Supplementary Material for "Zoom and Learn: Generalizing Deep Stereo Matching to Novel Domains"

Jiahao Pang[1] Wenxiu Sun[1] Chengxi Yang[1] Jimmy Ren[1] Ruichao Xiao[1] Jin Zeng[1] Liang Lin[1,2]
[1]SenseTime Research [2]Sun Yat-sen University
{pangjiahao, sunwenxiu, yangchengxi, rensijie, xiaoruichao, zengjin, linliang}@sensetime.com

1. Introduction

In this supplementary material, we provide more discussions on the phenomenon of scale diversity. We also derive the backward computation of the proposed graph Laplacian regularization loss. We then showcase more visual comparisons of the proposed method, zoom and learn (ZOLE), for self-adaptation.

2. More Discussions on Scale Diversity

In addition to stereo matching, scale diversity can also be observed in other pixel-wise regression/classification problems employing convolutional neural networks (CNNs), *e.g.*, optical flow estimation [2, 5] and semantic segmentation [6].

Optical flow estimation is a problem closely related to stereo matching, where given two frames at different time instants, a per-pixel optical flow field is estimated. Similar to the setup of stereo matching, we adopt the representative *FlowNet* [2] architectures—both the one with explicit 2D correlation (FlowNetC) and the one with only convolution (FlowNetS)—to see the scale diversity in optical flow estimation. Particularly, the released FlowNetC and FlowNetS models, trained with the synthetic Flying Chairs dataset [2], are adopted. Besides, we employ the training split of the Middlebury optical flow dataset [1] for investigation. All of its images are first resized to an original size of 576×448. To estimate the optical flow of an image pair at a finer scale, we first up-sample (zoom in) the image pair by r times ($r > 1$). The up-sampled image pair is then fed to the CNN, leading to a flow at a higher resolution. By down-sampling the flow field by r times and also re-scale its values by a factor of $1/r$, one obtains an optical flow estimate at a finer scale r.

Figure 1 shows several optical flow fields obtained by feeding image pairs at different resolutions, *i.e.*, $576r \times 448r$, $r \in \{1, 3, 5\}$, to the released FlowNetC model. Similar to the observations about stereo matching, we see that as r increases, more and more high-frequency details emerge on the output optical flow fields. However, a bigger r, *i.e.*, a finer scale, does not necessarily translate to a better performance. To see this, we measure the average endpoint error (EPE) between the output flows (estimated at different scales) and the ground-truth flows on the Middlebury optical flow dataset. The obtained objective performance are listed in Table 1. Similar to our observations about stereo matching, we see that as the resolution grows, the network performance first improves (by virtue of a finer-grain esti-

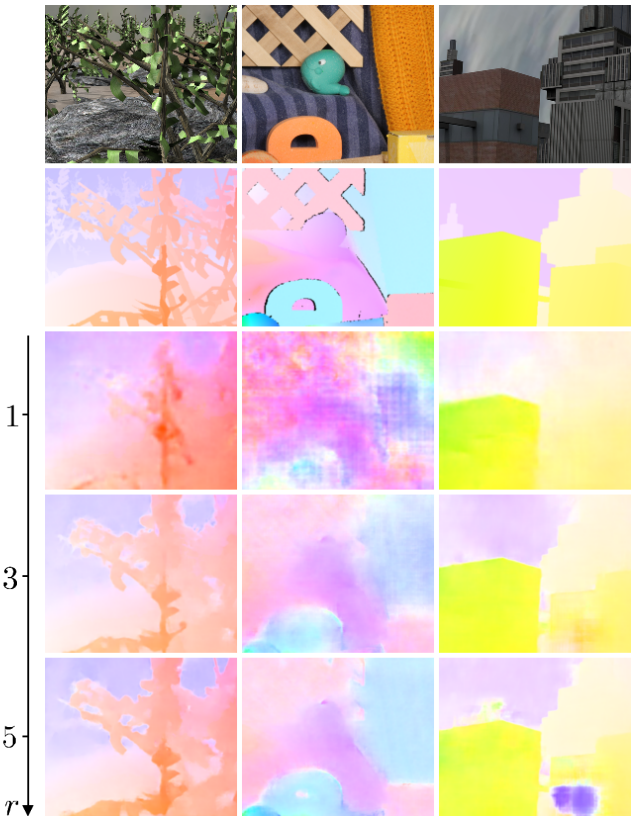


Figure 1. For the same image pair, feeding its zoomed-in version to an optical flow estimation CNN leads to results with extra details. The five rows are the first image, the ground-truth, and the resulting flows obtained with up-sampling ratios $r = 1, 3, 5$, respectively.

Table 1. The endpoint errors of the released FlowNetC and FlowNetS models on the training set of the Middlebury optical flow data. A resolution of r means the stereo pairs are up-sampled to $576r \times 448r$ before passing to the CNNs.

Network	Resolution					
	1	2	3	4	5	6
FlowNetC	1.30	0.86	0.71	**0.69**	0.79	0.87
FlowNetS	1.27	0.81	0.69	**0.64**	0.72	0.82

mation process) then deteriorates (due to the shrinking receptive field).

Different from stereo matching and optical flow estimation, semantic segmentation takes as input an image then performs pixel-wise classification. We find that the phenomenon of scale diversity also exists in semantic segmentation. For illustration, we adopt the notable fully convolutional neural network (FCN) architecture, FCN-8s [6], and perform the following tests. Particularly, we use the released FCN-8s model pre-trained on the PASCAL VOC dataset (with the 8498 training examples used by the work [4]). It is tested on the validation split (736 images) provided by the authors of [6]. To obtain the segmentation of an image at a finer scale, one simply up-samples the original image by a ratio of r and pass it to the network. The resulting segmentation is then down-sampled by a factor of r with nearest neighbor down-sampling, so that it has the same size as the original image. Figure 2 shows a few segmentations estimated with different scales, *i.e.*, $r \in \{1, 1.5, 2, 2.5\}$. We see that as r grows, more thin details are produced. However, for an r that is too large, *i.e.*, the network becomes too localized, wrong classifications start to emerge.

Hence, it is possible to exploit scale diversity to improve the performance of CNNs for optical flow estimation and semantic segmentation. We leave this topic for future research.

3. Backpropagation of the Graph Laplacian Regularization Loss

We hereby derive the backward pass of our proposed graph Laplacian regularization loss. Given a disparity map $S_i \equiv S(P_i; \mathbf{\Theta})$ generated by a deep stereo model $S(\cdot; \mathbf{\Theta})$, its graph Laplacian regularization loss is given by

$$\mathcal{L}_{\mathcal{G}}(S_i) = \lambda \cdot \sum_{j=1}^{M} \mathbf{s}_{ij}^{\mathrm{T}} \mathbf{L}_{ij}^{(k)} \mathbf{s}_{ij}.$$

Be reminded that λ is a constant, M is the number of patches tiling the whole disparity map, $\mathbf{s}_{ij} = \mathbf{R}_j \cdot \mathrm{vec}(S_i) \in \mathbb{R}^m$ (*i.e.*, the j-th patch of S_i), $\mathbf{L}_{ij}^{(k)}$ is the pre-computed graph Laplacian matrix for regularizing $\mathbf{s}_{ij}$ at the k-th iteration.

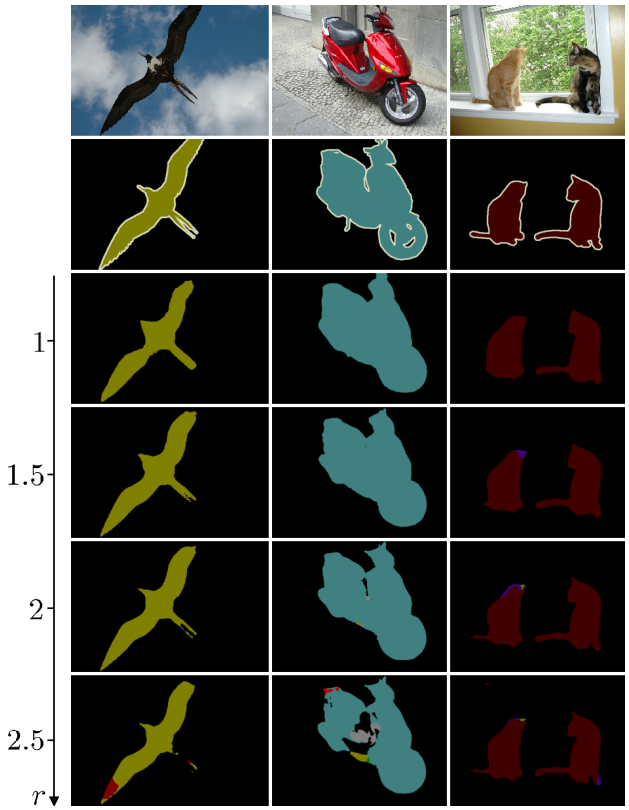


Figure 2. Feeding the up-sampled versions of an image to the pre-trained FCN-8s model leads to segmentation results with more fine details. However, the results deteriorate if the up-sampling ratio r is too large. The six rows are the original image, the ground-truth segmentation, and the segmentations obtained by up-sampling ratios $r = 1, 1.5, 2, 2.5$, respectively.

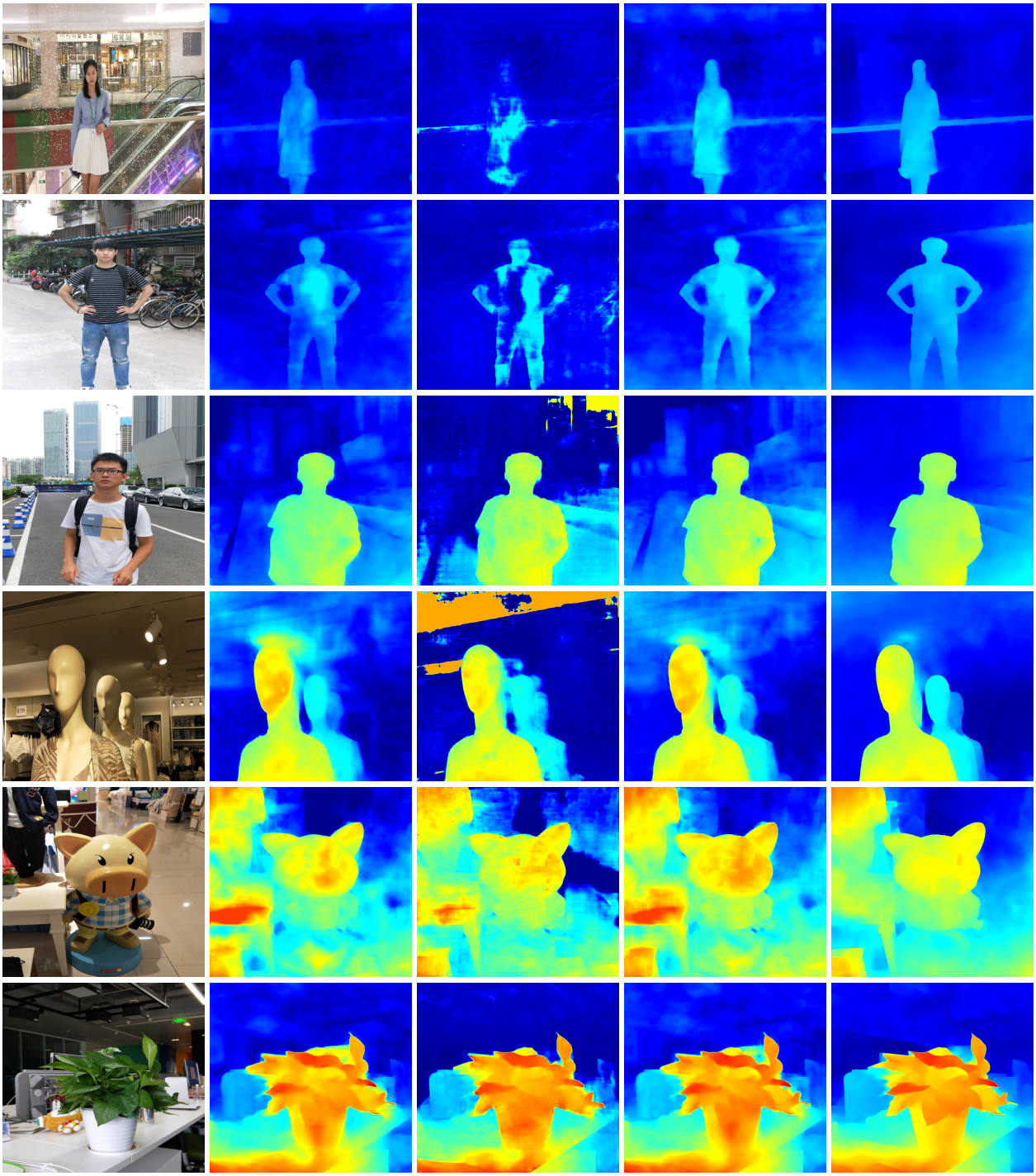
We denote the n-th pixel on a particular patch $\mathbf{s}_{ij}$ as $p_{ij}^{(n)}$, then the partial derivative of $\mathcal{L}_{\mathcal{G}}$ with respect to $p_{ij}^{(n)}$ can be derived as follows,

$$\frac{\partial \mathcal{L}_{\mathcal{G}}}{\partial p_{ij}^{(n)}} = \lambda \left(\frac{\partial \mathbf{s}_{ij}^{\mathrm{T}} \mathbf{L}_{ij}^{(k)} \mathbf{s}_{ij}}{\partial \mathbf{s}_{ij}} \right)^{\mathrm{T}} \cdot \frac{\partial \mathbf{s}_{ij}}{\partial p_{ij}^{(n)}} = 2\lambda \mathbf{s}_{ij}^{\mathrm{T}} \mathbf{L}_{ij}^{(k)} \mathbf{1}_n,$$

where $\mathbf{1}_n \in \mathbb{R}^m$ is an indication vector, its n-th entry equals one and the rest entries are zeros.

4. More Visual Results

In this section, we present more visual results of our proposed zoom and learn (ZOLE) approach. We employ the same settings as described in the paper, *i.e.*, we generalize the pre-trained DispNetC model to two different domains: daily scenes collected by smartphones and street views captured from the perspective of a driving car. Visual comparisons on the test split of our smartphone dataset are shown in Figure 3. As can be seen, our obtained disparity maps change smoothly for regions within the same object, while



Left image | Tonioni *et al.* [8] | DispNetC | ZOLE-S | ZOLE

Figure 3. Visual comparisons of different models on the test set of our collected smartphone data. This figure shows the left images and the corresponding disparity maps obtained with four different models. We see that the proposed ZOLE approach produces smooth disparity maps with sharp edges at object boundaries.

they have sharp edges at object boundaries. Our method is also capable of producing fine details that are far away from the cameras, *i.e.*, with small disparity values. Figure 4 shows the visual results on the training split (with sparse ground-truths) of the KITTI stereo 2015 dataset [3]. We see that compared to the other CNN models, the one ob-

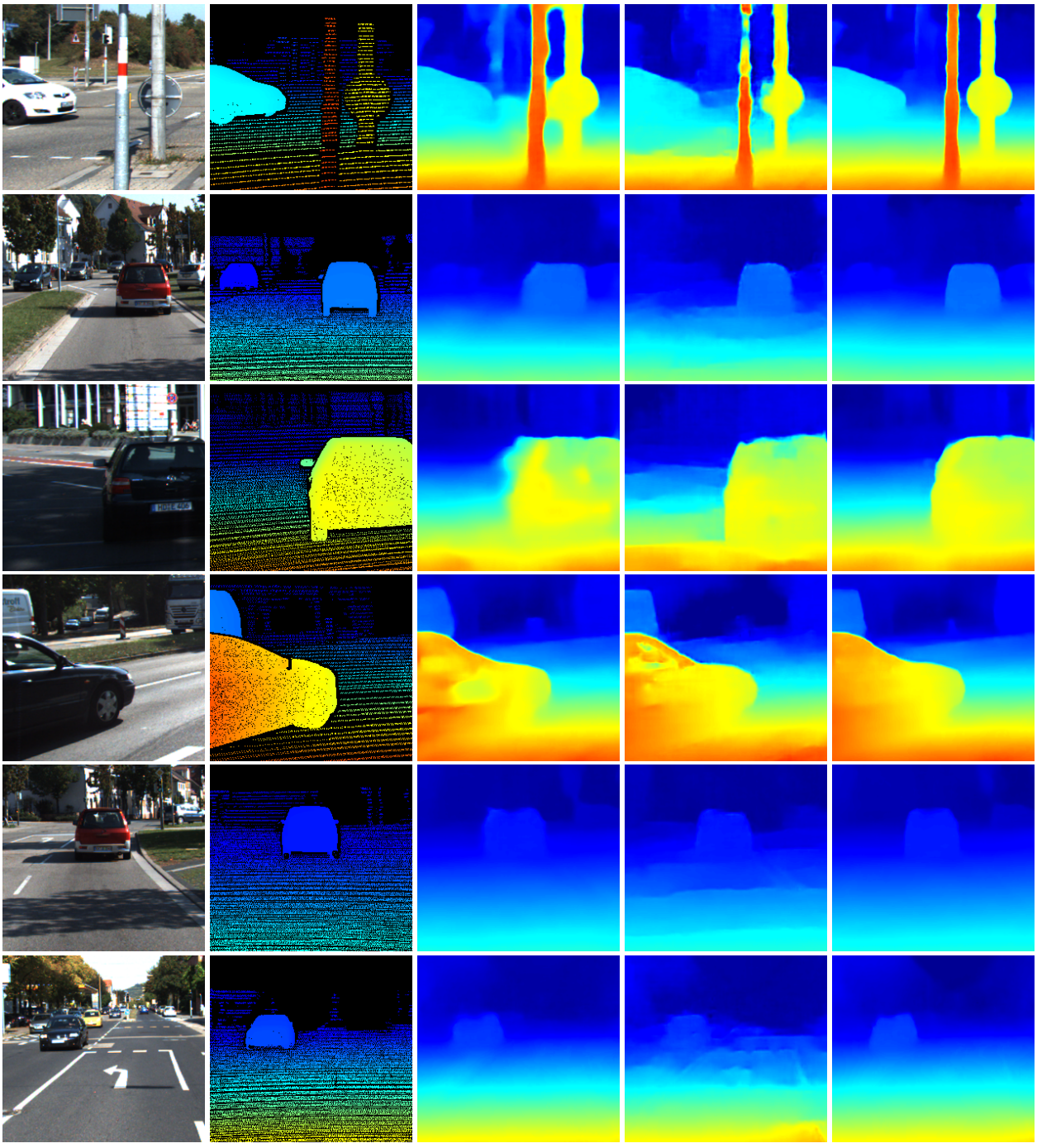


Figure 4. Visual comparisons of different models on the KITTI stereo 2015 dataset. The five columns of this figure shows the fragments of the left images, sparse ground-truth disparity maps, the disparity maps produced by [8], the released DispNetC model pre-trained with FlyingThings3D [7], and the proposed ZOLE approach, respectively. We see that our method provides high-fidelity disparity maps.

tained by our ZOLE approach provides high-fidelity disparity maps.

References

[1] S. Baker, D. Scharstein, J. Lewis, S. Roth, M. J. Black, and R. Szeliski. A database and evaluation methodology for optical flow. *International Journal of Computer Vision*, 92(1):1–

31, 2011. 1

[2] 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 *The IEEE International Conference on Computer Vision (ICCV)*, pages 2758–2766, 2015. 1

[3] 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 (CVPR)*, pages 3354–3361. IEEE, 2012. 3

[4] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik. Semantic contours from inverse detectors. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 991–998. IEEE, 2011. 2

[5] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. FlowNet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1

[6] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015. 1, 2

[7] 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 (CVPR)*, pages 4040–4048, 2016. 4

[8] A. Tonioni, M. Poggi, S. Mattoccia, and L. Di Stefano. Unsupervised adaptation for deep stereo. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 3, 4