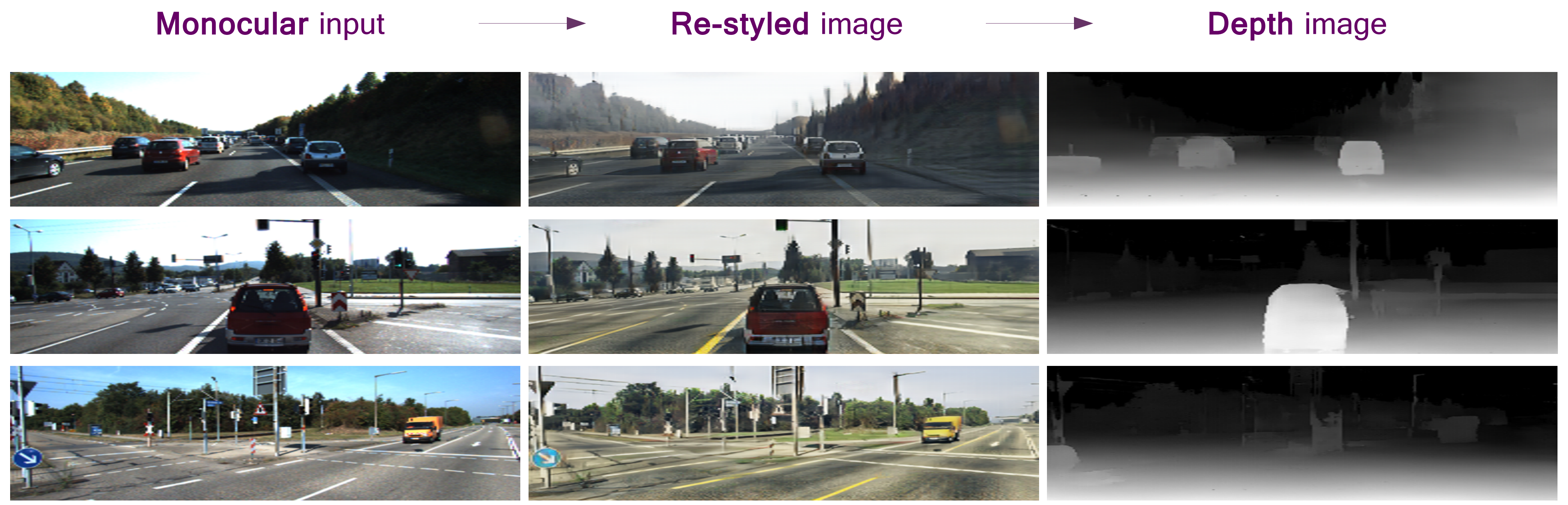
Real-Time Monocular Depth Estimation using Synthetic Data with Domain Adaptation via Image Style Transfer

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

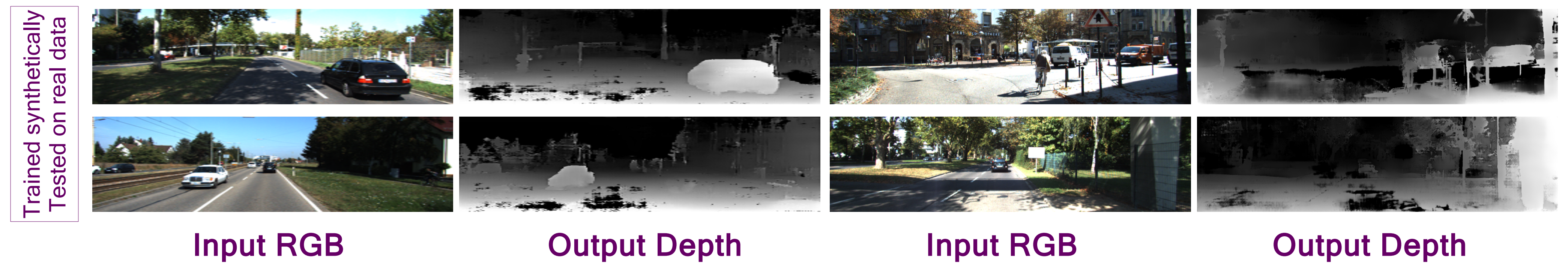


Monocular depth estimation model trained on synthetic data produces sharp and plausible depth when applied to real-world images transformed to the style of synthetic images.

Motivation:

Synthetic images captured from a **graphically-rendered virtual environment** primarily designed for gaming can be **employed to train a monocular depth estimation model**. However, this **will not generalize well to real-world images** as the supervised model easily **overfits to local features present within the training domain.**

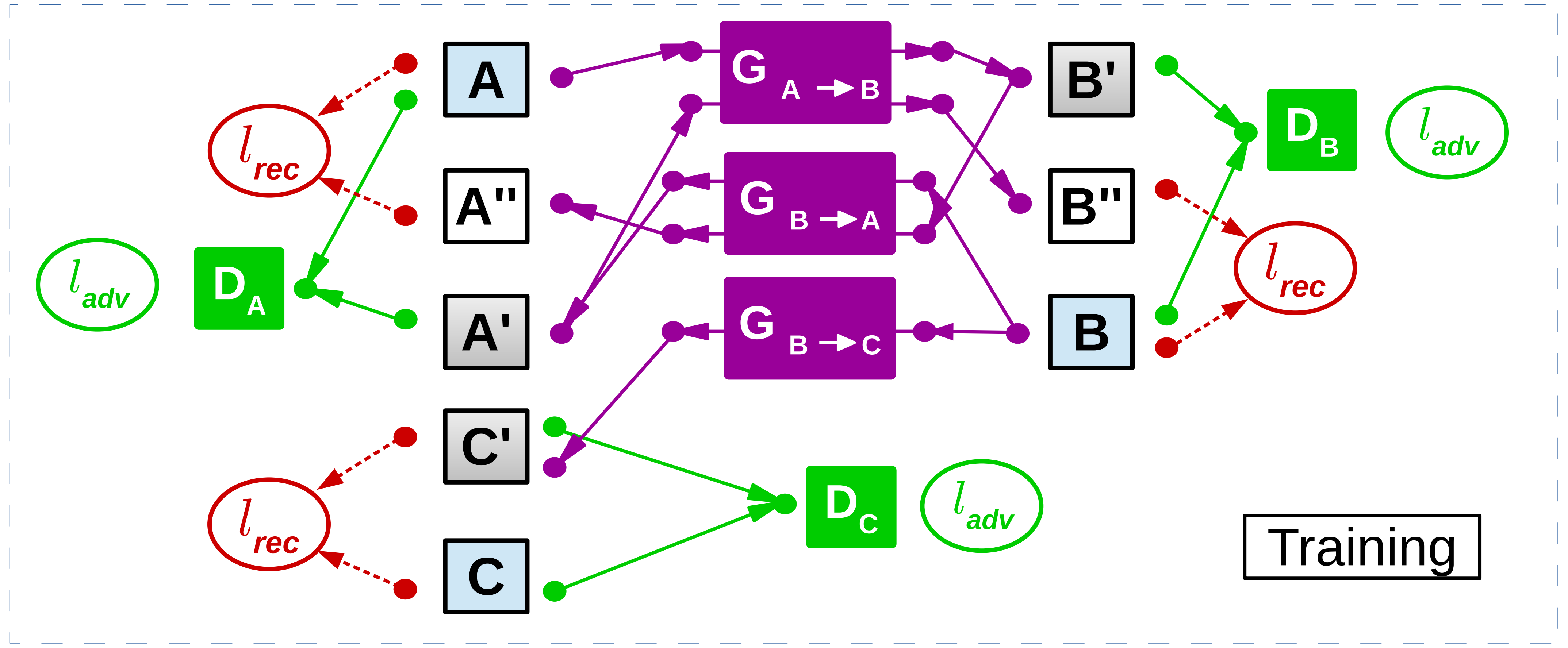
Without using domain adaptation:



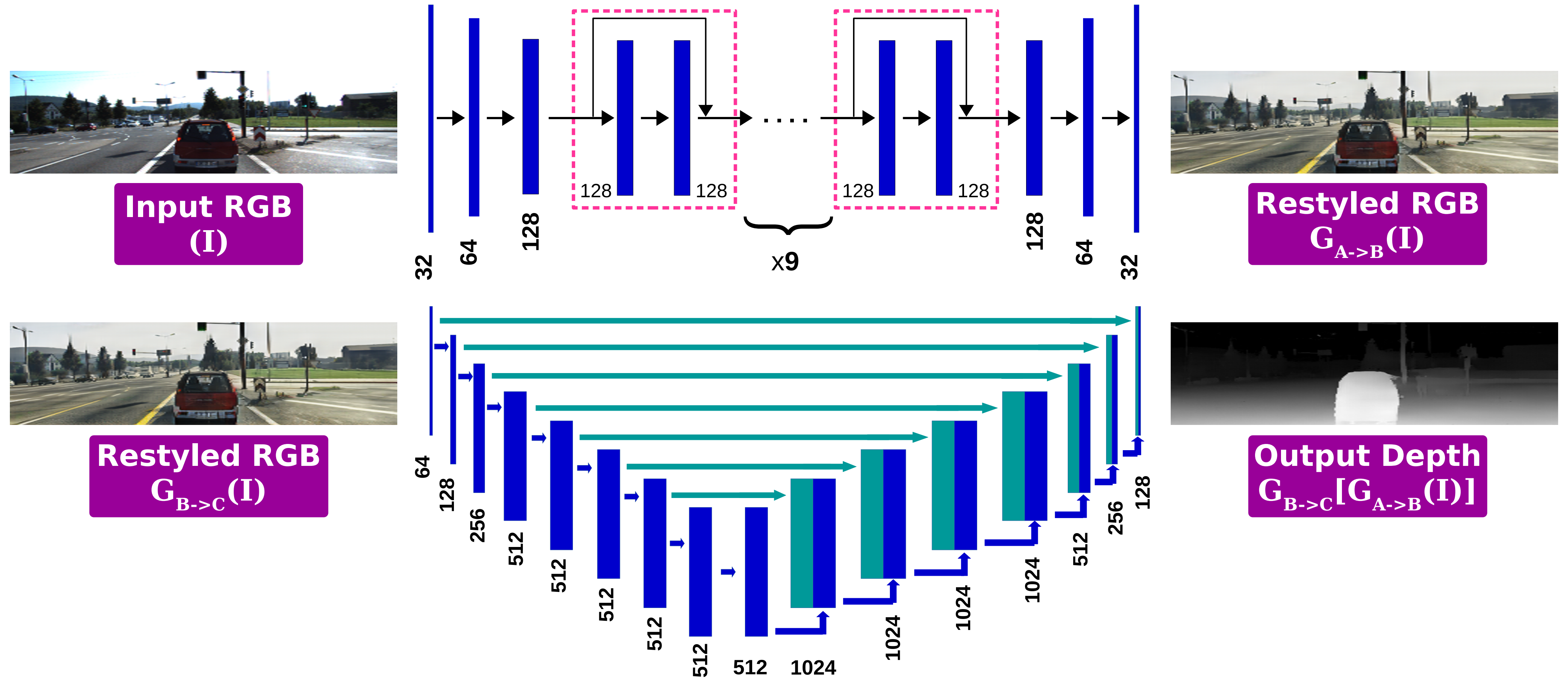
Style transfer has previously been theoretically **linked to domain adaptation** [5]. We utilize the **CycleGAN approach** presented in [4] **to re-style real-world images** to look similar **to the synthetic images** the model is originally trained on, hence **reducing the discrepancy between the two image domains** during inference.

Proposed Approach:

1) train a primary model to **estimate monocular depth based on synthetic images.** 2) use a secondary model to **transform real-world images to the synthetic style before their depth is estimated.**



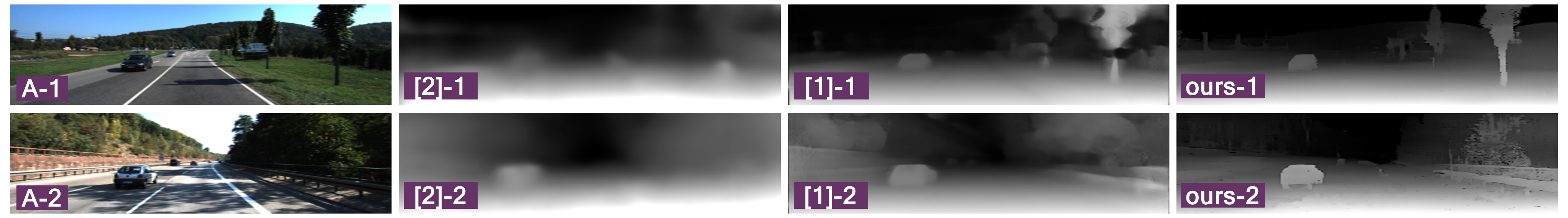
Run-time:- two forward passes required during inference – once through the **style transfer network** and once through the **depth estimation model.**


Input RGB (I)
32
64
128
128
128
128
128
x9
128
64
32
Restyled RGB $G_{A\to B}$(I)
Restyled RGB $G_{B\to C}$(I)
64
128
256
512
512
512
512
512
512
1024
1024
1024
1024
512
256
128
Output Depth $G_{B\to C}[G_{A\to B}(I)]$

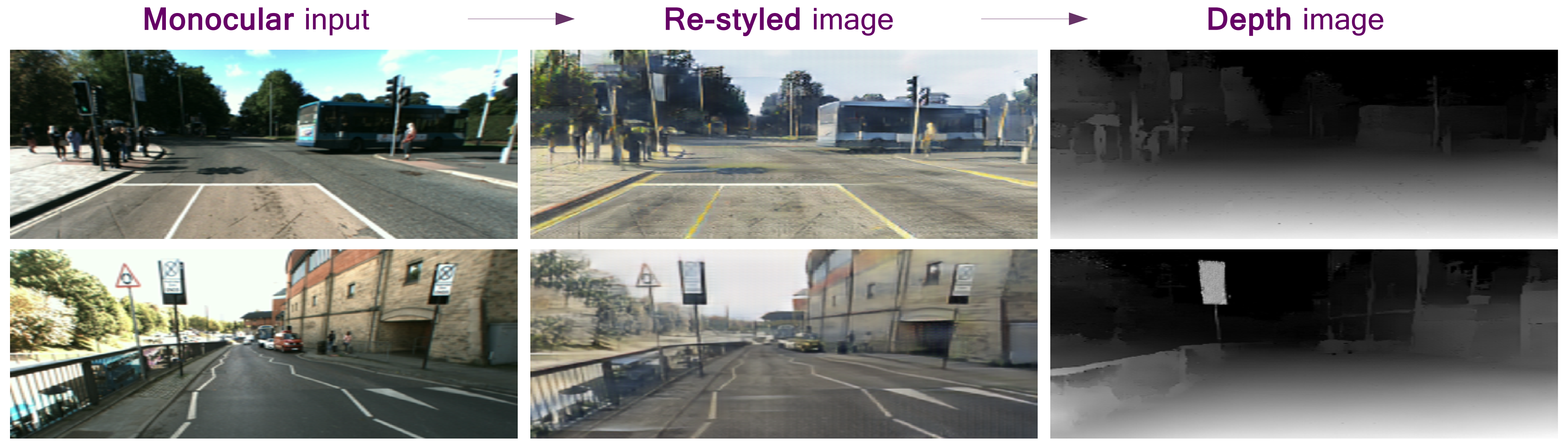
Results:

Our approach produces **superior qualitative (sharper)** and **quantitative (lower error) results** compared to the contemporary state-of-the-art.

Methods	Error Metrics				Accuracy Metrics		
	Abs. Rel.	Sq. Rel.	RMSE	RMSE log	$\sigma < 1.25$	$\sigma < 1.25^2$	$\sigma < 1.25^3$
Godard et al. [1]	0.124	1.076	5.311	0.219	0.847	0.942	0.973
Zhou et al. [2]	0.198	1.836	6.565	0.275	0.718	0.901	0.960
Ours (no adaptation)	0.498	6.533	9.382	0.609	0.712	0.823	0.883
Ours using [3]	0.154	1.338	6.470	0.296	0.874	0.962	0.981
Ours using [4]	**0.101**	**1.048**	**5.308**	**0.184**	**0.903**	**0.988**	**0.992**



Model **generalization** is **tested using unseen images** from Durham, UK.



[1] Godard et al., 'Unsupervised monocular depth estimation with left-right consistency'. CVPR, 2017.
[2] Zhou et al., 'Unsupervised learning of depth and ego-motion from video.' CVPR, 2017.
[3] Johnson et al., 'Perceptual losses for real-time style transfer and super-resolution.' ECCV, 2016.
[4] Zhu et al., 'Unpaired image-to-image translation using cycle-consistent adversarial networks.' ICCV, 2017.
[5] Li et al., 'Demystifying neural style transfer.' arXiv preprint arXiv:1701.01036, 2017.

Network inference **code and models available** here:

https://github.com/atapour/monocularDepth-Inference