Stereo Image Pairs

Xiuye Gu
xiuyegu@stanford.edu

1. More detailed qualitative results

For the training with mask version, qualitative results on SceneFlow [3] are shown in Fig 1 and Fig 2. The qualitative results on KITTI [1, 2] are shown in 3, including samples from both training set and testing set.

When talking about the error when using fixed mean and variance during training, one bad example is shown in Fig 4.

To try to obtain better results, we remove the masks, and train the model from scratch. This time, the visualized results are better, and are shown in Fig 5. But time is limited, we cannot obtain the new qualitative results before deadline, so we represent some temporary qualitative results (The model is still training, we show the training samples prediction results).

2. Some thoughts

Through my exploring on the two stereo matching approaches, one traditional, one using deep neural network. The traditional one requires much trials and errors, and it is hard to outperform the deep approaches (according to the leaderboards, the methods ranked higher are all CNN-heavy, methods without using CNN's ranks are larger than 10). However, the deep approach, based on my experience, is very tricky, and requires great deep neural networking tuning techniques. The model architectures are the same, most details are the same, why the results are so different is beyond me (currently, the training results of the training without mask version seem still cannot reproduce the high performance of the original paper).

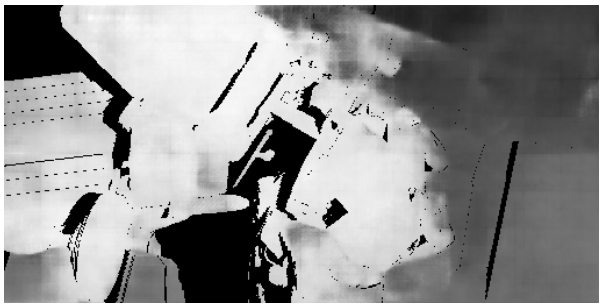
References

[1] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

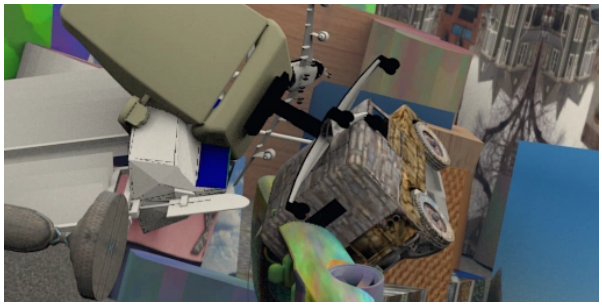
[2] M. Menze and A. Geiger. Object scene flow for autonomous vehicles. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[3] N.Mayer, E.Ilg, P.Häusser, P.Fischer, D.Cremers, A.Dosovitskiy, and 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. arXiv:1512.02134.

(a) Ground truth 1

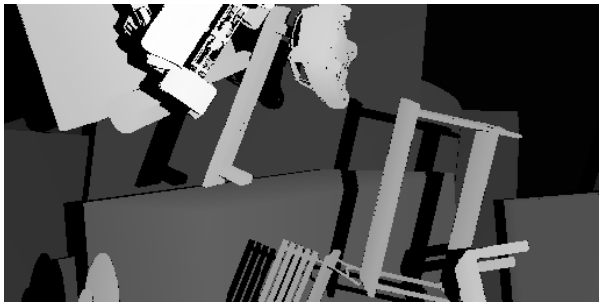


(b) Masked prediction 1

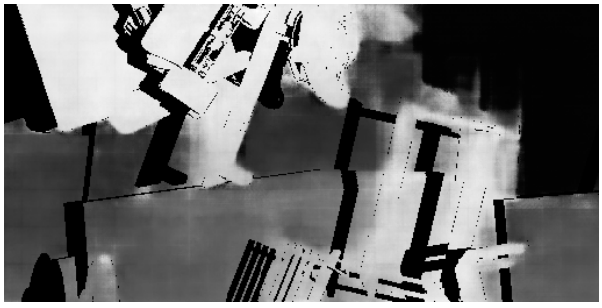


(c) Original image 1

(d) Unmasked prediction 1

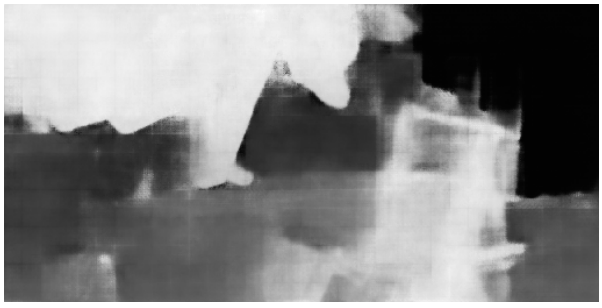


(e) Ground truth 2



(f) Masked prediction 2

(g) Original image 2



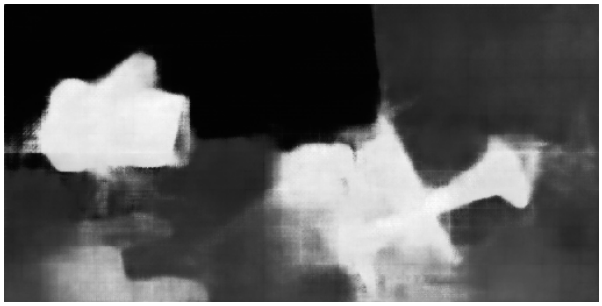
(h) Unmasked prediction 2

Figure 1. GC-Net qualitative results on SceneFLow, train with masks.

(a) Ground truth 3

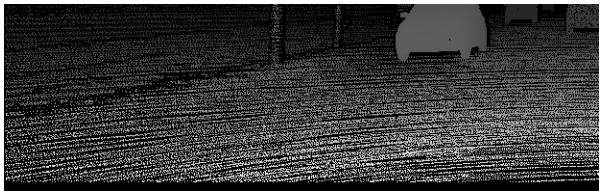
(b) Masked prediction 3

(c) Original image 3

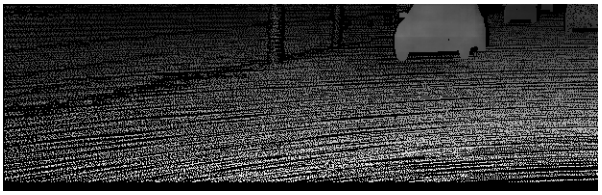


(d) Unmasked prediction 3

Figure 2. GC-Net qualitative results on SceneFLow, train with masks. Continued.

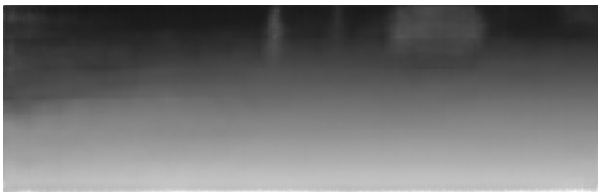


(a) Perform on training set, ground truth 1



(b) Perform on training set, masked prediction 1

(c) Perform on training set, original image 1



(d) Perform on training set, masked prediction 1

(e) Perform on training set, ground truth 2

(f) Perform on training set, masked prediction 2

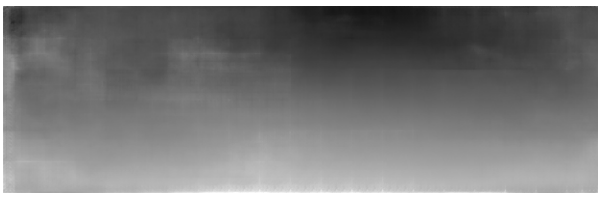
(g) Perform on training set, original image 2

(h) Perform on training set, masked prediction 2

(i) Perform on test set, original image 1

(j) perform on test set, prediction 1

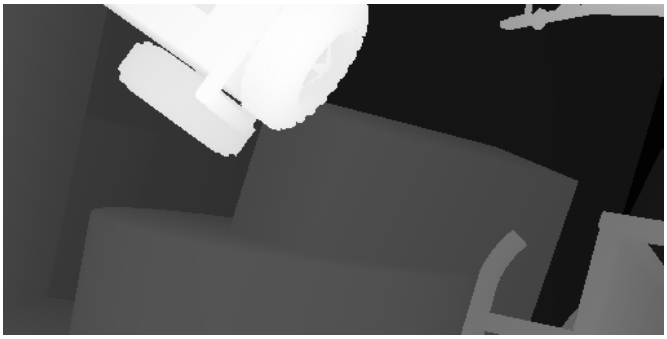
(k) Perform on test set, original image 2



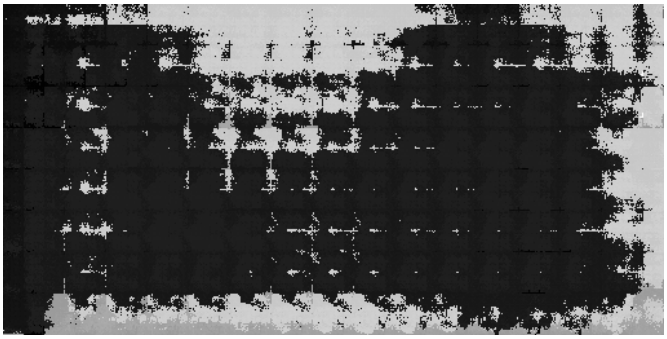
(l) Perform on test set, prediction 2

Figure 3. GC-Net qualitative results on KITTI, train with masks.

(a) Train with mask, bad when using fixed mean/var, original image



(b) Train with mask, bad when using fixed mean/var, groundtruth

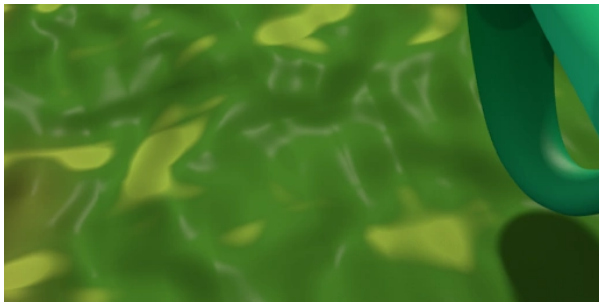


(c) Train with mask, bad when using fixed mean/var, prediction

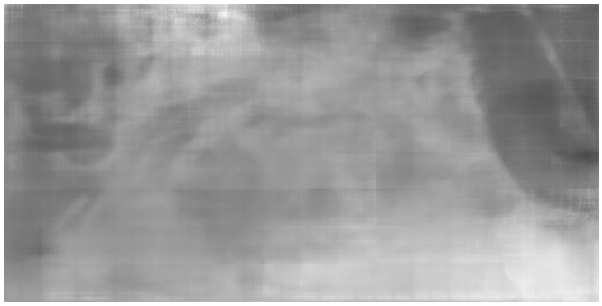
Figure 4. Bad example when fixing mean and variance on SceneFLow.

(a) Original image 1

(b) Prediction 1



(c) Original image 2



(d) Prediction 2

Figure 5. GC-Net qualitative results on SceneFLow, train without masks, temporary prediction results.