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. Another contribution they did is publishing a new Mannequin Challenge Dataset, which will be discussed later. More specifically speaking, this method is a data driven approach on building a deep neural network.

2. Motivations

   Existing methods for recovering depth for dynamic or non-rigid objects from monocular video impose strong assumptions on the objects' motion and may only recover sparse depth. So those authors want to develop a new method better than those state of art monocular depth prediction method to recover dense depth. Also, people usually use a hand-help camera viewing a dynamic scene. To recover the dense depth in this case is very difficult. Because the moving objects will violate the epipolar constraint used in 3D vision and create a lot of noise in existing structure-from-motion (SfM) and multi-view-stereo (MVS) methods. But the depth prediction of human will not be easily changed even your camera or other objects is moving in your scene. So, these authors come up an idea that, they can maintain a feasible interpretation of the objects' geometry and depth ordering by the stable human depth. And since they do not want to impose the assumptions on the shape or deformation of people, they use a data driven approach to learn these priors from data. 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

      Before those author begin building their own network, they need sufficient and valid training data, validation data and testing data. So 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. These videos are all from YouTube so they can include both indoor and outdoor scenes to be trained. To satisfy their model input, that is not enough. They also assume the scenes are static and obtain accurate camera poses and depth information with SfM and MVS algorithms. 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 for each frame. Their main method for tracking the movement and poses of camera is ORB-SLAM2[1]. This is a complete SLAM system working on

standard CPUs in a wide variety of environments. At the camera poses estimation stage, these authors first process lower-resolution video for efficiency. Then they reprocess each sequence at a higher resolution using a visual SfM system[2]. Finally, they will remove non-smooth camera motion sequences.

After they get the estimated camera poses for each frame, they use COLMAP to get the dense depth map, which is a state of the art MVS system. But the raw outputs from MVS is too hard to be used because there are some camera motion blurs, shadows and reflections. They come up with a careful depth filtering mechanism. The algorism is shown below which is not hard:

$$\Delta(\mathbf{p}) = \frac{|D_{\text{MVS}}(\mathbf{p}) - D_{\text{pp}}(\mathbf{p})|}{D_{\text{MVS}}(\mathbf{p}) + D_{\text{pp}}(\mathbf{p})}$$

Δ(p) stands for the normalized error for every valid pixel **P**. $\mathbf{D}_{\text{Mvs}}$ stands for the depth map obtained by MVS and $\mathbf{D}_{\text{pp}}$ is the depth map computed from two-frame motion parallax. If the **Δ(p)** is larger than 0.2, depth value is removed. Then, they remove some unqualified frames where less than 20% of pixels have valid MVS depth. They also remove frames where the estimated radial distortion coefficient larger than 0.1 or where the estimated focal length is less than 0.6 or larger than 1.2. They keep sequences which have more than 30 frames with ratio 16:9 and have a width more than 1600 pixels. Finally they also removed some wrongly reconstructed images through SfM system. The MC dataset has total 4690 sequences with more than 170k valid image-depth pairs.

2) Depth Prediction Model

An important thing here is, they use the depth generated by the MVS pipeline as the target result of the model. The full input to their network is shown below in Fig1.

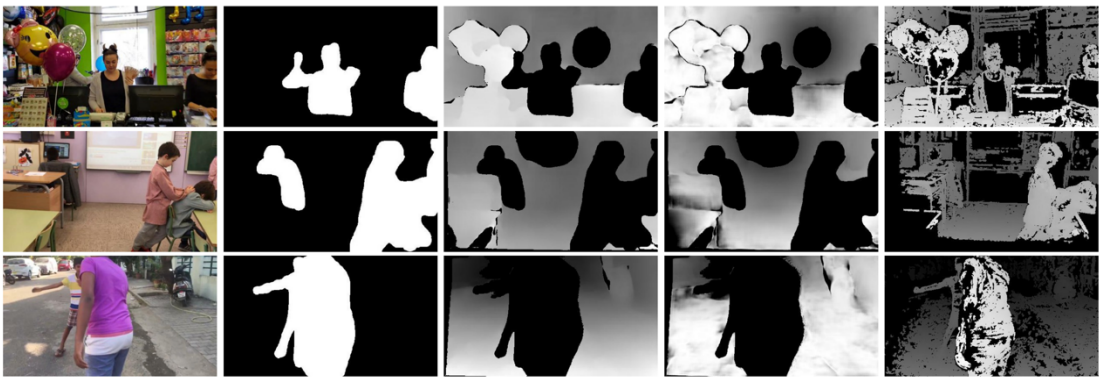

Fig 1-a,b,c,d are inputs, e is target result

$\mathbf{I}^{\mathbf{r}}$ is the reference image which is an original single RGB picture without any modification. M is the binary mask of human regions which will mask out some depth information in the human region in the $\mathbf{D}_{\text{pp}}$ figure and condidence figure **C**. The idea to do so is they want their network inpaint the depth in the human regions, refine the depth of environments and keep the depth consistent.

As for the $\mathbf{D}_{\text{pp}}$ 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. Be aware about that the initial depth map does not include the human regions, so does confidence map. The confidence value is defined as the following formula:

$$C(\mathbf{p}) = C_{lr}(\mathbf{p})C_{ep}(\mathbf{p})C_{pa}(\mathbf{p}).$$

$\mathbf{C}_{lr}$ means "left-right" consistency between the forward and backward fields. $\mathbf{C}_{ep}$ measures how well the flow field complies with the epipolar constraint. $\mathbf{C}_{pa}$ stands for low confidence to pixels for which the parallax between the view is small. As for exactly how to calculate those parameters, you can view the forth page of original paper for more information. When **C(p)** is less than 0.25, the pixel in the confidence map will be masked out.

The network architecture is quite similar to the hourglass network of the network used in "Single-image depth perception in the wild" [3]. They just replace the nearest-neighbor upsampling layers by the bilinear upsampling layers.

Finally, they use their own loss computation to train their model. They use scale-invariant depth regression loss, which is computed on log-space depth values and consists of three items.

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

MSE loss is computed by squared, log-space difference in depth between two pixels in the prediction and the ground truth.

$\mathbf{L}_{grad}$, which is the L1 difference between the predicted log depth derivatives (in x and y directions) and the ground truth log depth derivatives, at multiple scales.

$\mathbf{L}_{sm}$ stands for the smoothness. To encourage smooth interpolation of depth in texture-less regions where MVS fails to recover depth, they use $\mathbf{L}_{sm}$ which penalizes L1 norm of log depth derivatives based on the first- and second-order derivatives of images and is applied at multiple scales.

4. Results and Evaluations

To evaluate the model on dense depth prediction, they need to publish a metric for both their own model and state-of-the-art model. In this paper, they use scale-invariant RMSE (si-RMSE) as their error metric. Si-RMSE 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 prediction accuracy and **si-inter** for pixels between human region environment region depth prediction accuracy. As they are using their own dataset which means, they need their model run another dataset to compare with other models. They evaluate their model on the MC dataset first to check how does the input influence the dense depth prediction. This is shown 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

From these comparisons, 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. After filtering, they got total 11 sequences and 1815 valid images which can be run on their own model. The comparisons result is showing below 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

Dataset "-" means this is not a supervised method. **RMSE** stands for standard RMSE error metric and **Rel** stands for relative error metric. Lower is better for all the error metrics. They conclude that their model strongly resembles the ground truth and show high level of details and sharp depth discontinuities. Their own model has better accuracy than other state-of-the-art methods. Furthermore, they test the w/o d. cleaning, which means without the depth cleaning method just using the raw MVS depth predictions as supervision. They conclude that depth cleaning will support better performance.

To further prove their model is practical, they test their model and other models on Internet videos of dynamic scenes. They generate sequences ranging from 5 seconds to 15 seconds with smooth camera trajectories by their SLAM/SfM pipeline. And 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 map are shown in fig2.

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

From those result they conclude that DORN has very limited generalization. Chen *et al.* is not able to capture accurate depth and DeMoN 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. These works could be seen in fig3.

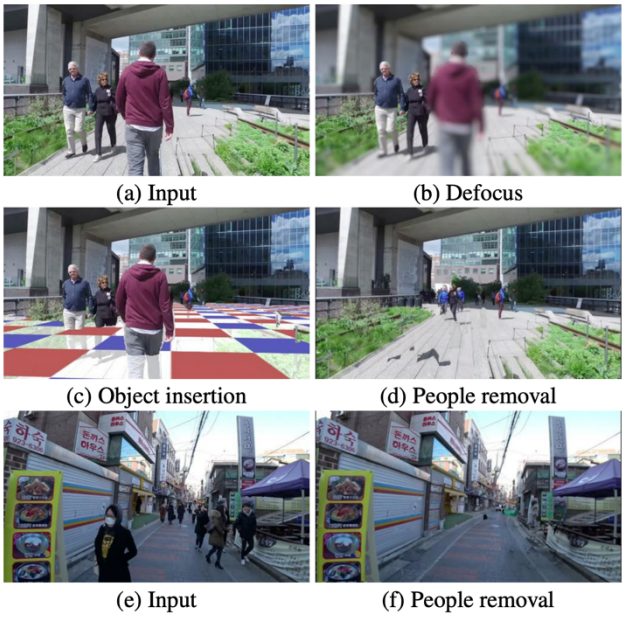

Fig 3 Depth-based visual effects

6. Disadvantages

   This data driven depth prediction method is not perfect right now, as those authors mentioned. 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. Related work

   Learning-based depth prediction cannot predict the depth of dynamic objects. Some

other work like depth estimation for dynamic scenes is not generable, they cannot be applied general people in any kind of scene. As for the dataset, all the previous RGBD data for learning depth cannot provide depth supervision for moving people in natural environment and they are limited to indoor scenes. Also, human shape and pose detection does not care about the depth information.

8. Conclusion

   From what I have learned from the paper, I will talk about my own understanding and how does this paper inspire me on machine vision field. 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. Furthermore, they have tried many recent methods in this area and utilize them to form a better idea such as creating the supervision depth map by MVS system. And how do they evaluate their model and compare it with others is very challenging.

   However, their method also has some limitations like the moving car and shadows brings wrong predictions. Also, 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.