Depth Estimation from a Single Shot Image Using Feature Pyramid Network

Yudai Fukuda*, Takuro Oki* and Ryusuke Miyamoto†
*Dept. of Computer Science, Graduate School of Science and Technology, Meiji University, Japan
†Dept. of Computer Science, School of Science and Technology, Meiji University, Japan
Email:{fukuda, o_tkr, miya}@cs.meiji.ac.jp

***Abstract*—Depth estimation is a popular research topic in the field of computer vision. Recent schemes based on deep learning are showing good results for this task, although hand-crafted features and Markov random field were popular several years ago. This paper introduces a feature pyramid network extracting global features from input images into depth estimation, which was originally proposed for object detection. To show the validity of the feature pyramid network, a neural network for depth estimation from a single shot image composed of ResNet-50 and the feature pyramid network was implemented. Experimental results using the KITTI dataset showed that RMSE was improved by about 5% by the proposed scheme with an acceptable decrease of computational speed, resulting in a processing speed of about ten frames per second on a NVIDIA GV100 GPU with 32GB memory.**

I. Introduction

Several schemes have attempted to solve depth estimation from a single shot image using hand-crafted features prior to the emergence of deep learning. In [1], Saxena et al. proposed a scheme that uses feature vectors composed of seventeen types of features: two color features and fifteen types of local features obtained by spatial filters[2]. Make3D[3] adopted the Markov random field model to combine superpixels obtained by segmentation, considering that depth values of neighbouring pixels tend to be similar.

After the emergence of deep learning, Eigen et al. showed good accuracy for depth estimation from a single image[4] using two different convolutional neural networks to treat global and local features for depth estimation appropriately. In the scheme, these two networks were trained independently. One network was trained to produce a rough depth map using global features, and the other network enhanced edges of the depth map using local features. To achieve a further improvement in estimation accuracy, a new scheme was proposed that adopted three independent networks for depth estimation[5]. The scheme proposed by Kuznietsov et al. [6] adopted skip connections that transfer feature maps generated at each convolutional layer in an encoder directly to a decoder, in order to improve the estimation accuracy without an additional computational cost.

Semantic segmentation, whose accuracy was drastically improved by deep learning, has also been adopted for depth estimation from a single image. DORN[7] is one of the schemes that introduced the idea of a pyramid pooling module, which showed excellent results for semantic segmentation used in PSPNet[8] and DeepLab[9]. DORN tried to obtain global features by atrous spatial pyramid pooling using dilated convolution with various pixel intervals. Experimental results with DORN show that pyramid pooling is also effective for the depth estimation task.

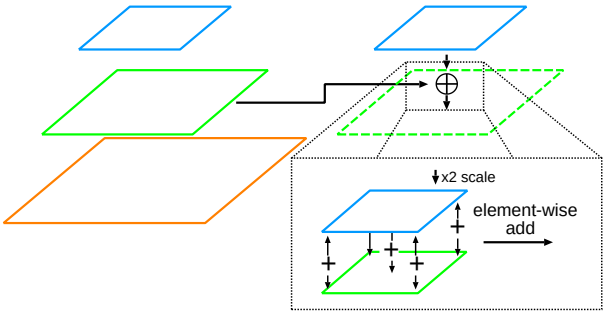

Fig. 1. Upsampling operation in a feature pyramid network.

Pyramid pooling is a valid method for depth estimation. However, this operation requires huge computational costs to apply to existing schemes. Hu et al. proposed a novel scheme that attempted to use a feature pyramid without additional computational costs[10]. This scheme reused feature maps with different scales generated in the middle of the convolution process. This operation was termed multi-scale feature fusion, and enabled accurate depth estimation considering global features obtained from several scales of feature maps.

However, this reuse of feature maps may cause degradation of estimation accuracy because shallow layers cannot obtain good feature maps with effective information extracted by the early stages of convolution layers. To solve this problem, this paper proposes a novel network for depth estimation that uses a Feature Pyramid Network[11] (FPN) to extract global features. Fig. 1 shows an upsampling process by FPN on several scales of feature maps. To validate the effect of the proposed scheme, estimation accuracy is evaluated using the KITTI dataset and compared with state-of-the-art schemes.

II. Feature Pyramid Network

The Feature Pyramid Network (FPN) was originally proposed for improved accuracy of object detection. The object detection task requires several scales of input images or feature spaces to find various sizes of detection targets. Generally, an increase in the number of scales used in the detection

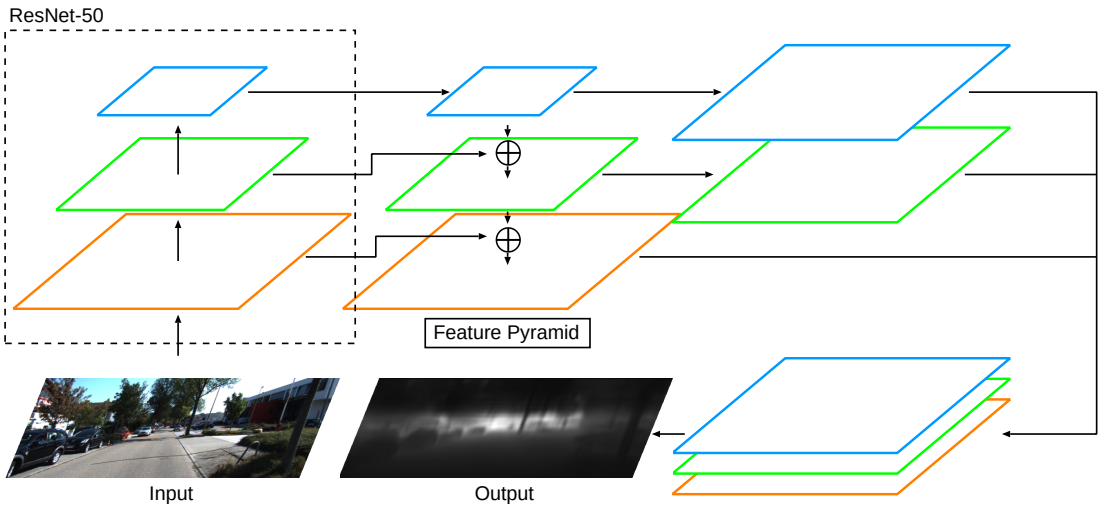

Fig. 2. Overview of a depth estimation network with a feature pyramid network.

improves the detection accuracy. However, the computational costs required for detection increases according to the number of scales and the processing speed becomes slower. The trade-off between accuracy and computational costs therefore becomes quite important to achieve practical object detection. In object detection schemes without deep learning, FPDW[12] and the 100fps human detection scheme[13] are famous for their efforts to construct efficient detectors considering multi-scale detection. The FPN also tries to improve the efficiency of object detection using deep learning.

The FPN uses a feature pyramid obtained from several scales of feature maps, however it does not adopt simple concatenation of generated feature maps at multiple layers. Smaller feature maps are upsampled using the nearest neighbor interpolation to enlarge the size of features to be the same size as the largest feature map. After upsampling, these feature maps are concatenated to generate a feature pyramid. Fig. 1 shows the operation flow of generating a feature pyramid using the FPN. This feature pyramid enables efficient computation for multi-scale object detection.

III. How to Apply Feature Pyramid Network to Depth Estimation

Fig. 2 shows an overview of the proposed network architecture that introduces the feature pyramid network for end-to-end depth estimation. In the proposed architecture, ResNet-50 was adopted for feature extraction from an input image. A feature pyramid is constructed from feature maps generated corresponding to three layers having different scales in order to obtain global features using the FPN. As these layers have different scales, the third, fourth, and fifth convolutional layers of ResNet-50 are adopted: they are named C_3, C_4, and C_5. Layers in a constructed feature pyramid corresponding to C_3, C_4, and C_5 are named P_3, P_4, and P_5. The sizes of layers having the same number is the same and the size of a layer having one greater number is twice as large as the size of the layer having one smaller number.

Fig. 3 shows how to construct a feature pyramid from a feature map generated from ResNet-50. At Fig. 3(a), C_5 obtained from convolution by ResNet-50 are set as P_5 which is the top layer of a feature pyramid. Next, P_4 is generated from C_4 obtained from ResNet-50 and upsampled P_5 as shown in Fig. 3(b). In the same way, P_3 is generated from C_3 obtained from ResNet-50 and upsampled P_4 as shown in Fig. 3(c). These additional operations for feature map generation do not require a large computational cost; only simple element-wise addition is required in addition to the original computation flow. Finally, a depth value is estimated by concatenating P_3, upsampled P_4, and upsampled P_5 as shown in Fig. 3(d): the scaling factors for P_4 and P_5 are two and four, respectively.

To train the proposed network, the following loss function is adopted, which represents the square of the error between the correct depth and the estimated depth.

$$Loss = \frac{1}{N} \cdot \frac{1}{w \cdot h} \sum_{n=1}^{N} \sum_{i=1}^{w} \sum_{j=1}^{h} (y_{n,i,j} - y^*_{n,i,j})^2, \qquad (1)$$

where y, $y*$, N, w, and h represent an estimated depth value, a correct depth value, the number of images, width of images, and height of images, respectively.

IV. Evaluation

This section evaluates the proposed scheme using the KITTI dataset.

A. Dataset

The KITTI dataset was used for evaluation of the proposed scheme. This dataset has RGB images and corresponding depth images obtained by an on-board camera and a depth sensor, respectively. The KITTI dataset includes outdoor scenes at various locations such as city, residential, and so on. 22,600 images were used for training and 697 for evaluation. The resolution of the input images was converted to 621×188, which was about half of the original size.

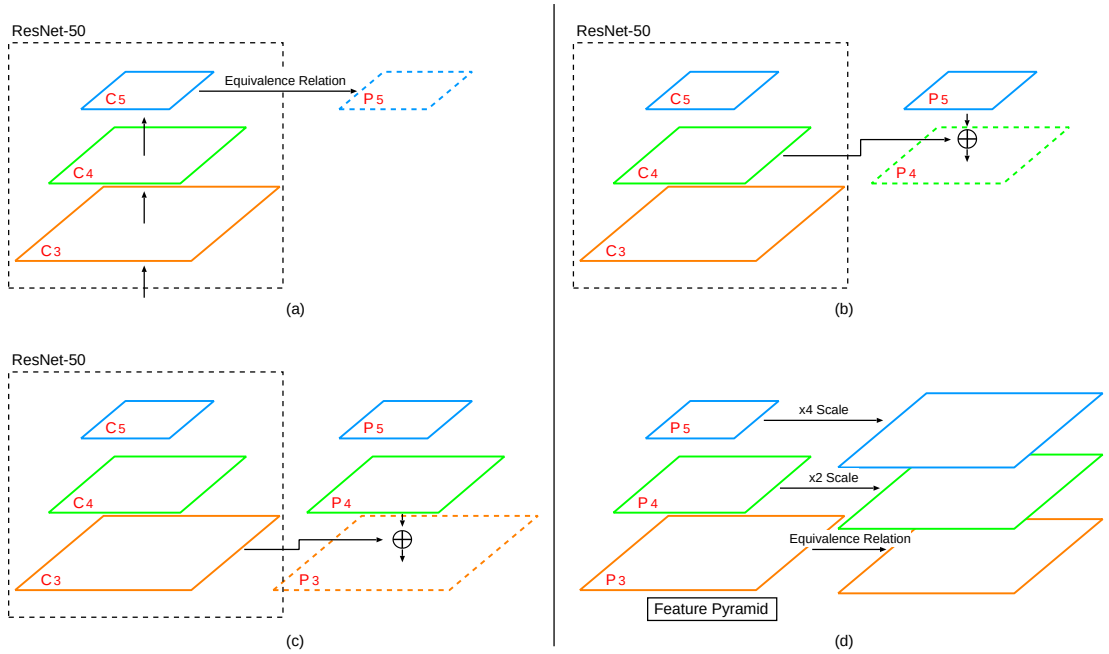

Fig. 3. Depth estimation with feature pyramid network.

B. Experimental procedures

To evaluate the validity of the proposed scheme, accuracies by two different networks were compared: the proposed scheme that tries to obtain global features using a feature pyramid created by several scales of feature maps generated from ResNet-50 and a network without a feature pyramid network. ResNet-50 maps used for both networks were fine-tuned for depth estimation from a pre-trained model constructed with ImageNet[14]. As an optimizer, Momentum SGD[15] was adopted.

The proposed network was implemented using Caffe[16], which is a widely used deep learning framework. The size of the mini-batch was eight and the initial value of learning rate was 1.0×10^{-8}. The learning rate was changed to 1.0×10^{-9} when the number of iterations reached 30,000. Momentum and weight decay were 0.9 and 0.005, respectively. For training and inference of the evaluation, NVIDIA QUADRO GV100 with 32GB graphics memory was used. Table I shows the specification of the PC containing the GPU used for the evaluation.

TABLE I
EXPERIMENTAL ENVIRONMENT.

CPU	Xeon(R) E5-2697 v4 ×2
Main memory	128GB
GPU	NVIDIA Quadro GV100
GPU memory	32GB
OS	Ubuntu 16.04.5 LTS

C. Evaluation criteria

To evaluate the accuracy of the proposed network and the reference network, the following criteria were adopted:

$$\text{RMSE} : \sqrt{\frac{1}{|T|}\sum_{y\in T}||y-y^*||^2}, \quad (2)$$

$$\text{RMSE(log)} : \sqrt{\frac{1}{|T|}\sum_{y\in T}||\log y-\log y^*||^2}, \quad (3)$$

$$\text{Abs Relative difference} : \frac{1}{|T|}\sum_{y\in T}|y-y^*|/y^*, \quad (4)$$

$$\text{Squared Relative difference} : \frac{1}{|T|}\sum_{y\in T}||y-y^*||^2/y^*, \text{and} \quad (5)$$

$$\text{Accuracy} : \%\ \text{of}\ y_i\ \text{s.t.}\ \max(\frac{y_i}{y_i^*},\frac{y_i^*}{y_i}) = \delta < thr. \quad (6)$$

In these equations, T, y, and y^* represent a set of pixels corresponding to a depth image, an estimated value of depth, and the correct depth obtained from ground truth, respectively.

RMSE (root mean squared error) is a widely used criterion. RMSE(log) is Root Mean Squared Logarithmic, which can reduce the influence of large outliers. RMSE and RMSE(log) can evaluate the absolute error but they do not show the error rate relative to the estimated value. Therefore, absolute relative difference and squared relative difference were also evaluated. These are relative errors considering the error ratio to the estimated value: absolute relative difference and squared relative difference represent the absolute error and the squared error, respectively.

Accuracy relates to the percentage of correct answers. An estimated value is regarded as correct if the result obtained by division of a larger value by a smaller value between the estimated value and ground truth was less than a threshold

TABLE II
ACCURACY EVALUATION USING THE KITTI DATASET.

	Error (Lower is better)				Accuracy (Higher is better)		
Model	rmse	log_rmse	abs_rel	sq_rel	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
ResNet50	4.670	0.197	0.131	0.854	0.828	0.951	0.984
ResNet50 + FPN	4.443	0.190	0.125	0.793	0.838	0.955	0.986
Eigen et al.[4]	7.156	0.270	0.190	1.515	0.692	0.899	0.967
Liu et al.[17]	7.046	–	0.217	–	0.656	0.881	0.958
Xu et al.[18]	4.384	0.188	0.127	0.811	0.841	0.955	0.985
DORN[7]	4.006	0.181	0.115	0.716	0.872	0.956	0.981

TABLE III
COMPUTATION TIME FOR THE KITTI DATASET.

Model	time(s)
Xu et al.[18]	0.101
DORN [7]	1.522
ResNet50	0.070
ResNet50 + FPN	0.097

value. In the evaluation, 1.25, 1.25^2, and 1.25^3 were used as threshold values.

D. Experimental results

Table. II summarizes the experimental results using criteria described above by the proposed network, the reference network, and several other existing schemes. These results shows that the feature pyramid network can improve the estimation accuracy compared with the reference network that does not have a feature pyramid. The proposed scheme also outperforms [4] and [17], and the accuracy is comparable to [18]. However, the proposed scheme does not show better accuracy than [7]: the proposed scheme outperforms [7] in accuracy when the threshold was 1.25^3.

Fig. 4 shows some examples of estimation results by the proposed scheme and the reference. The upper part of the depth images in Fig. 4 is dark because the ground truth is obtained by a laser sensor whose viewing angle is narrower than that of the visible camera used to create the dataset. It can be seen that the feature pyramid network improves the estimation accuracy compared with the results given by the reference scheme. In particular, contours of objects such as cars and trees can be estimated clearly.

Table III shows the computation speed for the KITTI dataset. The results show that the computation speed of the proposed scheme was slower than the reference scheme owing to the additional computation by the feature pyramid network. However, the computation speed itself is not unreasonably slow, as more than ten frames can be computed per second.

V. CONCLUSION

This paper introduces a feature pyramid network that was originally proposed for accuracy improvement of object detection in order to improve the accuracy of depth estimation from a single shot image using appropriate global features. To validate the effect of the feature pyramid network, a neural network for depth estimation composed of ResNet-50 and the feature pyramid network was implemented using Caffe and evaluated using the KITTI dataset. Experimental results showed that the feature pyramid network improves RMSE by about 5% relative to the reference network that did not have the feature pyramid network and the accuracy also improved.

The estimation performance was improved by the feature pyramid network. However, the accuracy obtained did not outperform the state-of-the-art scheme proposed by Fu et al.[7]. The proposed scheme showed a better result for accuracy only when the threshold was 1.25^3. To achieve further improvement of depth estimation, further efforts should be applied to the current network. In the future, the authors will try to apply Multi-Level Feature Pyramid Network[19] which showed good results for object detection and Visual Attention[20] that has been proposed for fashion analysis.

REFERENCES

[1] A. Saxena, S. H. Chung, and A. Y. Ng, "Learning depth from single monocular images," in *Proc. Advances in Neural Information Processing Systems*, 2005, pp. 1161–1168.

[2] Laws and Kenneth I, "Textured image segmentation," Tech. Rep., University of Southern California Los Angeles Image Processing INST, 1980.

[3] A. Saxena, M. Sun, and A. Y. Ng, "Make3D: Learning 3D scene structure from a single still image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 5, pp. 824–840, 2009.

[4] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Proc. Advances in Neural Information Processing Systems*, 2014, pp. 2366–2374.

[5] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2650–2658.

[6] Y. Kuznietsov, J. Stückler, and B. Leibe, "Semi-supervised deep learning for monocular depth map prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2215–2223.

[7] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2002–2011.

[8] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6230–6239.

[9] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. European Conference on Computer Vision*, 2018, pp. 833–851.

[10] J. Hu, M. Ozay, Y. Zhang, and T. Okatani, "Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries," in *Proc. IEEE Winter Conf. on Applications of Computer Vision*, 2019, pp. 1043–1051.

[11] T. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 936–944.

[12] P. Dollár, S. Belongie, and P. Perona, "The fastest pedestrian detector in the west," in *Proc. Brit. Mach. Vis. Conf.*, 2010.

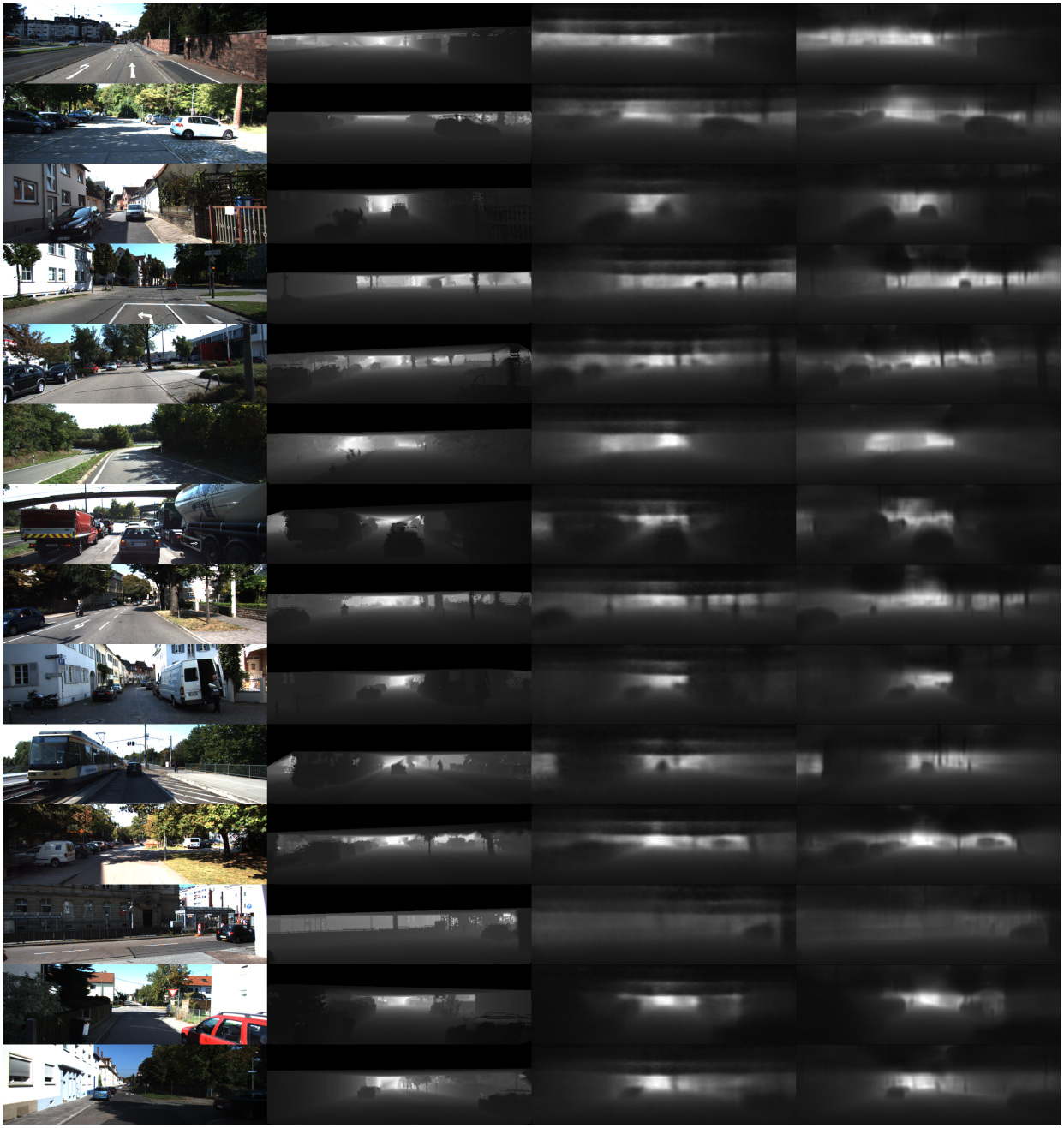

Fig. 4. Examples of depth estimation. RGB image, ground truth, the reference network, and the proposed network.

[13] R. Benenson, M. Mathias, R. Timofte, and L. Van Gool, "Pedestrian detection at 100 frames per second," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, June 2012, pp. 2903–2910.
[14] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.
[15] D. E. Rumelhart, G. E. Hinton, and R. J. Wilson, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, 1986.
[16] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. B. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *ACM International Conference on Multimedia*, 2014, pp. 675–678.
[17] F. Liu, C. Shen, G. Lin, and I. D. Reid, "Learning depth from single monocular images using deep convolutional neural fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2024–2039, 2016.
[18] D. Xu, W. Wang, H. Tang, H. Liu, N. Sebe, and E. Ricci, "Structured attention guided convolutional neural fields for monocular depth estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3917–3925.
[19] Q. Zhao, T. Sheng, Y. Wang, Z. Tang, Y. Chen, L. Cai, and H. Ling, "M2Det: A single-shot object detector based on multi-level feature pyramid network," *CoRR*, vol. abs/1811.04533, 2018.
[20] W. Zhonghao, G. Yujun, Z.Y. Ru, Z. Jun, and G. Xiao, "Clothing retrieval with visual attention model," in *Proc. IEEE International Conference on Visual Communications and Image Processing*, 2017, pp. 1–4.