PARALLAX CONTEXTUAL REPRESENTATIONS FOR STEREO MATCHING

Hui Deng[1], Qingmin Liao[1], Zongqing Lu[1], Jing-Hao Xue[2]

[1]Shenzhen International Graduate School, Tsinghua University, China
[2]Department of Statistical Science, University College London, UK

ABSTRACT

In this work, we study the context aggregation in stereo matching from a new parallax perspective. Unlike previous works, we propose to characterize and augment a pixel with its parallax contextual representation (PCR), which has not been explored before. We also propose a new concept called disparity prototype to describe the overall representation of a disparity plane. Our proposed PCR module consists of three steps: 1) divide disparity planes for a rough estimation of disparity; 2) estimate the disparity prototypes for each disparity plane; 3) derive PCR-augmented representations with disparity prototypes. Extensive experiments on various datasets using different networks validate the effectiveness of our proposal.

Index Terms— Stereo matching, parallax contextual representation, disparity prototype, disparity plane

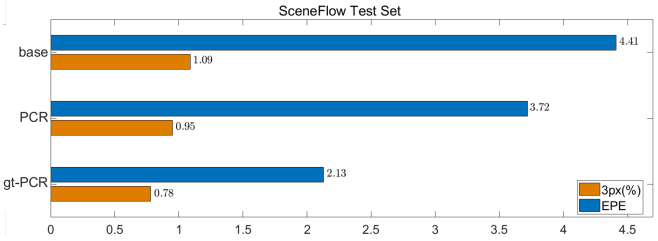

Fig. 1. Effectiveness of the proposed PCR module, compared with the baseline network (PSMNet). For illustration and comparison, the ground-truth disparity map is used to build a method called gt-PCR, which forms completely correct matching probability maps. We can observe that the PCR module performs better than the baseline. Moreover, gt-PCR can produce remarkably excellent results, which shows the potential of PCR. The 3-pixel threshold error rate 3px(%) and the average end point error (EPE) are used to evaluate the performance of these methods.

1. INTRODUCTION

Recent efforts using CNNs to learn powerful representations from data significantly promote the performance of stereo matching. DispNetC [1] builds the first end-to-end network for stereo matching and proposes a correlation layer, following which many methods directly regress disparity maps [2, 3]. GCNet [4] takes a different approach, which concatenates left features and shifted right features to generate a 4D feature volume and employes 3D CNN to aggregate contextual cues. Based on 3D convolution, many other methods, including PSMNet [5], StereoNet [6], AnyNet [7], GANet [8], and EMCUA [9], achieve state-of-the-art performance on benchmarks. Particularly, GwcNet [10] and AMNet [11] explore a middle way between using the correlation of features and directly concatenating the features.

The 3D convolution based methods generally outperform the 2D methods due to the use of global semantic context. However, relying solely on fixed-size convolutional kernel to model the contextual information is inefficient. Firstly, fixed-size 3D filtering inevitably involves irrelevant pixels, causing a well-known edge-fattening issue in object boundaries and thin structures [12]. Secondly, local contextual representation usually fails to address long-range dependencies [13, 14]. Considering these drawbacks, an intuitive solution is to investigate a better contextual representation scheme with higher weights for similar features.

Semantic segmentation, another fundamental task in computer vision, also requires a good understanding of context. Several recent representative semantic-segmentation methods [15, 16, 17] try to extract the global context from a categorical perspective: to exploit an aggregated representation of the object region that a pixel belongs to. Motivated by their success, this paper aims to answer a question for stereo matching: *Can the representations aggregating pixels at the same disparity level help stereo matching?*

Disparity planes are the counterpart of a set of discrete disparity candidates. Empirically speaking, pixels from the same disparity plane have strong correlation since they are assigned to close disparity values. Such a parallax contextual information, a factor largely ignored before, should be explored for better stereo matching. Hence, *for the first time*, we present a parallax contextual representation (PCR) approach in this work. Fig.1 well verifies the effectiveness of this new representation augmentation scheme.

The novelties and contributions of our proposed approach can be described as follows.

1) We propose a new parallax contextual representation

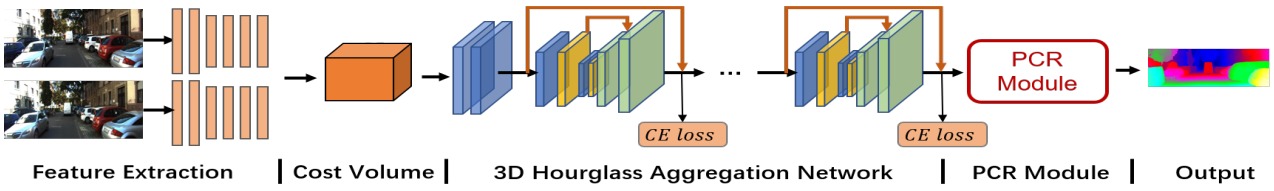

Fig. 2. Overview of the network architecture. The PCR is added to the PSMNet [5] backbone for parallax contextual aggregation. Given a stereo pair, PSMNet extracts the features of the left and right images and forms three cost volumes with a stacked hourglass architecture. For each cost volume, we supervise it with a cross-entropy loss by using quantified disparity ground truth. The inputs of PCR is a 4D feature volume and a 3D cost volume from the last hourglass module. The output of PCR module is an augmented representation for each pixel, which is fed to 1×1 convolutions to obtain a new refined cost volume.

(PCR), which augments the representation of each pixel with its contextual information extracted from disparity planes.

2) We propose the concept of disparity prototype to describe the overall representation of a disparity plane. Specifically, the disparity prototype is the aggregation of all features of pixels belonging to a disparity plane.

3) Different from existing methods [4, 5], which minimize the distance between the ground truth and the mean value of estimated disparity, we supervise the model to learn the cost distribution peaking around the ground truth, which provides distinguish disparity planes for the PCR module.

4) We integrate our PCR into different stereo networks [5, 10] and evaluate them on various stereo benchmarks. Ablation studies on the SceneFlow dataset [1] demonstrate the effectiveness of the proposed PCR, and our method largely improves the performance of previous methods.

2. METHODS

Fig.2 shows the overall network architecture used in this paper. We choose PSMNet [5] as the backbone network for illustration. The parts of feature extraction and cost volume construction retain the identical structure with PSMNet, but we make some modifications to the stacked hourglass modules, which will be discussed in Section 2.2.

2.1. PCR Module

As shown in Fig.3, the inputs of the proposed PCR module includes a 4D feature volume $\mathbf{A} \in \mathbb{R}^{H \times W \times D \times C}$ and a 3D cost volume $\mathbf{B} \in \mathbb{R}^{H \times W \times D}$, both of which are outputs from last hourglass module.

The proposed parallax contextual representation scheme contains three steps: 1) structure all the pixels in the reference image I into D disparity planes and compute the matching probability map M_d; 2) estimate the disparity prototype $\boldsymbol{x}_d$ for each disparity plane by aggregating the representations of all the pixels in the disparity plane π_d; and 3) derive the PCR-augmented representation of each pixel by incorporating the D disparity prototypes.

Computing the matching probability maps. Matching probability maps are denoted by $\{M_0, \ldots, M_{D-1}\}$, where a value on each 2D map M_d indicates how likely a pixel p_i belongs to the disparity plane π_d, and we generate such a group of disparity planes under the supervision of quantified disparity ground truth. We partition the 3D cost column conveyed from hourglass module into D slice $\{c_0, \ldots, c_{D-1}\}$, with each cost slice corresponding to a disparity plane. For pixel p_i, the matching probability to plane π_d is the softmax output of c_d^i over the disparity dimension:

$$m_d^i = \frac{\exp(c_d^i)}{\sum_{d'=0}^{D-1} \exp(c_{d'}^i)}. \tag{1}$$

Estimating the disparity prototypes. It is computationally expensive to get contextual representations from original 4D feature volume, and it contains redundant information, because only a small portion of disparity candidates $\{0, \ldots, D-1\}$ have major contributions to final prediction (with most disparity candidates having a very low matching probability). Therefore, we conduct a simple 4D-to-3D conversion Γ:

$$\Gamma : \eta(\mathrm{cat}(V(d, C, x, y))), \tag{2}$$

where $V(d, C, x, y)$ is the 3D part for disparity d in the 4D feature volume. The Γ operation first concatenates C features over disparity dimension (using $\mathrm{cat}(\cdot)$) for dimension reduction, and then uses the 1×1 convolution $\eta(\cdot)$ for feature adaption to obtain a resultant 3D feature volume $F \in \mathbb{R}^{\hat{C} \times H \times W}$.

Then we can aggregate all pixels weighted by the matching probability to estimate the disparity prototype $\mathbf{x}_d$:

$$\mathbf{x}_d = \sum_{i \in I} \widetilde{m}_d^i \mathbf{f}_i, \tag{3}$$

where $\mathbf{f}_i$ is the feature of pixel p_i in F and $\widetilde{m}_d^i$ is the normalized degree for matching probability of each pixel p_i across spatial dimension; $\mathbf{x}_d$ has the same size of $\mathbf{f}_i \in \mathbb{R}^{\hat{C} \times 1 \times 1}$. In this way, $\mathbf{x}_d$ can reflect the intrinsic information in a disparity plane, and thus we call it the disparity prototype of disparity plane π_d.

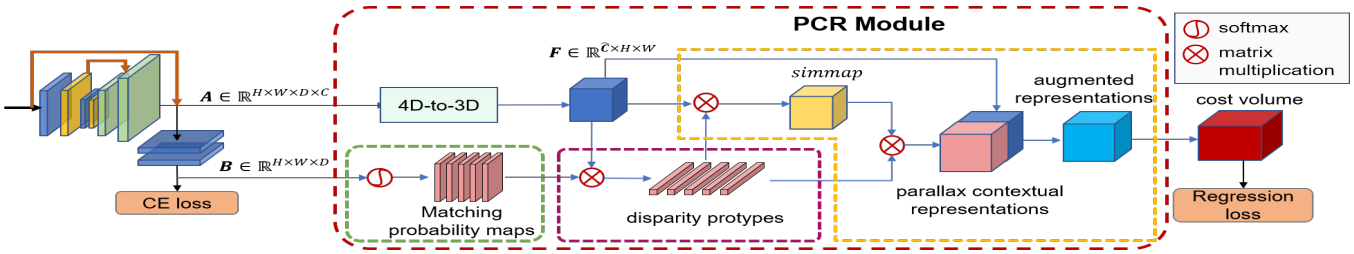

Fig. 3. Illustration of the whole pipeline of PCR. The far left part depicts the two output branches fed into PCR from the last hourglass module. The architecture of proposed PCR: (i) green dashed box: Computing matching probability maps. (ii) purple dashed box: Estimating disparity prototypes. (iii) orange dashed box: Deriving PCR-augmented representation. Finally, we transform the augmented representation to a new refined cost volume.

Deriving the PCR-augmented representations. Different pixels should selectively attend to different disparity planes. The relation between each pixel and each disparity prototype acts as the "parallax contextual attention". Inspired by the concept of self-attention [18], we calculate the parallax contextual representation $\mathbf{y}_i$ for pixel p_i as

$$\begin{aligned} \mathbf{y}_i &= \rho\Big(\sum_{d=0}^{D-1} \omega_d^i \delta(\mathbf{x}_d)\Big), \\ \omega_d^i &= \frac{e^{\gamma(\mathbf{f}_i, \mathbf{x}_d)}}{\sum_{d'=0}^{D-1} e^{\gamma(\mathbf{f}_i, \mathbf{x}_{d'})}}, \end{aligned} \tag{4}$$

where $\gamma(\mathbf{f}, \mathbf{x}) = \phi(\mathbf{f})^T \psi(\mathbf{x})$ is the unnormalized relation function; and $\phi(\cdot)$, $\psi(\cdot)$, $\delta(\cdot)$ and $\rho(\cdot)$ are all transformation function implemented by 1×1 convolution followed by BN and ReLU.

Finally, we concatenate $\mathbf{y}_i$ and original feature $\mathbf{f}_i$ together to obtain the PCR-augmented representation of pixel p_i:

$$\mathbf{z}_i = g([\mathbf{f}_i^T \mathbf{y}_i^T]^T), \tag{5}$$

where $g(\cdot)$ is also 1×1 convolution to fuse the original representation and the PCR. The whole pipeline of our PCR module is depicted in Fig.3.

In this way, we can adaptively aggregate relevant context from different disparity planes for each pixel. For a given pixel p_i, its original feature $\mathbf{f}_i$ may have ambiguity in the locality of disparity plane, but with the help of disparity prototypes this ambiguity could be reduced and the pixel will be divided into the correct disparity plane.

2.2. Loss Functions

Replacing regression loss with cross-entropy loss. As shown in Fig.3, unlike most existing works using regression loss only, we propose to replace original regression loss with a cross-entropy loss for every hourglass network and add only one regression loss at the end of the PCR module. Our design is for two purposes. First, the cross-entropy loss is favorable to get a clean separation of disparity planes; Second, the cross-entropy loss is helpful to alleviate the well-know multi-modal problem brought by the weighted average operation [19]. The cross-entropy loss L_{ce} is defined as

$$L_{ce}(\hat{y}_{gt}, p) = -\sum_{d=0}^{D-1} \hat{y}_{gt} \cdot \log p(d), \tag{6}$$

where $p(\cdot)$ is the estimated probability distribution of possible disparity candidates; and $\hat{y}_{gt}$ is the one-hot ground-truth label after quantification. The quantified ground truth $\hat{d}_{gt}$ is obtained by minimizing $|d_{gt} - \hat{d}|$, where d_{gt} is continuous disparity ground truth and $\hat{d}$ is the integral value of candidate disparity indexes.

Total loss function. We use a combination of the cross-entropy loss L_{ce} and the regression loss L_{reg} to supervise the training of our network:

$$L = L_{ce} + \lambda L_{reg}, \tag{7}$$

where $\lambda(= 0.2)$ in our work is used to balance the two losses; the smooth L1 loss is adopted as regression loss; and the cross-entropy loss L_{ce} contains three parts, each of which acts on a 3D cost volume on each hourglass output and keeps the same weighting setting as PSMNet.

3. EXPERIMENTS

3.1. Datasets and Implementation Details

Datasets. We evaluate our methods on the SceneFlow [1] and KITTI [20, 21] datasets. The SceneFlow dataset is a large scale synthetic dataset with dense ground-truth disparity maps. The KITTI2012 and KITTI2015 datasets are real-world datasets providing sparse ground-truth disparity.

Networks and training. We validate our proposal embedded in two 3D CNN based disparity networks: PSMNet and GwcNet. We use the Adam optimizer, with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. All the data processing and training strategies are the same as those in the original papers [5, 10]. We train our model with a batch size of 8 on 4 NVIDIA 2080Ti GPUs.

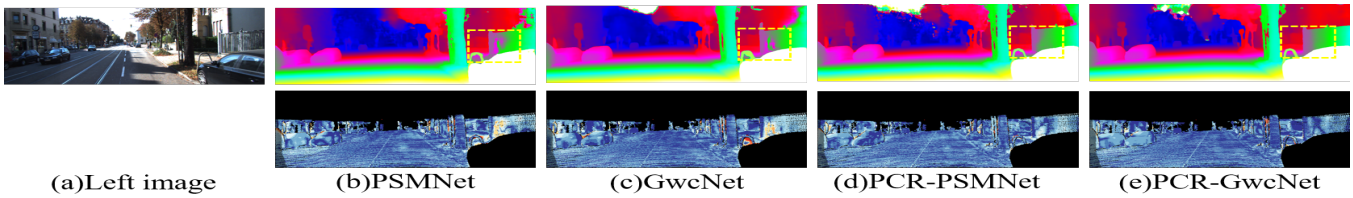

Fig. 4. Visualization results on KITTI2015. The left-hand panel shows the left input image of the stereo image pair. The right-hand panels show (upper) the disparity maps and (lower) the error maps obtained by different methods.

3.2. Ablation Studies

We compared three models: the baseline network(PSMNet), the baseline with the cross-entropy loss replacing the regression loss (-ce), and our approach (PCR-PSMNet). From Table 1, we can see that the cross-entropy loss helps the network achieve better performance. With the help of cross-entropy loss, the quality of disparity prototypes is guaranteed thus the PCR-augmented method achieves a significant performance boost with EPE drops from 1.09 to 0.94.

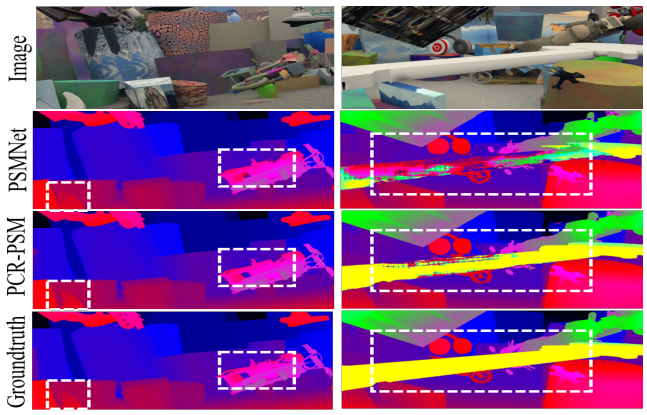

Fig. 5. Visualization results on SceneFlow.

Table 1. Performance comparison under different settings and metrics on the SceneFlow test set.

Method	SceneFlow			
	1px(%)	2px(%)	3px(%)	EPE
PSMNet	10.80	6.06	4.41	1.09
PSMNet-ce	9.54	5.57	4.24	1.00
PCR-PSMNet	**9.00**	**5.01**	**3.74**	**0.94**

Fig.5 gives some visual comparisons on the SceneFlow test set. We can find that our proposed approach produces sharper edges and better recovery of thin structures (left-hand column) and corrects some regions partitioned to wrong disparity planes(right-hand column). Such improvements could be attributed to the discriminative division of disparity planes and the long-range dependency with the help of PCR.

3.3. Benchmark Results

We embed our proposed PCR into PSMNet and GwcNet to build two new models: PCR-PSMNet and PCR-GwcNet. We compare them with PSMNet, GwcNet and some other methods on the test set for the KITTI submission, as shown in Tables 2 and 3. According to the online leader board, the PCR-enhanced models achieve better performance in all the evaluation metrics compared with the original ones.

Table 2. Evaluation Results on KITTI2015.

Method	All(%)			Noc(%)		
	bg	fg	all	bg	fg	all
GCNet [4]	2.21	6.16	2.87	2.02	5.58	2.61
PSMNet [5]	1.86	4.62	2.32	1.71	4.31	2.14
SegStereo [22]	1.88	4.07	2.25	1.76	3.70	2.08
GwcNet [10]	1.74	3.93	2.11	1.61	3.49	1.92
PCR-PSMNet	1.53	3.62	1.88	1.39	3.32	1.71
PCR-GwcNet	**1.49**	**3.51**	**1.83**	**1.36**	**3.17**	**1.66**

Table 3. Evaluation Results on KITTI2012.

Method	>2px(%)		>3px(%)		>5px(%)	
	Noc	All	Noc	All	Noc	All
GCNet [4]	2.71	3.46	1.77	2.30	1.12	1.46
PSMNet [5]	2.44	3.01	1.49	1.89	0.90	1.15
SegStereo [22]	2.66	3.19	1.68	2.03	1.25	1.52
GwcNet [10]	2.16	2.71	1.32	1.70	0.80	1.03
PCR-PSMNet	2.11	2.63	1.30	1.65	0.79	1.01
PCR-GwcNet	**1.97**	**2.51**	**1.23**	**1.60**	**0.75**	**0.98**

Some resultant maps downloaded from the KITTI evaluation server are visualized in Fig.4, showing that our PCR-enhanced models perform remarkably better in some ill-posed regions and keep the object structures very well.

4. CONCLUSION

In this work, we propose the concept of disparity prototype to exploit the disparity-level context. With the parallax contextual representation (PCR), the augmented pixel representation enables a more accurate prediction of disparity maps. We further propose to integrate advantages of cross-entropy loss and the widely-used soft-argmin operation. Experiments on various public datasets and visualization results verify the effectiveness of our proposal.

5. REFERENCES

[1] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *CVPR*, 2016, pp. 4040–4048. 1, 2, 3

[2] Jiahao Pang, Wenxiu Sun, Jimmy SJ Ren, Chengxi Yang, and Qiong Yan, "Cascade residual learning: A two-stage convolutional neural network for stereo matching," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 887–895. 1

[3] Haofei Xu and Juyong Zhang, "AANet: Adaptive aggregation network for efficient stereo matching," in *CVPR*, 2020, pp. 1959–1968. 1

[4] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry, "End-to-end learning of geometry and context for deep stereo regression," in *ICCV*, 2017, pp. 66–75. 1, 2, 4

[5] Jia-Ren Chang and Yong-Sheng Chen, "Pyramid stereo matching network," in *CVPR*, 2018, pp. 5410–5418. 1, 2, 3, 4

[6] Sameh Khamis, Sean Fanello, Christoph Rhemann, Adarsh Kowdle, Julien Valentin, and Shahram Izadi, "StereoNet: Guided hierarchical refinement for real-time edge-aware depth prediction," in *ECCV*, 2018, pp. 573–590. 1

[7] Yan Wang, Zihang Lai, Gao Huang, Brian H Wang, Laurens Van Der Maaten, Mark Campbell, and Kilian Q Weinberger, "Anytime stereo image depth estimation on mobile devices," in *ICRA*, 2019, pp. 5893–5900. 1

[8] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip H.S. Torr, "GA-Net: Guided aggregation net for end-to-end stereo matching," in *CVPR*, Long Beach, CA, USA, 2019, pp. 185–194, IEEE. 1

[9] Guang-Yu Nie, Ming-Ming Cheng, Yun Liu, Zhengfa Liang, Deng-Ping Fan, Yue Liu, and Yongtian Wang, "Multi-level context ultra-aggregation for stereo matching," in *CVPR*, 2019, pp. 3283–3291. 1

[10] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li, "Group-wise correlation stereo network," in *CVPR*, 2019, pp. 3273–3282. 1, 2, 3, 4

[11] Xianzhi Du, Mostafa El-Khamy, and Jungwon Lee, "Amnet: Deep atrous multiscale stereo disparity estimation networks," *arXiv preprint arXiv:1904.09099*, 2019. 1

[12] Daniel Scharstein and Richard Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision*, vol. 47, no. 1-3, pp. 7–42, 2002. 1

[13] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He, "Non-local neural networks," in *CVPR*, 2018, pp. 7794–7803. 1

[14] Meng Li, William Hsu, Xiaodong Xie, Jason Cong, and Wen Gao, "SACNN: Self-attention convolutional neural network for low-dose CT denoising with self-supervised perceptual loss network," *IEEE Transactions on Medical Imaging*, 2020. 1

[15] Fan Zhang, Yanqin Chen, Zhihang Li, Zhibin Hong, Jingtuo Liu, Feifei Ma, Junyu Han, and Errui Ding, "ACFNet: Attentional class feature network for semantic segmentation," in *ICCV*, 2019, pp. 6798–6807. 1

[16] Yuhui Yuan, Xilin Chen, and Jingdong Wang, "Object-contextual representations for semantic segmentation," *arXiv:1909.11065*, 2019. 1

[17] Ruigang Niu, "HMANet: Hybrid multiple attention network for semantic segmentation in aerial images," *arXiv:2001.02870*, 2020. 1

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *NeurIPS*, 2017, pp. 5998–6008. 3

[19] Chuangrong Chen, Xiaozhi Chen, and Hui Cheng, "On the over-smoothing problem of cnn based disparity estimation," in *ICCV*, 2019, pp. 8997–9005. 3

[20] Andreas Geiger, Philip Lenz, and Raquel Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *CVPR*, 2012, pp. 3354–3361. 3

[21] Moritz Menze and Andreas Geiger, "Object scene flow for autonomous vehicles," in *CVPR*, 2015, pp. 3061–3070. 3

[22] Guorun Yang, Hengshuang Zhao, Jianping Shi, Zhidong Deng, and Jiaya Jia, "Segstereo: Exploiting semantic information for disparity estimation," in *ECCV*, 2018, pp. 636–651. 4