# Recurrent Convolutional Neural Network for Object Recognition

Xuewen Yang

July 26 2018

## Abstract

*Learning to estimate 3D geometry in a single image by watching unlabeled videos via deep convolutional network is attracting significant attention. In this paper,the author introduce a 3D as-smooth-as-possible (3D-ASAP) prior inside the pipeline, which enables joint estimation of edges and 3D scene, yielding results with significant improvement in accuracy for fine detailed structures. Specifically, they define the 3D-ASAP prior by requiring that any two points recovered in 3D from an image should lie on an existing planar surface if no other cues provided.They design an unsupervised framework that Learns Edges and Geometry (depth, normal) all at Once (LEGO). The predicted edges are embedded into depth and surface normal smoothness terms, where pixels without edges in-between are constrained to satisfy the prior. In their framework, the predicted depths, normals and edges are forced to be consistent all the time.*

## 1. Introduction

Recently, impressive progress [3] [5] has been made to mimic detailed 3D reconstruction by training a deep network taking only unlabeled videos or stereo images as input and testing on monocular image, yielding even better depth estimation results than those of supervised methods [2] in outdoor scenarios. The core underlying idea is the supervision by view synthesis, where the frame of one view (source) is warped to another (target) based on the predicted

## 2. Related Work

Geometric based methods estimate 3D from a given video with feature matching, such as SFM, SLAM and DTAM, which could be effective and efficient in many cases.Deep neural networks (DCN) developed in recent years, e.g.VGG and ResNet, provide strong feature representation. Dense geometry, i.e., pixel-wise depth and normal maps, can be readily estimated from a single image [1]. The learned CNN model shows significant improvement compared to other methods based on hand-crafted features.Motivated by traditional methods, videos, which are easier to obtain and hold richer 3D information. Motivated by traditional methods like SFM and DTAM, lots of CNN based methods are proposed to do single view geometry estimation with supervision from vieos, and yield impressive progress.Long range and non-local spatial regularization has been vastly explored in classical graphical models like CRF [4], where nodes beyond the neighboring are connected, and the smoothness in-between are learned with high-order CRF or densely-connected CRF.

## 3. Conclusion

In this paper,the author proposed LEGO, an unsupervised framework for joint depth, normal and edge learning. A novel 3D-ASAP prior is proposed to better regularize the learning of scene layout. This regularization jointly considers the three important descriptors of 3D scene and improves the results on all tasks: depth, normal and edge estimation.They conducted comprehensive experiments to present the performance of LEGO. On KITTI dataset, LEGO achieves SOTA performance on both depth and normal evaluation. For edge evaluation, LEGO outperformes the other methods by a large margin on Cityscapes dataset.

## References

[1] D. Eigen and R. Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *ICCV*, pages 2650–2658, 2015. 1

[2] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NIPS*, 2014. 1

[3] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2016. 1

[4] J. D. Lafferty, A. Mccallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, pages 282–289, 2001. 1

[5] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017. 1