Unsupervised Monocular Depth and Latent Structure

Kenneth Chaney*
chaneyk@seas.upenn.edu

Bernadette Bucher*
bucherb@seas.upenn.edu

Evangelos Chatzipantazis
vaghat@seas.upenn.edu

Jianbo Shi
jshi@seas.upenn.edu

Kostas Daniilidis
kostas@seas.upenn.edu

The GRASP Laboratory, University of Pennsylvania

Abstract

We propose a novel method to learn a latent representation of the structural information of a scene. We use this separation of structural versus semantic data for novel viewpoint synthesis. We achieve competitive results through a selection of meaningful loss functions. We train and test our model using stereo data from the KITTI data set to generate novel viewpoints of scenes in which objects undergo forward and backward motion. We achieve a stable latent representation of the structural information in our scenes under these transformations.

1. Introduction

The ability to recognize shape and structure of 3D objects and scenes is a well-studied problem in computer vision. More recently, the question of novel viewpoint synthesis has emerged from the advances in this domain. Novel viewpoint synthesis addresses the task of rendering an object or scene from a different viewpoint than the one given. In particular, research in this area advances the ability to accurately render elements of shape and structure not visible from the original viewpoint. The changes in viewpoint typically target a specific collection of transformations such as object rotation, stereo camera views, or depth changes.

In this work, we present the following contributions.

- We present a method for separating structure and semantic information in images. This method produces depth information from a single image of a scene.
- We use our learned structural information to generate the opposite image of a stereo pair given a single view of a scene.
- We demonstrate the ability to extrapolate forward motion within a scene using our learned depth result.

2. Related Work

Full metric structure estimation requires the memorization capabilities of deep networks to be heavily leveraged. Existing methods use photo-consistency and cycle-consistency losses at the end of the network enabling end-to-end training of depth [6, 11], optical flow [8], pose, or a combination of the three [10].

Similar concepts have been used for novel viewpoint

*These authors contributed equally.

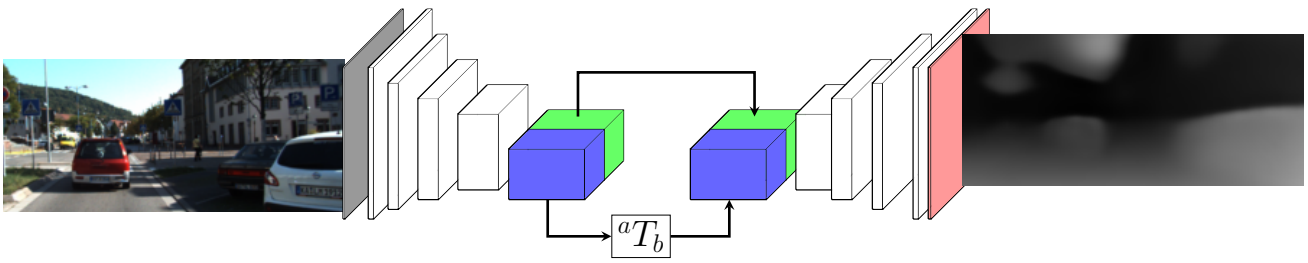

Figure 1: Architecture overview of the proposed depth estimation network. The latent space is divided into two regions: structure and semantic. The structural, blue, is enforced through the use of a transformation into another arbitrary frame (e.g. the stereo pair or another frame in time). While the semantic, green, is kept constant indicating that the same scene is being viewed.

generation. In [4], a novel viewpoint is generated with the constraint that the majority of the new viewpoint has been seen. This constraint is implicit in the chosen representation (a re-projection of depth and semantic information). In [3], the latent representation of the network is used to construct a viewpoint agnostic scene representation. This representation is assembled from multiple frames and then queried for a novel viewpoint anywhere within the scene. This work shows the ability of a deep network to understand the full context of the object including portions that are not explicitly visible.

Constraining the latent representation learned by the network can be used to generate new and valid outputs. Done in a meaningful way, enforcing constraints on the latent space can yield useful properties. For example, the Variational Autoencoder [2] pushes the latent representation of the network onto a more continuous manifold than given by a traditional autoencoder. This continuity property allows for traversal across the manifold.

Recently, there has been work in more complex manipulations of the latent space which apply standard object-centric rotations to generate new images. In [9], the latent space is partitioned into two discrete segments: the first encompassing the semantic information in the scene and the second representing the human pose. Rhodin, et al. utilize a multi-camera setup to train the network. The latent representation is enforced in training, and the structural portion of the resulting representation can be arbitrarily rotated in human pose experiments. In particular, their results demonstrate images of the same scene with only the human rotated.

In work concurrent to ours, Chen, et al. present a structural and semantic partition of the latent space. [1] Their similar approach performs end-to-end training with a reconstruction loss where rotations and translations are explicitly applied in the latent space of their network. They use their method in novel viewpoint synthesis experiments and demonstrate the usefulness of their generated views in a SLAM system.

3. Method

We propose a network to enable both novel viewpoint construction and future frame prediction for scenes. We use a collection of losses both in the latent space and at the end of the network to enforce a separation of structural and semantic information within the latent space. The output of our network is a disparity image which also allows for monocular depth prediction. We visualize this conceptual view of our model in Figure 1.

The primary structure of our network is adapted from the model described in [6]. The underlying architecture is a ResNet with the latent space segmented into structural and semantic components [7]. A description of the variables we use throughout the description of our model is provided below.

I_i^L, I_i^R	left and right stereo images at frame time i
S^L, S^R	left and right semantic components of the latent space
X_i^L, X_i^R	left and right structural components of the latent space indexed by a frame in time
S	semantic component of the latent space after training
X_i	structural component of the latent space after training
$T_{i,k}$	ground truth temporal transformation between time indexed frames
$T_{L,R}$	ground truth spatial transformation between stereo pairs
D_i^L, D_i^R	learned left and right disparity images indexed by a frame in time
$W(\cdot,\cdot)$	inverse warps an image given a disparity

To train our network, we use time-sequenced stereo images. The execution of our network during training is visualized in Figure 2.

For a given time step, we feed in each image in the stereo pair separately into our network. Since each image in the stereo pair is training the same network, we want the structural and semantic components of our latent space to be the same regardless of whether we are using the left or right image. Furthermore, learning the equivalence of the scene from the left and right viewpoints will embed a 6-DoF representation of the scene into the latent space of our network. This logic leads us to develop the losses for our semantic information,

$$L_1(S^R, S^L), \tag{1}$$

and our structural information,

$$L_1(X_i^L, T_{L,R}X_i^R), \tag{2}$$

applied in the latent space of our network during training.

The transformation $T_{L,R}$ which we apply to the structural component of the latent space is computed from the separately input stereo images at each training step.

We recall that our goal is not to output a disparity image at our current time step. We want to output a disparity image associated with a future time step k. To achieve this result, we apply temporal transformation $T_{i,k}$ to the structural component of the latent space before our disparity image is output. In both the testing and training of our model, this temporal transformation is computed directly from images at different time steps. However, if we wished to predict a transformation via a separate process, we could also use that prediction in our model.

After we output the disparity images computed from each stereo image in a training step, we enforce left-right

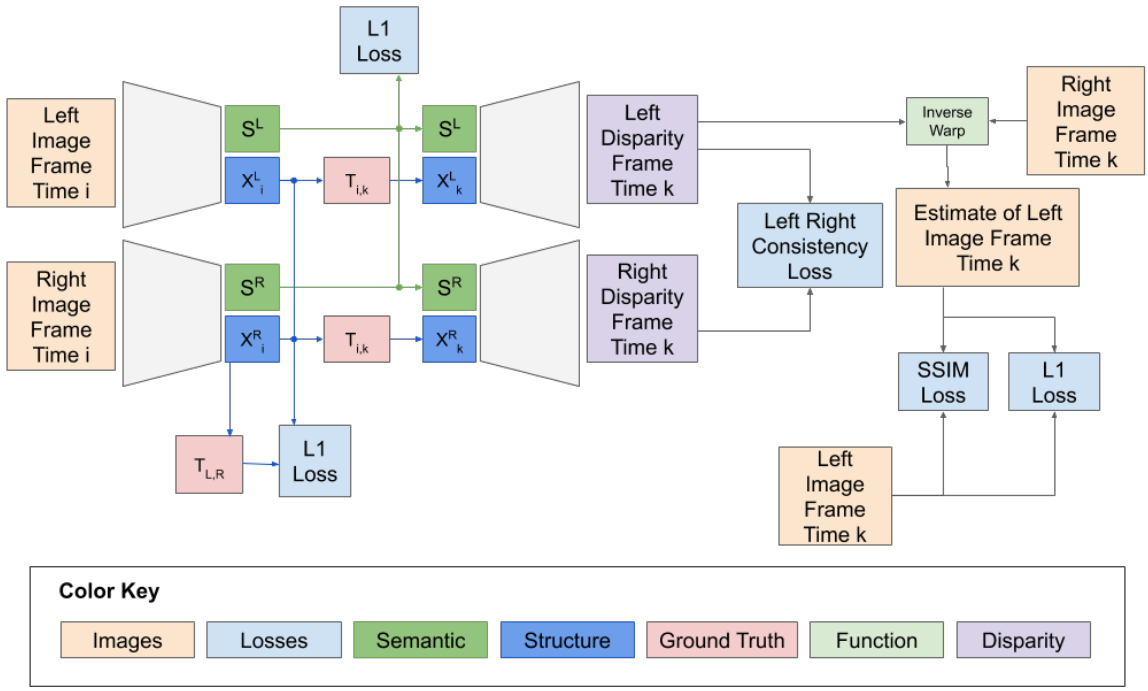

Figure 2: Overview of the model architecture.

consistency via an L_1 loss shown below.

$$L_1(D_k^L, W(D_k^L, D_k^R)) + L_1(D_k^R, W(D_k^R, D_k^L)) \quad (3)$$

For each resulting disparity image, we then compute two additional losses. We first create an estimate of the opposite image in the future stereo pair. Specifically, we compute

$$\widehat{I_k^L} = W(D_k^L, I_k^R) \quad (4)$$

for each stereo image using the predicted disparity and true future image frames. Then, we use an L_1 loss computed as

$$L_1(\widehat{I_k^L}, I_k^L) \quad (5)$$

for each stereo image as well as an $SSIM$ loss computed as

$$SSIM(\widehat{I_k^L}, I_k^L) \quad (6)$$

for each stereo image to enforce similarity between our predicted scene views and the true images.

To test our method, a single image is sent through the network. We use ground truth temporal transformations in the latent space. Then, the resulting disparity prediction is applied to a future view of the scene to generate a novel viewpoint.

4. Results

To test the success of our model, we performed an experiment using the KITTI dataset [5]. We chose this dataset since it is a common performance benchmark for other state of the art methods. We design our experiments to highlight our network's capability of inferring information from a scene not visible from a given viewpoint. To demonstrate this capability, we display results for scenes in which objects move forward and backward since the KITTI dataset primarily showcases objects undergoing forward motion.

To test the network, we first calculated disparity between close-in-time left image frames from a stereo pair. The sequentially first in time source images are shown in the first column of Figure 3. The source images shown here are all identical.

We predict the disparity between the time-offset frames and the source frames with our network. These disparity predictions are shown in the fourth column of Figure 3. We compare our results to the Monodepth Network shown in the fifth column of Figure 3.

Recall that our source frames are the left images from a stereo pair. We take the known right image stereo pair from the target time offset, and we warp this image frame using the predicted disparity from our network. This warping results in a prediction of the left image frame at the target time step. These time offset left image frame predictions are shown in the second column of Figure 3. The ground truth time offset left image frames for the same scene are shown in the third column of Figure 3. The time offset for these images is listed on each row.

Figure 3: Future frame novel viewpoint results. Predicted disparity comparison with the Monodepth network.

Index	0	2	4	6	Index	12	14	16	18
Monodepth Left	0.046	0.044	0.043	0.041	Monodepth Left	0.044	0.044	0.047	0.046
Monodepth Right	0.045	0.043	0.042	0.046	Monodepth Right	0.043	0.044	0.043	0.045
Ours Left	0.061	0.055	0.056	0.051	Ours Left	0.056	0.054	0.056	0.058
Ours Right	0.063	0.055	0.055	0.055	Ours Right	0.054	0.055	0.056	0.057

Table 1: Photometric error of the results shown in Figures 3. For the monodepth network each actual image at the index is fed in. In contrast, for our method we feed in the starting image and transform the latent space to the proper index. KITTI Day 2011/09/26 Sequence 0059 Frames 0-20. Total distance traveled was approximately 8m.

5. Conclusion

We show the ability to learn scene structure separate from semantic information in our network. We also demonstrate how to use this learned structure to generate novel views of scenes from different perspectives. Proving this capability opens up a number of future research questions.

The network should be used to predict motion beyond translations by expanding our training set and adapting the network to yield comparable results on a greater variety of viewpoint changes. Additional experiments should include utilizing image views other than stereo pairs. In particular, the Oxford Robocar dataset provides forward, side, and backward camera views which could yield interesting additional results with our model.

Acknowledgments: NSF-IIP-1439681 (I/UCRC), NSF-IIS-1703319, NSF MRI 1626008, ARL RCTA W911NF-10-2-0016, ONR N00014-17-1-2093, ARL DCIST CRA W911NF-17-2-0181, the DARPA-SRC C-BRIC, and by Honda Research Institute.

References

[1] X. Chen, J. Song, and O. Hilliges. NVS machines: Learning novel view synthesis with fine-grained view control. *CoRR*, abs/1901.01880, 2019. 2

[2] C. Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016. 2

[3] S. A. Eslami, D. J. Rezende, F. Besse, F. Viola, A. S. Morcos, M. Garnelo, A. Ruderman, A. A. Rusu, I. Danihelka, K. Gregor, et al. Neural scene representation and rendering. *Science*, 360(6394):1204–1210, 2018. 2

[4] J. Flynn, I. Neulander, J. Philbin, and N. Snavely. Deepstereo: Learning to predict new views from the world's imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5515–5524, 2016. 2

[5] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 3

[6] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2017. 1, 2

[7] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. 2

[8] S. Meister, J. Hur, and S. Roth. Unflow: Unsupervised learning of optical flow with a bidirectional census loss. *arXiv preprint arXiv:1711.07837*, 2017. 1

[9] H. Rhodin, M. Salzmann, and P. Fua. Unsupervised geometry-aware representation for 3d human pose estimation. *arXiv preprint arXiv:1804.01110*, 2018. 2

[10] S. Vijayanarasimhan, S. Ricco, C. Schmid, R. Sukthankar, and K. Fragkiadaki. Sfm-net: Learning of structure and motion from video. *arXiv preprint arXiv:1704.07804*, 2017. 1

[11] Y. Zhong, Y. Dai, and H. Li. Self-supervised learning for stereo matching with self-improving ability. *arXiv preprint arXiv:1709.00930*, 2017. 1