Temporally Consistent Depth Prediction with Flow-Guided Memory Units

Chanho Eom, Hyunjong Park, and Bumsub Ham, *Member, IEEE*

***Abstract*—Predicting depth from a monocular video sequence is an important task for autonomous driving. Although it has advanced considerably in the past few years, recent methods based on convolutional neural networks (CNNs) discard temporal coherence in the video sequence and estimate depth independently for each frame, which often leads to undesired inconsistent results over time. To address this problem, we propose to memorize temporal consistency in the video sequence, and leverage it for the task of depth prediction. To this end, we introduce a two-stream CNN with a flow-guided memory module, where each stream encodes visual and temporal features, respectively. The memory module, implemented using convolutional gated recurrent units (ConvGRUs), inputs visual and temporal features sequentially together with optical flow tailored to our task. It memorizes trajectories of individual features selectively and propagates spatial information over time, enforcing a long-term temporal consistency to prediction results. We evaluate our method on the KITTI benchmark dataset in terms of depth prediction accuracy, temporal consistency and runtime, and achieve a new state of the art. We also provide an extensive experimental analysis, clearly demonstrating the effectiveness of our approach to memorizing temporal consistency for depth prediction.**

***Index Terms*—Depth video prediction, recurrent neural networks, convolutional gated recurrent units**

I. Introduction

DEPTH prediction from images plays a significant role in autonomous driving and advanced driver assistance systems, which helps understanding a geometric layout in a scene, and can be leveraged to solve other tasks, including vehicle/pedestrian detection [1], [2], traffic scene segmentation [3], and 3D reconstruction [4]. Stereo matching is a typical approach to recovering depth that finds dense correspondences between a pair of stereo images [5], [6], [7]. Stereo matching methods compute similarities between local patches [8] or optimize global objective functions to consider smoothness priors penalizing large derivatives of depth [9], [10], [11]. These approaches show state-of-the-art performance, but capturing pairs of stereo images requires multiple cameras calibrated, making it difficult to apply them in practice. An alternative is to predict depth from a monocular video sequence, and it is of great interests in recent years [12], [13], [14], [15], [16], [17], [18], [19]. This approach builds upon the insight that human can perceive depth using monocular depth cues (*e.g.*, occlusion, perspective, motion parallax) only [20]. Eigen *et al.* [12] first propose a supervised learning method for predicting depth from a single still image using CNNs. Zhou *et al.* [14] and Wang *et al.* [15] recently propose CNN architectures for predicting depth from a monocular video, where two networks are trained separately to estimate depth and camera pose. These methods are limited in that they predict depth independently for each frame, discarding temporal coherence in the video sequence. That is, they give temporally inconsistent results, causing serious temporal flickering artifacts. Recurrent neural networks (RNNs) have been widely used to model temporal dependency across sequential data (*e.g.*, video and text), and they have shown the effectiveness in various applications including action recognition [21] and machine translation [22]. They, however, still show a limited capability of handing the flickering artifacts [18], [23].

In this paper, we present a simple yet effective method for a temporally consistent depth prediction from a monocular video sequence (Fig. 1). We transfer temporal consistency in the video to RNNs explicitly, particularly using convolutional gated recurrent units (ConvGRUs) [24]. To implement this idea, we propose a flow-guided memory unit using optical flow specific to our task, maintaining a long-term temporal consistency in depth prediction results. Our module uses spatial and temporal features extracted by a two-stream CNN. We have two main reasons for decoupling these features. First, it has been proven that learning spatiotemporal features jointly from a stack of frames does not capture the motion well [25]. Second, optical flow itself provides an important clue for motion parallax, which is helpful to infer depth from a monocular video sequence. For example, objects closer to a camera move faster than distant ones. We show that our method outperforms the state of the art in terms of temporal consistency, and shows a good trade-off between depth prediction accuracy and runtime. The main contributions of this paper can be summarized as follows:

- We present an effective ConvGRU encoder-decoder module for a temporally consistent depth prediction from a monocular video sequence. To our knowledge, this is the first approach based on convolutional/recurrent networks to considering temporal consistency in depth prediction.
- We propose a flow-guided memory unit that retains a long-term temporal consistency explicitly for individual pixels.
- We present state-of-the-art results on the KITTI [26] benchmark. We additionally provide an extensive experimental analysis, clearly demonstrating the effectiveness of our

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (No.2016-0- 00197, Development of the high-precision natural 3D view generation technology using smart-car multi sensors and deep learning).

C. Eom, H. Park, B. Ham are with the School of Electrical and Electronic Engineering, Yonsei University, Seoul 03722, South Korea (e-mail: cheom@yonsei.ac.kr; hyunpark@yonsei.ac.kr; bumsub.ham@yonsei.ac.kr). Corresponding author: Bumsub Ham.