Depth maps correction based on neighboring frames

A. Varlamova
Lomonosov Moscow State University
Moscow, Russia
avarlamova1996@gmail.com

V. Kitov
Lomonosov Moscow State University
Moscow, Russia
v.v.kitov@yandex.ru

***Abstract*—**The use of convolutional neural networks in image processing tasks often allows achieving significantly better results in comparison with traditional methods. For the problem of calculating a real-time depth map, the state-of-the-art method is the MiDaS method - a convolutional neural network trained on a large dataset that is not publicly available. However, this approach does not include the ability to use information from neighboring frames which can improve the prediction. The method proposed in this paper uses the depth maps generated by MiDaS for several frames of a video sequence and their further refinement, which makes it possible to achieve an improvement in quality without a significant decrease in the algorithm performance.

Keywords— depth maps, convolutional neural networks, supervised learning, structure-from-motion.

1. INTRODUCTION

Depth information obtained from one input image or a set of them is an important part of solving a wide range of problems. For example, an algorithm that allows you to get depth maps in real-time would make it possible to replace LIDAR systems in self-driving vehicles, which could lead to a reduction in their cost. Depth maps can be useful in segmentation and object detection problems [1, 2]. They even can be applied to the style transfer problem, significantly improving the quality of the generated image [3, 4].

The best results are achieved by the neural networks-based approach. The current state-of-the-art model – MiDaS [5] is capable of generating accurate depth maps. However, in the case of a video sequence, one can extract additional spatial information from neighboring frames, which can be used for further refinement.

Papers focusing on depth prediction using video sequence [6, 7] typically use it to organize learning in an unsupervised manner with the neural network taking one frame at a time as its input. The paper [8] considers an approach based on two frames, however, the method loses [5] both in terms of performance and quality.

This article focuses on the possibility of using information from consequent frames of a video sequence simultaneously.

Various input data configurations are explored and a link between the quality of prediction and frames distance from each other in a sequence is studied. This paper also proposes a modified loss function capable of improving the prediction. The resulting method gives an increase in quality while maintaining comparable performance.

2. RELATED WORK

Depth maps calculation is traditionally associated with structure-from-motion algorithms, however, there are some serious disadvantages such as significant processing time, incomplete depth maps, and inconsistent results in the case of moving objects. Another group of methods is based on training on video data in an unsupervised manner, however, they use several frames only to calculate the loss function, leaving the prediction stage with one image [6, 7].

3. METHOD

The main goal of this paper is to improve the inverse depth maps obtained using the state-of-the-art MiDaS method for the case of a video sequence leveraging information about neighboring frames.

Simple convolution was trained for several data representation options. To ensure that the values are non-negative, the RELU activation function was used. Two versions of the loss function were used. The first one was identical to MiDaS's:

$$\mathcal{L} = \frac{1}{N}\sum_{n=1}^{N}\mathcal{L}_{ssi}(c^n,(c^t)^n) + \alpha\mathcal{L}_{reg}(c^n,(c^t)^n), \quad (1)$$

where

$$\mathcal{L}_{reg}(c,c^t) = \frac{1}{M}\sum_{k=1}^{K}\sum_{i=1}^{M}(|\nabla_x R_i^k| + |\nabla_y R_i^k|),$$

and

$$\mathcal{L}_{ssi}(c,c^t) = \frac{1}{2M}\sum_{i=1}^{M}(c-c^t)^2,$$

where N is the size of the training set, M – the size of the image, $R_i^k = c_i\text{-}c_i^t$ at k scale. c_i,c_i^t are predicted and ground true inverse depth maps correspondingly which were transformed to have equal scales and translations.

However, during the experiments, it was noticed that the structure of the loss function is such that a larger penalty is given to closer objects. To balance the difference in values, an additional version of the loss function is proposed:

$$\mathcal{L}_{proximity}(c,c^t) = \frac{1}{2M}\sum_{i=1}^{M}(\frac{c_i - c_i^t}{c_i^t})^2 \quad (2)$$

Thus objects farther away (with smaller values) are given a larger penalty for deviating from real values.

The median transformation was used as an approach to align predicted and original scale and translation parameters with each other. For each true inverse depth map and each prediction shift values were calculated as the median value and scale values as the mean absolute differences between the inverse depth map values and the computed shift values.

Three types of input data were considered:

- MiDaS's inverse depth maps for C_i, C_{i+k}, C_{i-k} (ND_k, neighbor depth).
- Inverse depth map for C_i, averaged inverse depth maps for k steps forward and backward (AD_k, average depth).

- Inverse depth map for C_i, differences between depth map for $C_i$and C_{i-k}, C_{i+k} (DD_k, depth difference).

A. Training details

The training was performed on 25000 images from a synthetic dataset [9] during 1 epoch. α parameter was set to 0.3, batch size was equal to 18 for ND and DD input types and 9 for AD.

Б. Evaluation

Quality was estimated on the validation set (5500 images) for different values of k parameter: 1, 2, 4, and 8 for ND and DD, and 2, 3, 4 for AD.

Abs Rel, Sq Rel, RMSE, log RMSE, δ metrics were used for evaluation. The best result was achieved by the method with ND input type with k=4.

For further experiments, a simple U-Net-like model with skip-connection was trained on the same dataset for data input type. The architecture of the model can be seen in Fig. 1.

For the comparison, Sintel [10] and ETH3D [11] datasets were used. Results are presented in Table 1. Better performance using all metrics for the Sintel dataset and almost all (except for δ) for the ETH3D dataset was achieved.

4. CONCLUSION

The best-performing methods for real-time depth map calculation are based on using neural networks. The State-of-the-art MiDaS method can be improved upon by using information from neighboring frames in tasks where video sequence is involved.

TABLE I. METHOD COMPARISON

Method	**Metrics**				
	Abs Rel	***Sq Rel***	***RMSE***	***log RMSE***	δ
ETH3D, ND_4	**0.108**	**0.282**	**1.386**	**0.148**	0.118
ETH3D, MiDaS	0.112	0.431	1.61	0.152	0.117
Sintel, ND_4	**0.276**	**2.546**	**5.111**	**0.438**	**0.38**
Sintel, MiDaS	0.295	3.254	5.401	0.468	0.386

In this paper, several options for presenting input data were proposed and it was shown how their form affects the final quality. The link between the quality of the prediction and the distance between adjacent frames during training is shown. Additionally, a modified loss function was proposed, which made it possible to make the model more consistent for distant objects. The algorithm obtained as a result of combining the approaches made it possible to improve the quality of prediction on datasets that did not participate in training and validation. At the same time, the possibility of using it in real-time was retained.

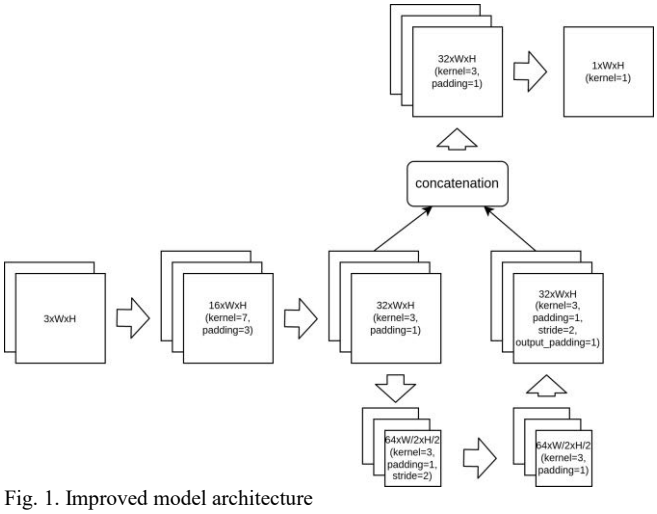

Fig. 1. Improved model architecture

REFERENCES

[1] Couprie, C. Indoor semantic segmentation using depth information / C. Couprie, C. Farabet, L. Najman, Y. LeCun // ArXiv preprint: 1301.3572, 2013.

[2] Qi, C.R. Frustum pointnets for 3d object detection from rgb-d data / C.R. Qi, W. Liu, C. Wu, H. Su, L.J. Guibas // Proceedings of the IEEE conference on computer vision and pattern recognition, 2018. – P. 918-927.

[3] Liu X.-C. Depth-aware neural style transfer / X.-C. Liu, M.-M. Cheng, Y.-K. Lai, P.L. Rosin // Proceedings of the Symposium on Non-Photorealistic Animation and Rendering. – 2017. – P. 1-10.

[4] Liao, R. Depth-preserving style transfer / R. Loao, Y. Xia, X. Zhang // Massachusetts Inst. Technol., Cambridge, MA, USA, Tech. Rep. – 2016.

[5] Lasinger, K. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer / K. Lasinger, R. Ranftl, K. Schindler, V. Koltun // ArXiv preprint: 1907.01341, 2019.

[6] Godard, C. Digging into self-supervised monocular depth estimation / C. Godard, O. Mac Aodha, M. Firman, G.J. Brostow // Proceedings of the IEEE international conference on computer vision. – 2019. – P. 3828-3838.

[7] Luo, X. Consistent video depth estimation / X. Luo, J.-B. Huang, R. Szeliski, K. Matzen, J. Kopf // ArXiv preprint: 2004.15021, 2020.

[8] Wang, C. Web Stereo Video Supervision for Depth Prediction from Dynamic Scenes / C. Wang, S. Lucey, F. Perazzi, O. Wang // CoRR. – 2019.

[9] Krahenbuhl, P. Free supervision from video games / P. Krahenbuhl // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. – 2018. – P. 2955-2964.

[10] Butler, D.J. A naturalistic open source movie for optical flow evaluation / D.J. Butler, J. Wulff, G.B. Stanley, M.J. Black // European Conf. on Computer Vision (ECCV). – 2012. – P. 611-625.

[11] Schops T. A Multi-View Stereo Benchmark with High-Resolution Images and Multi-Camera Videos / T. Schops, J.L. Schonberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, A. Geiger // Conference on Computer Vision and Pattern Recognition (CVPR). – 2017.