Paper Review On

Learning the Depths of Moving People by Watching Frozen People

Name: Dazhi Li

NetID: dl939

1. Introduction

   Learning the depths of moving people by watching frozen people is one of the awarded best papers in 2019 CVPR. High quality is the main reason for me to pick it and review on it. This paper is talking about a new method for people do dense depth prediction where both monocular camera and people in the scene are freely moving.

2. Motivations

   Those authors want to develop a new method better than those state of art monocular depth prediction method to recover dense depth. And they come up with an idea that, they can maintain a feasible interpretation of the objects' geometry and depth ordering by the stable human depth. As for the data origin, they derive data from YouTube videos in which people imitate mannequins to form their Mannequin Challenge dataset.

3. Approach/Method

   1) Dataset

      They come up with a new dataset which is called Mannequin Challenge dataset(MC) and publish them for people to use. The videos in the MC dataset would be like people freeze in place in an interesting pose while the camera operator moves around the scene filming them. As it is a supervised learning process, the authors need to make "tags" to the frames in the video clip. They start estimating the camera poses and tracks for each frame by ORB-SLAM2[1]. Then they use COLMAP to get the dense depth map, which is a state of the art MVS system[7]. They remove some noises and some wrongly reconstructed images through SfM system[2].

   2) Depth Prediction Model

      They use supervision model and the full inputs are in Fig1:

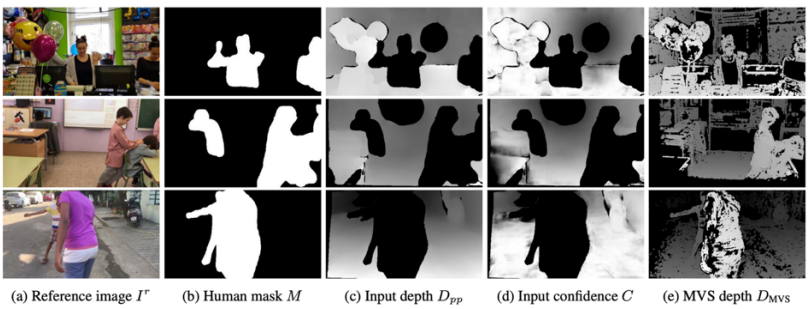

Fig 1-a,b,c,d are inputs, e is target result

I^r is the reference image which is not modified. M is the binary mask of human regions which will mask out depth information in the human region in the $\mathbf{D}_{pp}$ figure and condidence figure **C**.

As for the $\mathbf{D}_{pp}$ input, it is known as depth from motion parallax which is computed from two nearby frames in a video. They use the Plane-Plus-Parallax[4] representation to get the initial depth image.

Next important input is the confidence figure **C**. Since the optical flow used to calculate the $\mathbf{D}_{pp}$ is too noisy, they need a confidence map **C** to make the network relies more on the initial depth map in high confidence regions.

The network architecture is quite similar to the hourglass network of the Chen *et al.*[3]. They just replace the nearest-neighbor upsampling layers by the bilinear upsampling layers.

Finally, they use their own invariant depth regression loss computation to train their model.

$$\mathcal{L}_{\text{si}} = \mathcal{L}_{\text{MSE}} + \alpha_1 \mathcal{L}_{\text{grad}} + \alpha_2 \mathcal{L}_{\text{sm}}.$$

4. Results and Evaluations

To evaluate the model, they use scale-invariant RMSE (si-RMSE) as their error metric which is equivalent to $\sqrt{Lmse}$. They evaluate si-RMSE on 5 different regions: **si-full** for overall depth accuracy, **si-env** for environment regions depth accuracy, **si-hum** for human regions depth prediction accuracy, **si-intra** for human regions but independent to the environment depth and **si-inter** for pixels between human and environment region depth. They evaluate their model on the MC dataset first to check how does the input influence the dense depth prediction in table1.

	Net inputs	si-full	si-env	si-hum	si-intra	si-inter
I.	I	0.333	0.338	0.317	0.264	0.384
II.	$IFCM$	0.330	0.349	0.312	0.260	0.381
III.	$ID_{\text{pp}}M$	0.255	0.229	0.264	0.243	0.285
IV.	$ID_{\text{pp}}CM$	0.232	**0.188**	0.237	0.221	0.268
V.	$ID_{\text{pp}}CMK$	**0.227**	**0.189**	**0.230**	**0.212**	**0.263**

Table 1 Quantitative Comparisons on the MC Dataset

They conclude that by adding the initial depth of environment will both increase the non-human regions and human regions depth prediction accuracy. And they can even add human key point location(**K**) to get a better performance.

To compare with other state of the art depth prediction method, those authors try their own model and other models on TUM RGBD dataset shown in table2.

Methods	Dataset	two-view?	si-full	si-env	si-hum	si-intra	si-inter	RMSE	Rel
Russell *et al.* [31]	-	Yes	2.146	2.021	2.207	2.206	2.093	2.520	0.772
DeMoN [39]	RGBD+MVS	Yes	0.338	0.302	0.360	0.293	0.384	0.866	0.220
Chen *et al.* [3]	NYU+DIW	No	0.441	0.398	0.458	0.408	0.470	1.004	0.262
Laina *et al.* [17]	NYU	No	0.358	0.356	0.349	0.270	0.377	0.947	0.223
Xu *et al.* [46]	NYU	No	0.427	0.419	0.411	0.302	0.451	1.085	0.274
Fu *et al.* [7]	NYU	No	0.351	0.357	0.334	0.257	0.360	0.925	0.194
I	MC	No	0.318	0.334	0.294	0.227	0.319	0.840	0.204
$IFCM$	MC	Yes	0.316	0.330	0.302	0.228	0.323	0.843	0.206
$ID_{\text{pp}}M$	MC	Yes	0.246	0.225	0.260	0.233	0.273	0.635	0.136
$ID_{\text{pp}}CM$ (w/o d. cleaning)	MC	Yes	0.272	0.238	0.293	0.258	0.282	0.688	0.147
$ID_{\text{pp}}CM$	MC	Yes	0.232	0.203	0.252	0.224	0.262	0.570	0.129
$ID_{\text{pp}}CMK$	MC	Yes	**0.221**	**0.195**	**0.238**	**0.215**	**0.247**	**0.541**	**0.125**

Table 2 Results on TUM RGBD datasets

Lower is better for all the error metrics. Their own model has better accuracy than other state-of-the-art methods.

To further prove their model is genetable, they test their model and other models on Internet videos of dynamic scenes. They compare their full model (I D_{pp} C M K) with DORN[5], Chen *et al.* [3]and DeMoN[6] models. The depth prediction maps are shown in fig2.

(a) I^r (b) I^s (c) DORN [7] (d) Chen *et al.* [3] (e) DeMoN [39] (f) Ours (full)

Fig 2 Comparisons on Internet Video Clips with Moving Cameras and People

From those result they conclude that DORN[5] has very limited generalization. Chen *et al.* [3]is not able to capture accurate depth and DeMoN[6] often produces incorrect depth.

5. Applications

This accurate depth prediction can be applied in many visual effects such as depth-based defocus, insertion of synthetic 3D graphics and remove humans with inpainting.

6. Disadvantages

   Moving but non-human objects such as cars and shadows can cause bad predictions. And fine structures such as limbs may be blurred for distant people in challenging poses.

7. Conclusion

   This paper come up with a very interesting MC dataset which will help depth supervised learning improved in the future. And using the initial depth as input with a depth supervision is very new idea for people to try such kind of thing. Also, what they have done is proving another valid method for dense depth prediction in freely moving people and cameras. And how do they evaluate their model and compare it with others is very challenging.

   However, their method also has some limitations like they assume that they know the camera poses. But this paper gives researchers a new idea on how to achieve better depth predictions and improves some depth-based visual effects.

References

[1] R. Mur-Artal and J. D. Tardos. Orb-Slam2: An open-source ´ slam system for monocular, stereo, and RGB-D cameras. IEEE Transactions on Robotics, 33(5):1255–1262, 2017.

[2] J. L. Schonberger and J.-M. Frahm. Structure-from-motion revisited. In Proc. Computer Vision and Pattern Recognition (CVPR), 2016.

[3] W. Chen, Z. Fu, D. Yang, and J. Deng. Single-image depth perception in the wild. In Neural Information Processing Systems, pages 730–738, 2016.

[4] M. Irani and P. Anandan. Parallax geometry of pairs of points for 3d scene analysis. In Proc. European Conf. on Computer Vision (ECCV), pages 17–30, 1996.

[5] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao.Deep ordinal regression network for monocular depth estimation. In Proc. Computer Vision and Pattern Recognition (CVPR), 2018.

[6] B. Ummenhofer, H. Zhou, J. Uhrig, N. Mayer, E. Ilg, A. Dosovitskiy, and T. Brox. DeMoN: Depth and motion network for learning monocular stereo. In Proc. Computer Vision and Pattern Recognition (CVPR), 2017.

[7] J. L. Schonberger, E. Zheng, J.-M. Frahm, and M. Pollefeys. ¨Pixelwise view selection for unstructured multi-view stereo. In Proc. European Conf. on Computer Vision (ECCV), pages 501–518, 2016.