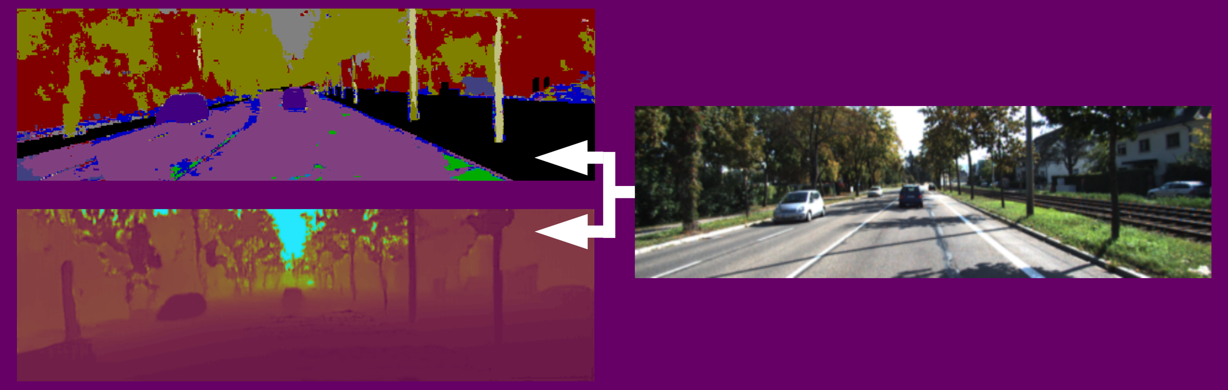
Veritatem Dies Aperit - Temporally Consistent Depth Prediction Enabled by a Multi-Task Geometric and Semantic Scene Understanding Approach

Amir Atapour-Abarghouei and Toby P. Breckon, Durham University, UK



Monocular input → { **Depth** image **Segmentation** image }

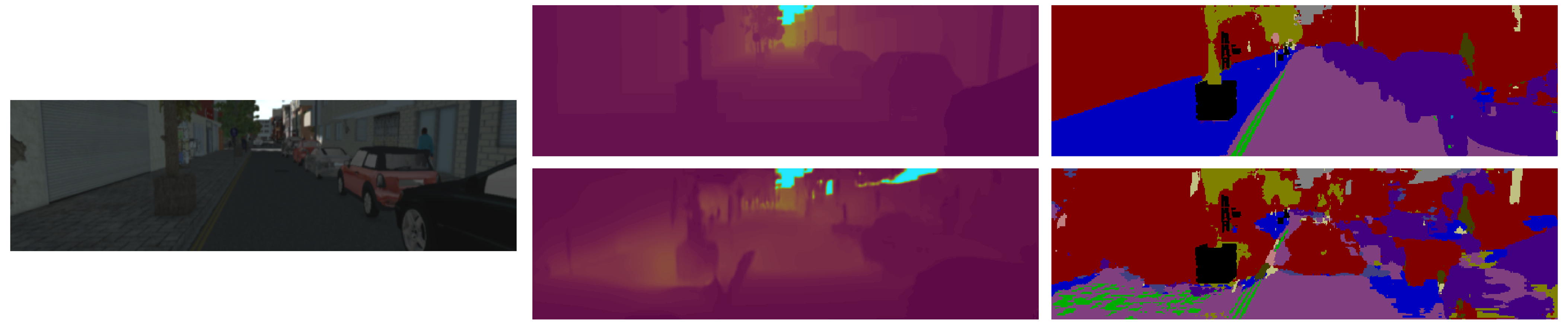
Results include sharp temporally consistent depth images along with semantic segmentation.

Contributions:

Primary objective: *depth prediction* (**monocular depth estimation** and **depth completion**). Via synthetic training images [1] from urban driving scenarios, our approach also performs **semantic scene segmentation** in a single **multi-task** deep network.

Model produces temporally consistent results via the temporal constraint of sequential frame **recurrence** and a per-trained **optical flow** network.

with temporal consistency



without temporal consistency

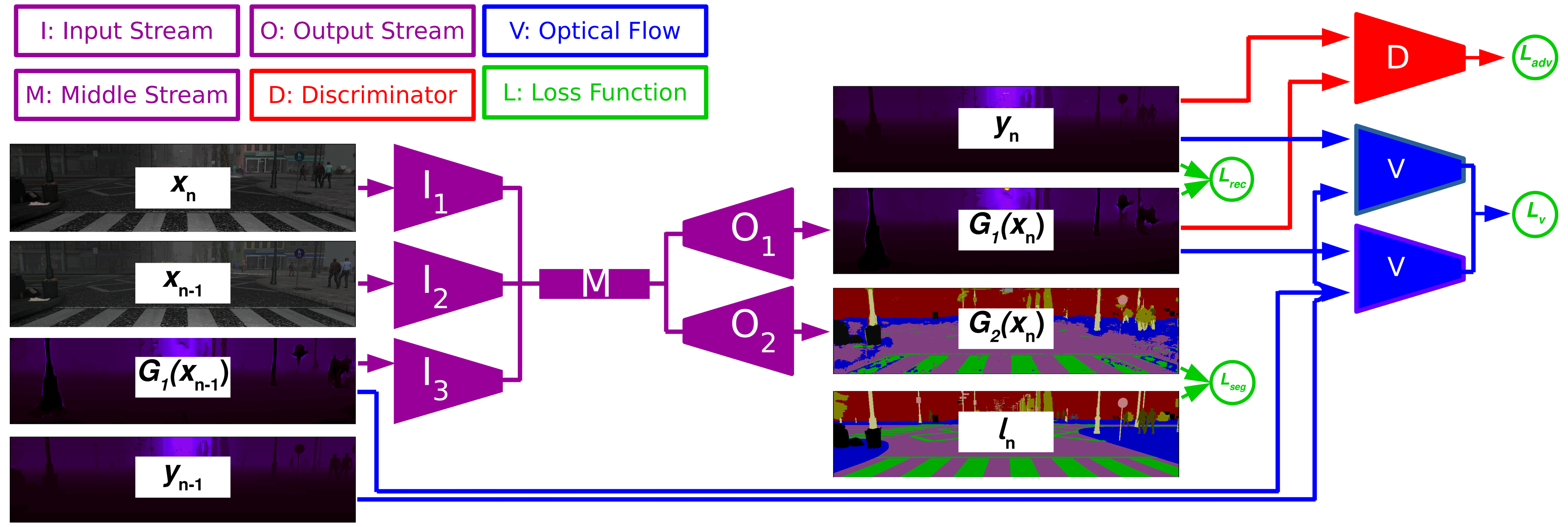
Via deeper and improved **scene representation** learned by the multi-task model, it generalises to **real-world imagery** without domain adaptation.

The approach is capable of producing superior results by preserving high-level spatial features using a complex network of **skip connections**.

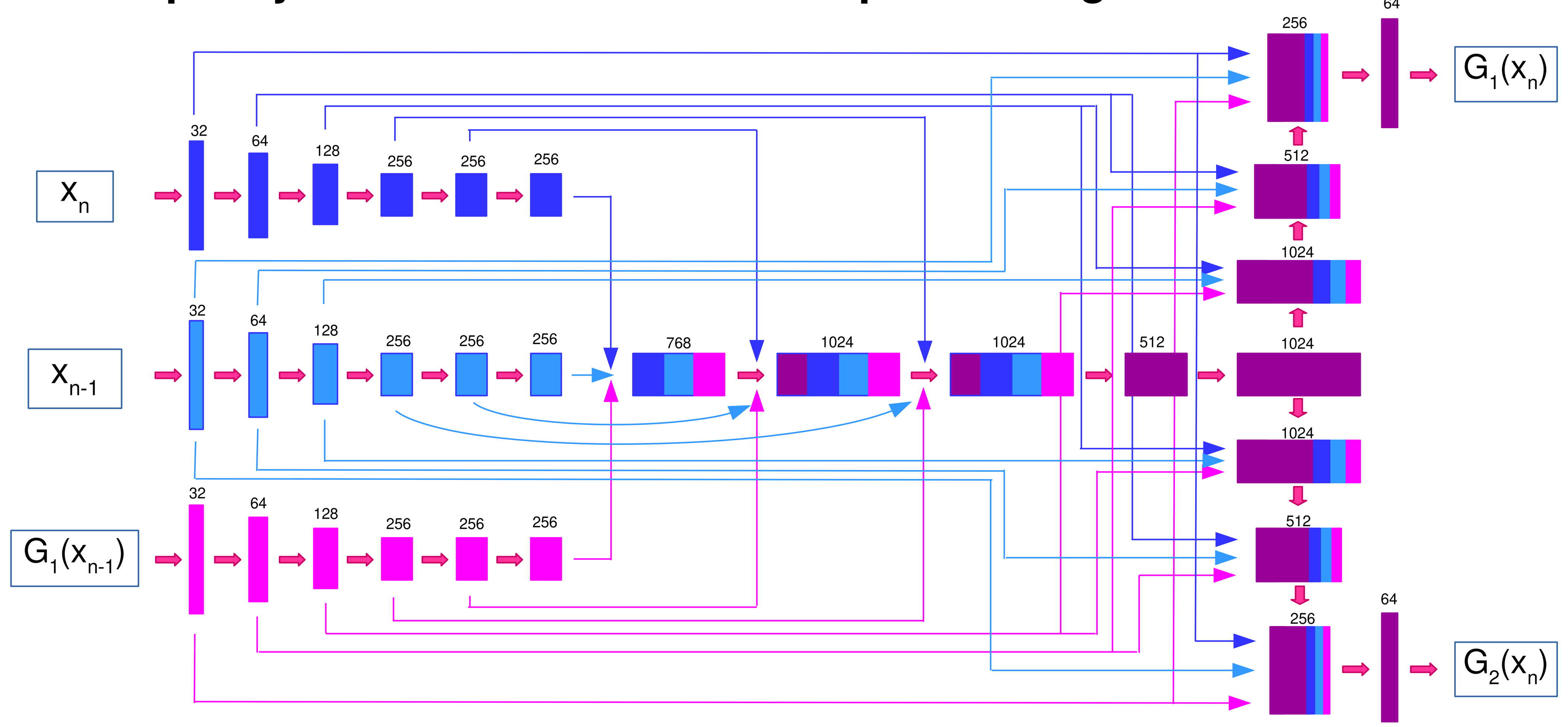
Training is performed using a robust objective function consisting of four components: depth **reconstruction** loss, **adversarial** loss, depth **smoothing** loss, **optical flow** loss, and semantic **segmentation** loss.

Proposed Approach:

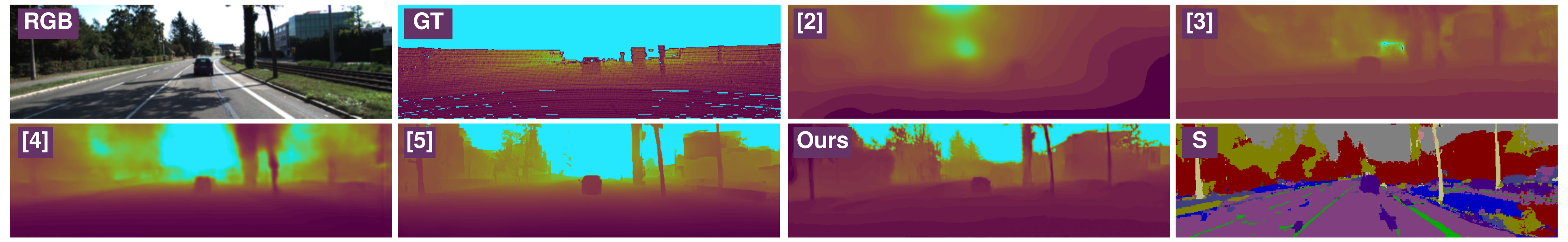
At each time step, the network takes the output generated at the previous time step as a recurrent input to preserve temporal continuity. Gradients from a pre-trained frozen optical flow network receiving output and ground truth depth images from current and previous steps also enforce temporal consistency.



A deep robust multi-stream architecture with complex skip connections, results in better high-level **contextual** and **geometric structural learning**, leading to better **quality** and **accurate** results for **depth** and **segmentation**.



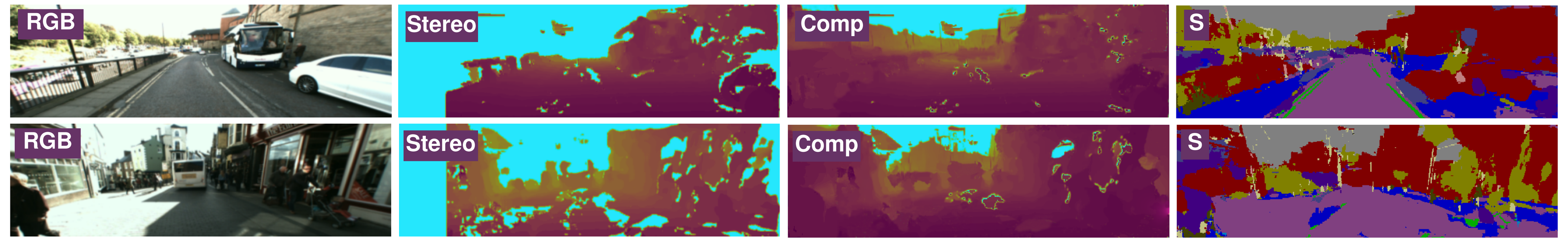
Results:



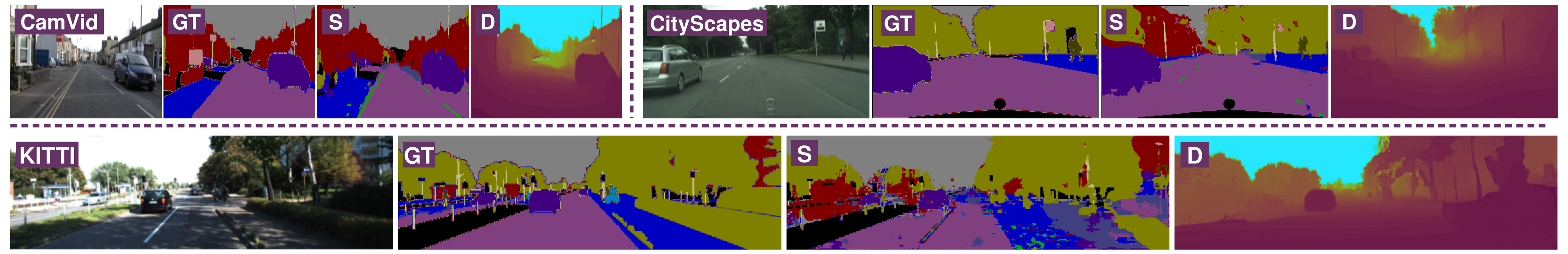
Superior **qualitative** results in **monocular depth estimation**.

Methods	Error Metrics				Accuracy Metrics		
	Abs. Rel.	Sq. Rel.	RMSE	RMSE log	$\sigma < 1.25$	$\sigma < 1.25^2$	$\sigma < 1.25^3$
Zhou et al. [2]	0.208	1.768	6.856	0.283	0.678	0.885	0.957
Godard et al. [3]	0.148	1.344	5.927	0.247	0.803	0.922	0.964
Our Results	0.193	1.438	5.887	0.234	0.836	0.930	0.958

Ours approach remains **quantitatively** competitive with the state of the art.



Plausible **depth completion** using **unseen** local **images** from Durham, UK.



Promising **semantic segmentation** results on well-established datasets.

[1] German et al., 'The SYNTHIA dataset: a large collection of synthetic images for urban scenes.' CVPR, 2016.
[2] Zhou et al., 'Unsupervised learning of depth and ego-motion from video.' CVPR, 2017.
[3] Godard et al., 'Unsupervised monocular depth estimation with left-right consistency'. CVPR, 2017.
[4] Kuznietsov, et al., 'Semi-supervised deep learning for monocular depth map prediction.' CVPR, 2017.
[5] Atapour et al., 'Real-time monocular depth estimation with domain adaptation via image style transfer.' CVPR, 2018.

Network **code and models available** here:

https://github.com/atapour/temporal-depth-segmentation