Supplementary Material for "Cascade Residual Learning: A Two-stage Convolutional Neural Network for Stereo Matching"

Jiahao Pang* Wenxiu Sun* Jimmy SJ. Ren Chengxi Yang Qiong Yan
SenseTime Group Limited
{pangjiahao, sunwenxiu, rensijie, yangchengxi, yanqiong}@sensetime.com

1. Introduction

In this supplementary material, we have included the specification of *DispFulNet* (our first-stage network) and some miscellaneous training details of our network. We have also showcased more visual results of the proposed *cascade residual learning* (CRL) scheme.

2. Specification of DispFulNet

The detailed specification of the first stage, DispFulNet, is provided in Table. 1. The disparity images are produced by the convolution layers with prefix `pr_` at multiple scales, they are supervised by the ground-truth by computing the ℓ_1 loss. The final disparity prediction of DispFulNet, d_1, is given by the layer `pr_1`; this layer is named as `pr_s1` in our paper (`s1` means stage 1). Note that different from [4], the proposed DispFulNet has an output resolution equals to the input resolution, gives rise to disparity images with extra details.

Having obtained the first-stage output d_1, it is then concatenated with the left image I_L, the right image I_R, the synthesized left image (using the warping layer) $\widetilde{I}_L$, and the error image $e_L = |I_L - \widetilde{I}_L|$. The obtained 3D array is then fed to the second-stage network (DispResNet) to further refine the disparity.

Layer	K	S	Channels	I	O	Input Channels
conv1a	7	2	3/64	1	2	left
conv1b	7	2	3/64	1	2	right
conv2a	5	2	64/128	2	4	conv1a
conv2b	5	2	64/128	2	4	conv1b
corr	1	1	256/81	4	4	conv2a+conv2b
conv_rdi	1	1	128/64	4	4	conv2a
conv3	5	2	145/256	4	8	corr+conv_rdi
conv3_1	3	1	256/256	8	8	conv3
conv4	3	2	256/512	8	16	conv3_1
conv4_1	3	1	512/512	16	16	conv4
conv5	3	2	512/512	16	32	conv4_1
conv5_1	3	1	512/512	32	32	conv5
conv6	3	2	512/1024	32	64	conv5_1
conv6_1	3	1	1024/1024	64	64	conv6
pr_64	3	1	1024/1	64	64	conv6_1
upconv6	4	2	1024/512	64	32	conv6_1
iconv6	3	1	1023/512	32	32	upconv6+conv5_1+pr_64
pr_32	3	1	512/1	32	32	iconv6
upconv5	4	2	512/256	32	16	iconv6
iconv5	3	1	769/256	16	16	upconv5+conv4_1+pr_32
pr_16	3	1	256/1	16	16	iconv5
upconv4	4	2	256/128	16	8	iconv5
iconv4	3	1	385/128	8	8	upconv4+conv3_1+pr_16
pr_8	3	1	128/1	8	8	iconv4
upconv3	4	2	128/64	8	4	iconv4
iconv3	3	1	193/64	4	4	upconv3+conv2a+pr_8
pr_4	3	1	64/1	4	4	iconv3
upconv2	4	2	64/32	4	2	iconv3
iconv2	3	1	97/32	2	2	upconv2+conv1a+pr_4
pr_2	4	1	32/1	2	2	iconv2
upconv1	4	2	32/16	2	1	iconv2
pr_1	5	1	20/1	1	1	upconv1+left+pr_2

Table 1. Detailed architecture of the proposed *DispFulNet*. The layer `corr` is the correlation layer of [4] with maximum displacement 40, while layers with prefix `pr_` are convolution layers output disparity images at multiple scales. **K** means kernel size, **S** means stride, and **Channels** is the number of input and output channels. **I** and **O** are the input and output downsampling factor relative to the input. The symbol + means concatenation.

3. Miscellaneous Training Details

In general, we train our network in a way similar to that of [1, 4]. Nevertheless, for simplicity, we have assigned different (fixed) weights to different ℓ_1 loss layers, in contrast to the loss weight schedule of [4]. Specifically, when training either the first stage (DispFulNet) or the second stage (DispResNet), we assign the highest resolution loss, *e.g.*, `pr_1` in Table. 1, with fixed weight 1. For other losses we let their weights be 0.2. When finetuning the overall network, we assign the highest resolution loss at the second stage with a weight 1; for other losses of the second stage and the loss of `pr_1` (the highest resolution loss of Disp-

*Both authors contributed equally.

FulNet), their losses have weights 0.2. We let the weights of the rest of the loss layers in DispFulNet be 0.1.

4. More Visual Results

Our paper has shown many disparity segments produced by the proposed method, so as to demonstrate its superior characteristics. As supplement, we hereby showcase a few intact disparity images. Fig. 1 and Fig. 2 show three groups of visual reuslts of the FlyingThings3D dataset [4] and the Middlebury 2014 dataset [6], respectively. DispNetC [4] is adopted as a baseline method for visual comparisons in these two figures. We remind that the Middlebury dataset has not been applied for training in this work. From the disparity images and the error images, one can see that our CRL scheme not only produces sharp edges but also provides more accurate disparity values at the inherently ill-posed regions.

Fig. 3 presents four groups of visual results on the test set of KITTI stereo 2015 dataset [5]. Our approach, CRL, ranks *first* (with a D1-all of 2.67%) in the KITTI 2015 stereo benchmark; while another approach, GC-NET [3], ranks second (with a D1-all of 2.87%). As a result, we compare our results with those of GC-NET here. In Fig. 3, the disparity images and the error images are obtained from the KITTI evaluation website, where the disparity images are shown using the color coding scheme of [2]. For the error images, warmer color indicate larger error. Again, our proposed CRL scheme produces sharp disparity images with high accuracy.

References

[1] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. van der Smagt, D. Cremers, and T. Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2758–2766, 2015. 1

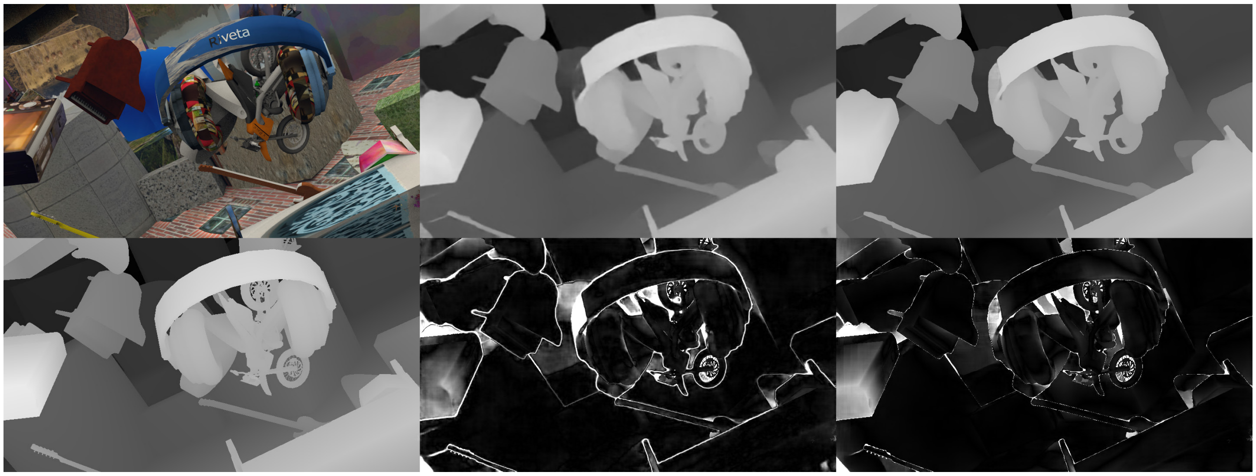
[2] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012. 2

[3] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, and A. Bry. End-to-end learning of geometry and context for deep stereo regression. *arXiv preprint arXiv:1703.04309*, 2017. 2

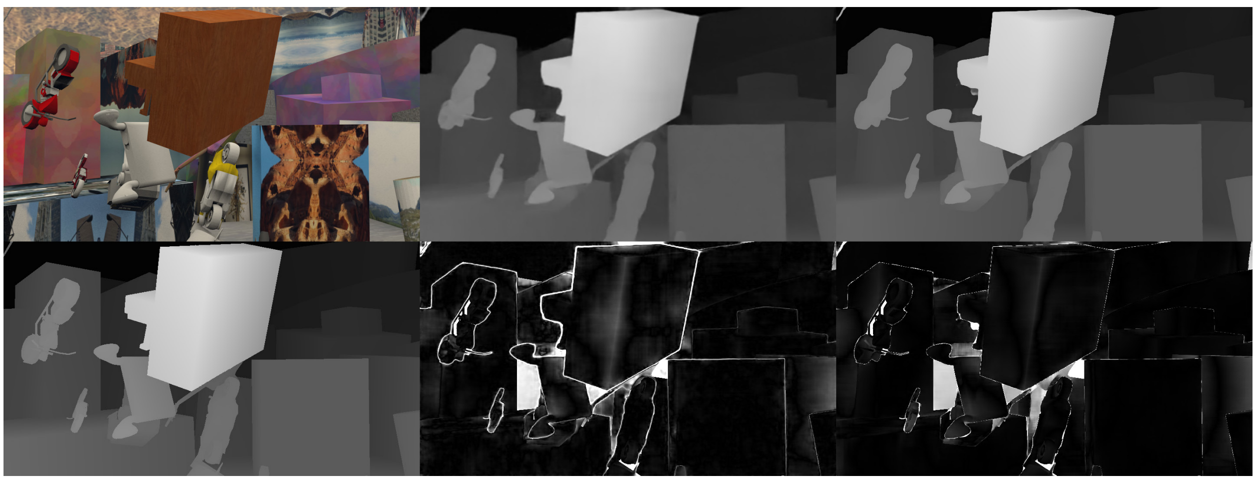
[4] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4040–4048, 2016. 1, 2

[5] M. Menze and A. Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 2

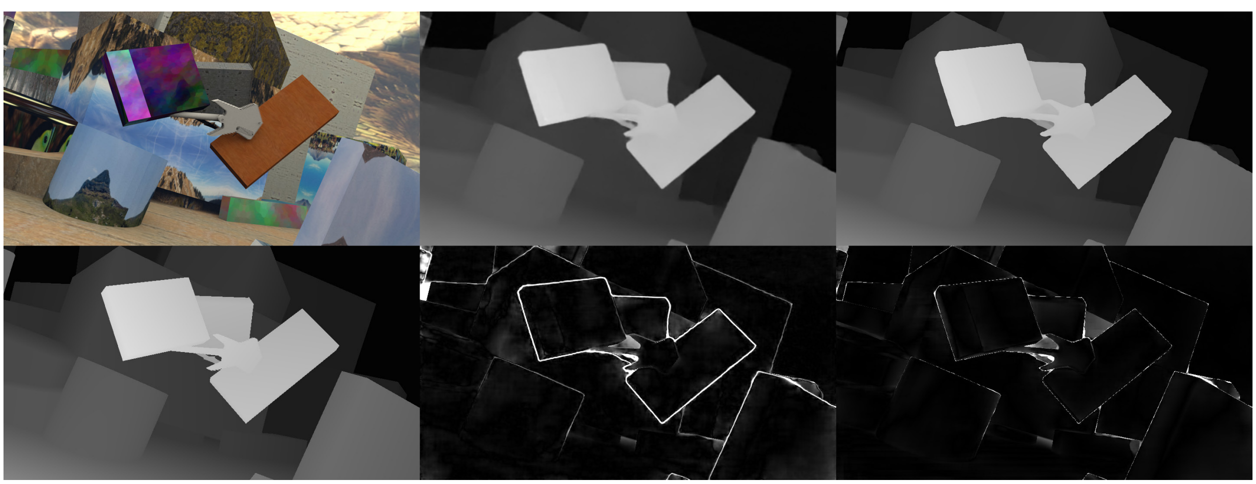
[6] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, and P. Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *German Conference on Pattern Recognition*, pages 31–42. Springer, 2014. 2



(a)



(b)



(c)

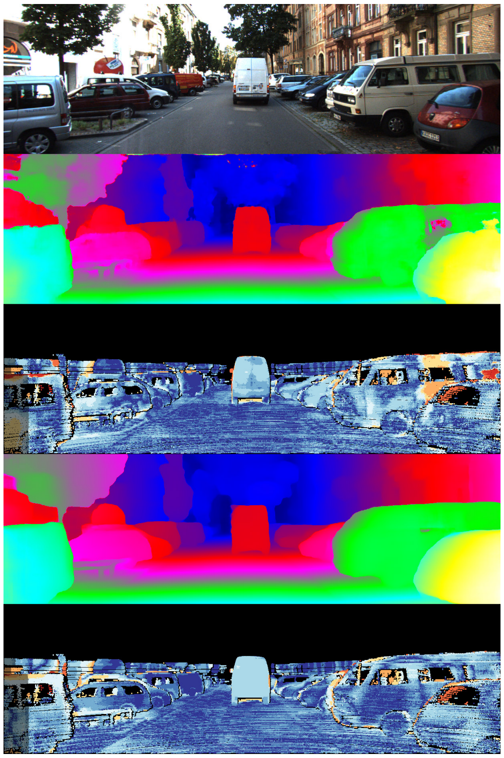
Figure 1. Three groups of results from the FlyingThings3D dataset are shown. The six images in each group, from left to right and from top to bottom, are the left image, the disparity images given by DispNetC and CRL (ours), the ground-truth disparity, the error images of DispNetC and CRL, respectively.

(a)

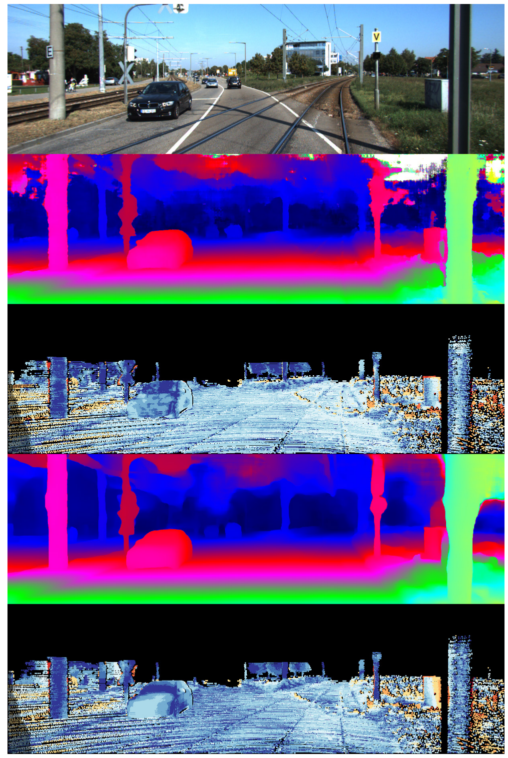
(b)

(c)

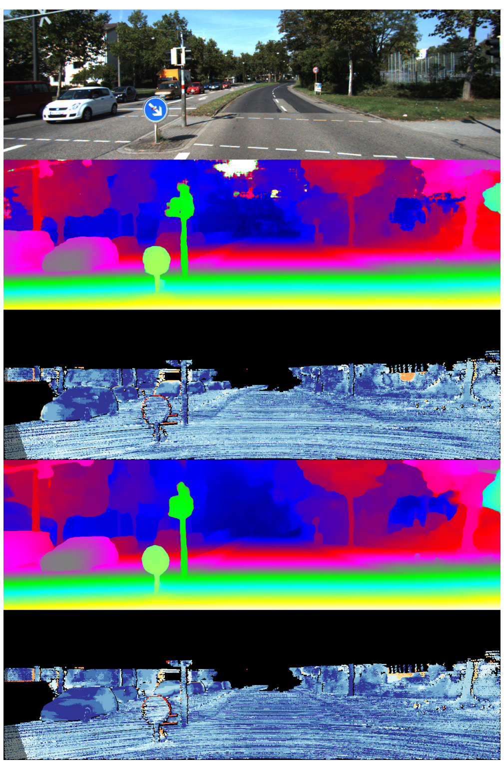
Figure 2. Three groups of results from the Middlebury 2104 dataset are shown. The six images in each group, from left to right and from top to bottom, are the left image, the disparity images given by DispNetC and CRL (ours), the ground-truth disparity, the error images of DispNetC and CRL, respectively.



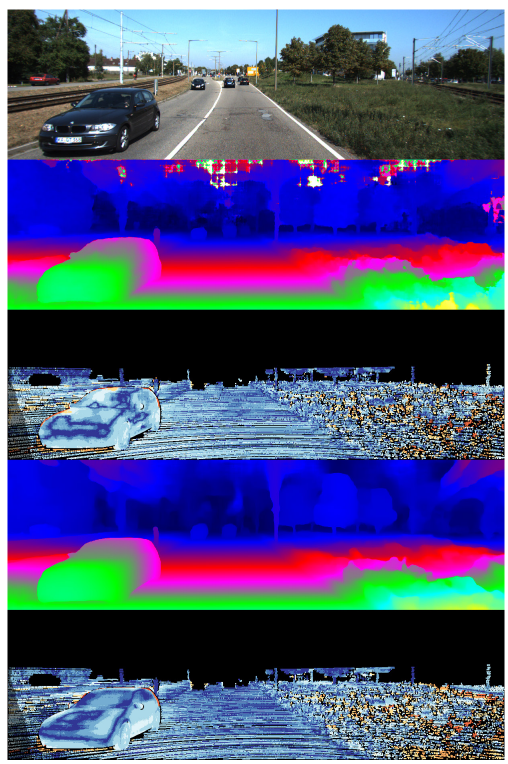
(a)



(b)



(c)



(d)

Figure 3. Four groups of results from the KITTI stereo 2015 dataset are shown. The five images in each group, from top to bottom, are the left image, the disparity image given by GC-NET and its error image, the disparity image given by the proposed CRL and its error image, respectively. For the error images, warmer colors indicate larger error.