Veritatem Dies Aperit - Temporally Consistent Depth Prediction Enabled by a Multi-Task Geometric and Semantic Scene Understanding Approach (Supplementary Material)

Amir Atapour-Abarghouei[1] Toby P. Breckon[1,2]
[1]Department of Computer Science – [2]Department of Engineering
Durham University, UK
{amir.atapour-abarghouei,toby.breckon}@durham.ac.uk

1. Introduction

In this document, we provide additional information that could not be placed within the main paper due to space restrictions. We kindly invite the readers to watch the **video** submitted as part of the supplementary material along with this document.

2. Hole Prediction Network

As mentioned in the main manuscript, when the model is being trained to perform depth completion, the input must be a four-channel RGB-D image, in which the depth channel contains holes that would naturally occur when sensed through imperfect capture technologies. However, the dataset used for training our model [11] consists of pixel-prefect depth images without any holes.

This synthetic dataset [11] does contain stereo image pairs, so a simple solution would be to calculate the disparity and subsequently the depth using a well-established stereo matching approach such as Semi-Global Matching [6] and use the resulting depth image (which will contain holes) as the input.

However, each image in a stereo pair in [11] (left and right) comes with its own corresponding (left and right) depth image, and half of the dataset (aligned RGB and depth images) will be rendered useless if stereo matching is used to calculate depth images with hole.

As a result, we opt for training an entirely separate model that would be responsible for creating holes in the depth images. Even though the details regarding the training or use of this network have no bearing on the approach proposed in the main manuscript, we will attempt to cover the inner workings and experimental evaluation of our *hole prediction* model here.

This *hole prediction* model is a fully convolutional encoder-decoder network inspired by [10] with skip connections between all corresponding layers in the encoder and the decoder. The last decoder layer is connected to a soft-max classifier. Each convolutional layer is followed by batch normalization [8] and a ReLU. The network architecture can be seen in Figure 1.

The training data for this *hole prediction* network is made up of $30,000$ pairs of stereo images from [4]. Disparity is calculated using Semi-Global Matching (SGM) [6] and a hole mask (M) is subsequently generated which indicates which pixels are holes. Although SGM is used here, this is interchangeable with any other passive or active depth capture approach. The left RGB images are thus used as inputs with the generated masks as ground truth labels. Binary cross-entropy is used as the loss function since the segmentation task involves only two classes: hole and non-hole.

Qualitative analyses reveal that holes are predicted where expected. From Figure 2, we see that in regions where camera overlap is absent or featureless surfaces, sparse shrubbery, unclear object boundaries, and very distant objects are present, such pixels are correctly classified as holes.

3. Additional Experiments

Following the conventions of the expansive literature on monocular depth estimation, we measure the performance of our approach against the KITTI dataset [4]. However, we have re-trained and tested all the comparators using the synthetic dataset of [11] but for brevity and due to our superior performance on the *unseen* KITTI dataset, against comparators *actually* trained on KITTI, we have not included these extra results in the main manuscript. Table 1 of this document presents the comparison of our approach against [5, 13] trained on the synthetic dataset of [11] under the exact same conditions as outlined in Section 3.4 of the main manuscript. Our approach outperforms the comparators by a large margin (Table 1).

Method	Error				Accuracy
	Abs. Rel.	Sq. Rel.	RMSE	RMSE log	$\sigma < 1.25^3$
[13]	0.401	1.601	6.598	0.363	0.788
[5]	0.334	1.556	6.304	0.302	0.852
Ours (full)	**0.208**	**1.402**	**6.026**	**0.269**	**0.926**

Table 1: Comparisons using synthetic data [11].

4. Figures

Due to the space restrictions, some of the figures within the main paper may be too small for appropriate viewing. While some of the results are better seen in the accompanying video, we also provide enlarged versions of some of the figures here. Please see Figures 3, 4, 5, 6, 7 and 8 in this document.
Video URL: https://vimeo.com/325161805

References

[1] Amir Atapour-Abarghouei, Gregoire Payen de La Garanderie, and Toby Breckon. Back to butterworth - a fourier basis for 3D surface relief hole filling within RGB-D imagery. In *Int. Conf. Pattern Recognition*, pages 2813–2818, 2016.

[2] Gabriel Brostow, Julien Fauqueur, and Roberto Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 30(2):88–97, 2009.

[3] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *IEEE Conf. Computer Vision and Pattern Recognition*, pages 3213–3223, 2016.

[4] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *Robotics Research*, pages 1231–1237, 2013.

[5] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *IEEE Conf. Computer Vision and Pattern Recognition*, pages 6602 – 6611, 2017.

[6] Heiko Hirschmuller. Stereo processing by semi-global matching and mutual information. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 30:328–341, 2008.

[7] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Trans. Graphics*, 36(4):107, 2017.

[8] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Int. Conf. Machine Learning*, pages 1–9, 2015.

[9] Junyi Liu, Xiaojin Gong, and Jilin Liu. Guided inpainting and filtering for kinect depth maps. In *Int. Conf. Pattern Recognition*, pages 2055–2058, 2012.

[10] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Int. Conf. Medical Image Computing and Computer-Assisted Intervention*, pages 234–241, 2015.

[11] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio Lopez. The SYNTHIA Dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *IEEE Conf. Computer Vision and Pattern Recognition*, pages 3234–3243, 2016.

[12] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *IEEE Conf. Computer Vision and Pattern Recognition*, pages 1–15, 2018.

[13] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and ego-motion from video. In *IEEE Conf. Computer Vision and Pattern Recognition*, pages 6612–6619, 2017.

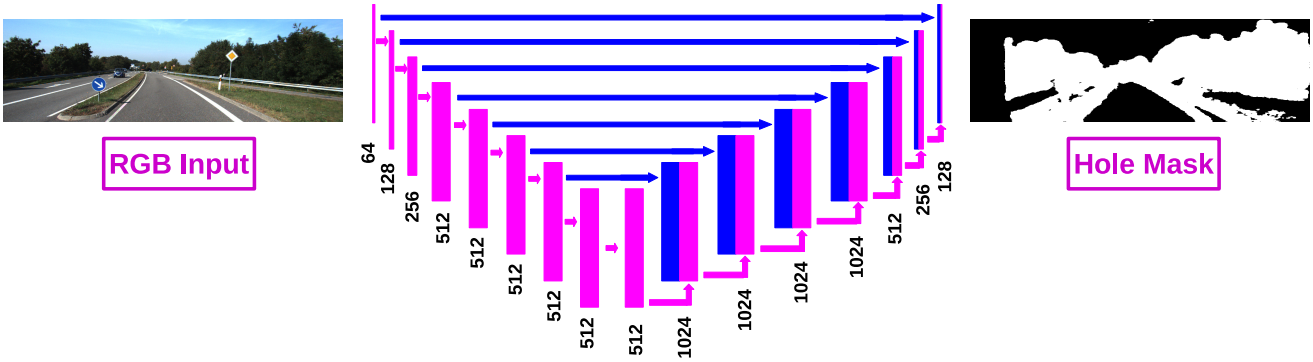

Figure 1: Overview of the architecture of the *hole prediction* network.

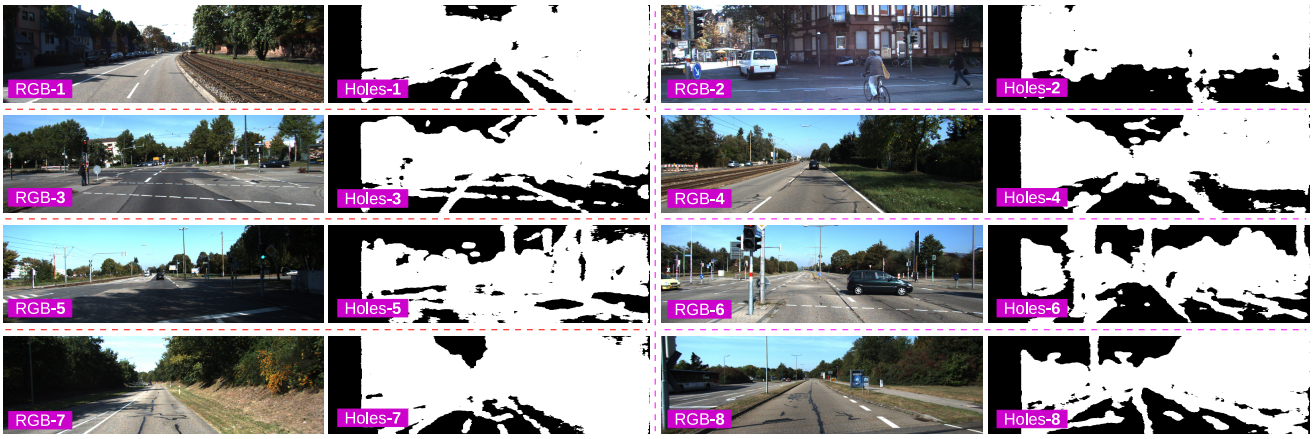

Figure 2: Examples of results of the *hole prediction* model applied to unseen images from [4].

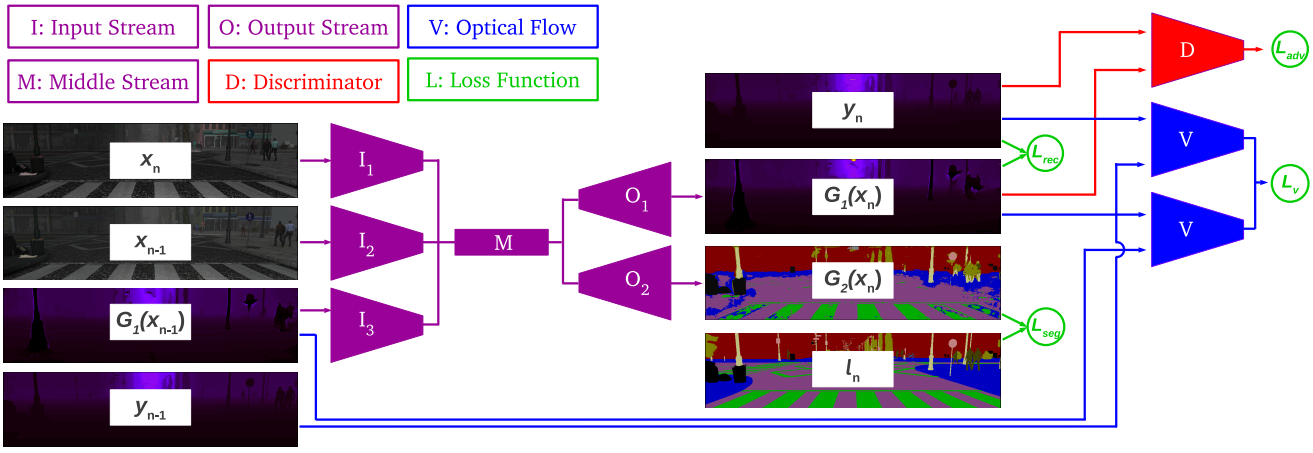

Figure 3: An outline of the training procedure of the main proposed approach.

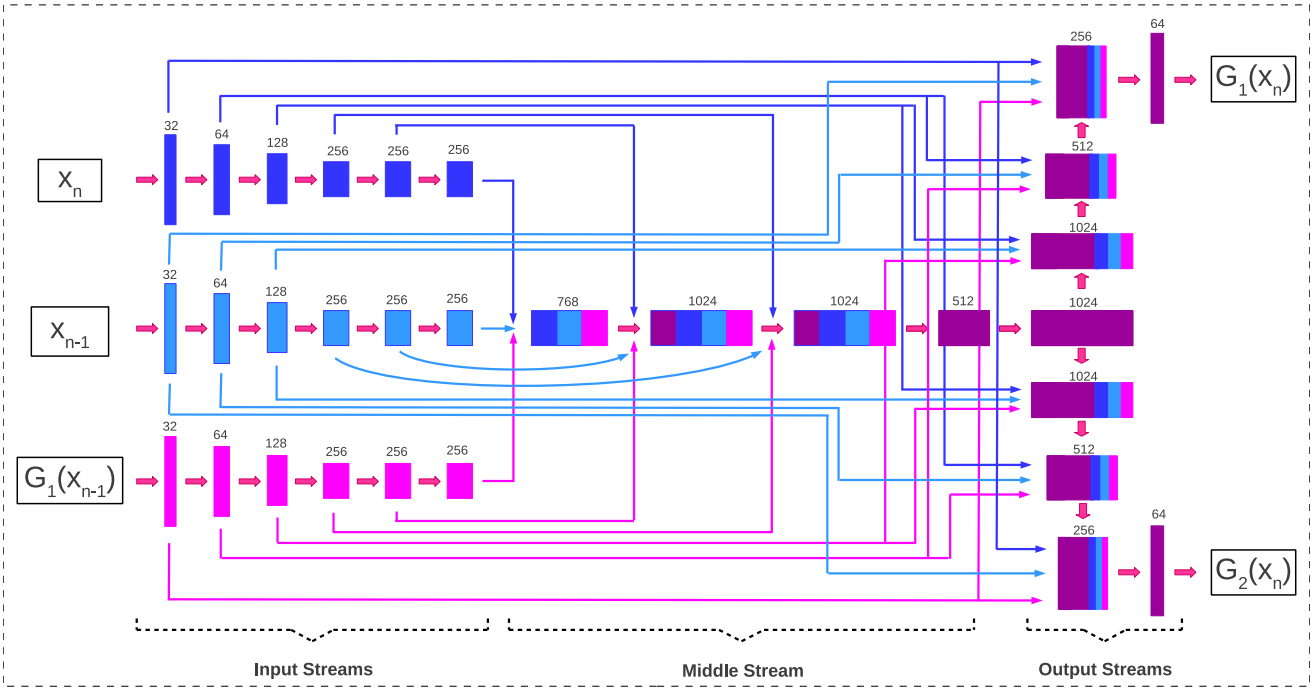

Figure 4: An overview of the generator architecture.

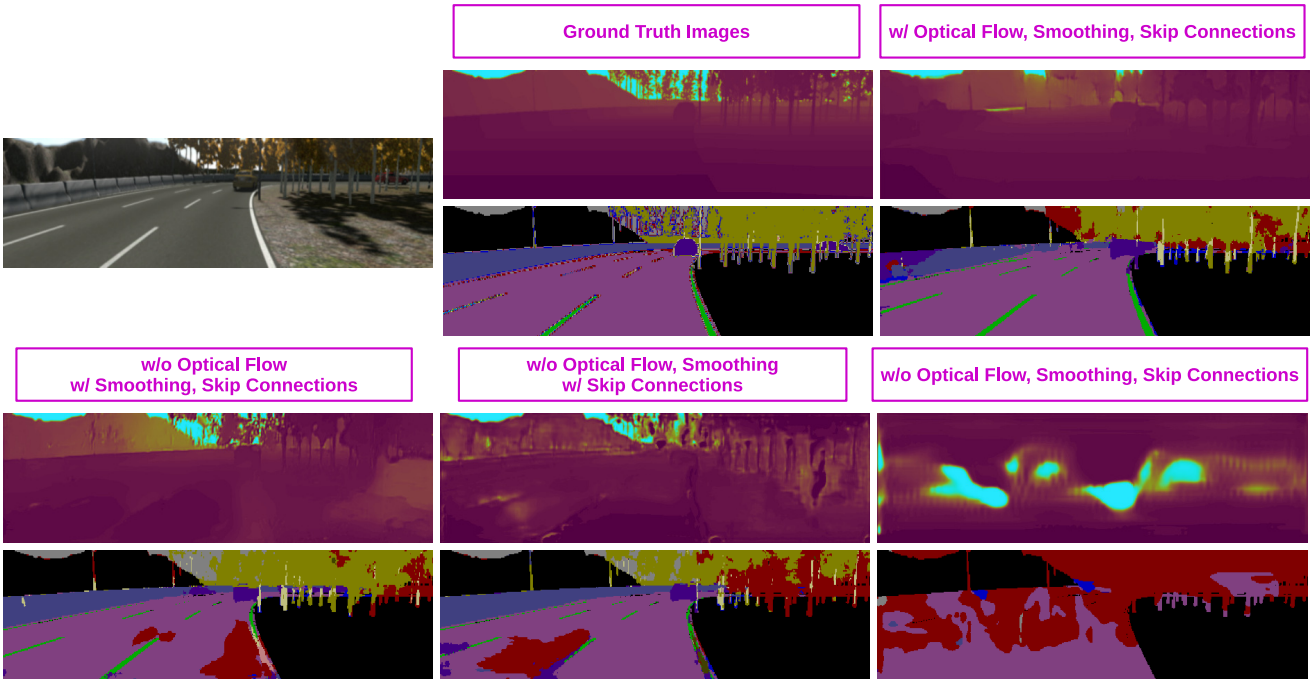

Figure 5: Comparing the results of our model (with monocular depth estimation) when different components of the approach are removed.

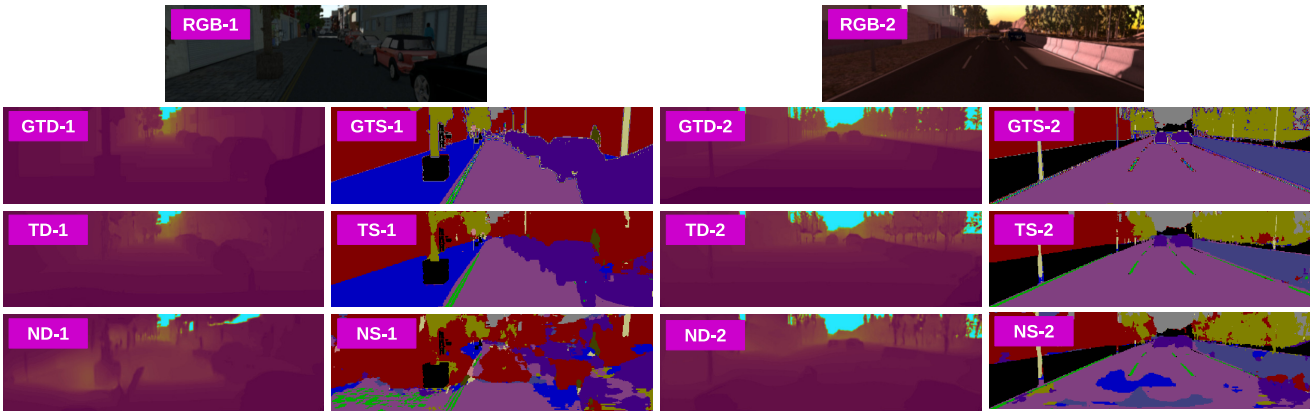

Figure 6: Comparing the results of the approach on the synthetic test set when the model is trained with and without temporal consistency. **RGB:** input colour image; **GTD:** Ground Truth Depth; **GTS:** Ground Truth Segmentation; **TS:** Temporal Segmentation; **TD:** Temporal Depth; **NS:** Non-Temporal Segmentation; **ND:** Non-Temporal Depth.

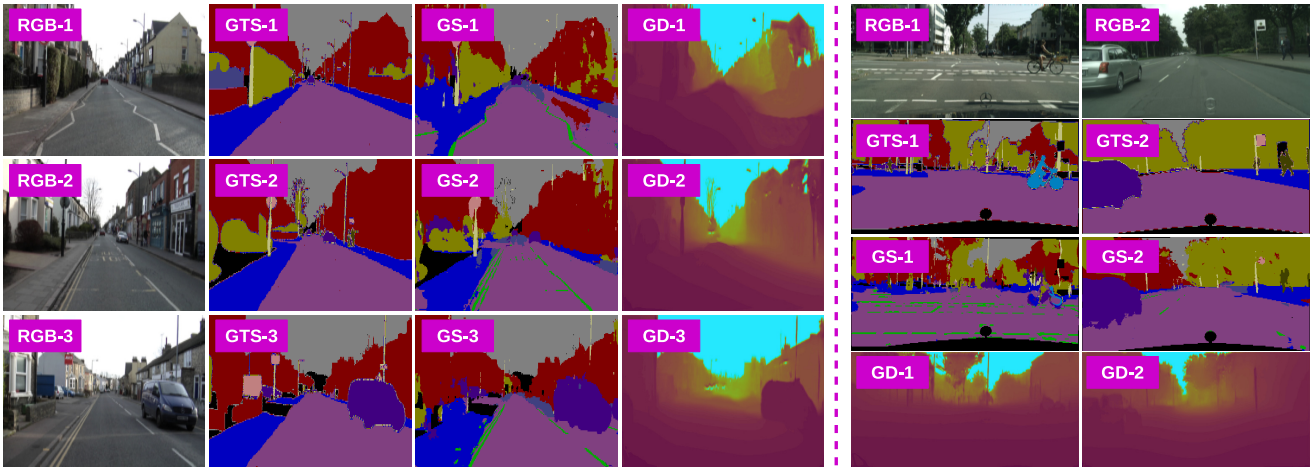

Figure 7: Results of our approach on CamVid [2] (left) and Cityscapes [3] (right) datasets. **RGB:** input colour image; **GTS:** Ground Truth Segmentation; **GS:** Generated Segmentation; **GD:** Generated Depth.

Figure 8: Comparison of depth completion methods applied to synthetic test set. **RGB:** input colour image; **GTD:** Ground Truth Depth; **DH:** Depth Holes; **FDF:** Fourier based Depth Filling [1]; **GTS:** Global and Local Completion [7]; **ICA:** Inpainting with Contextual Attention [12]; **GIF:** Guided Inpainting and Filtering [9].