Monocular Segment-wise Depth: Monocular Depth Estimation Based on a Semantic Segmentation Prior

Amir Atapour-Abarghouei and Toby P. Breckon

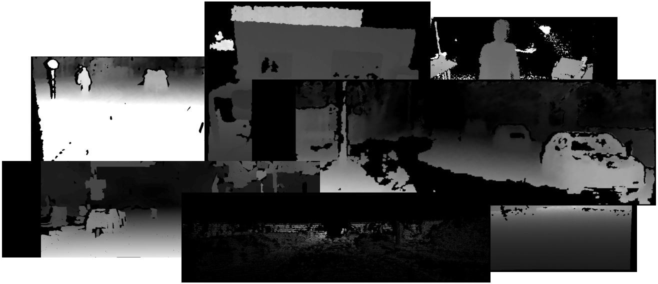
Department of Computer Science, Durham University

25 September 2019

Problem Domain

Current depth sensing technologies are flawed.

- Despite the various modern depth sensing technologies, acquired images are often noisy, corrupt and with large missing regions.

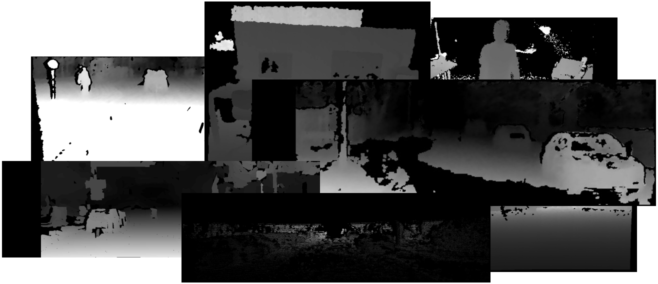


Durham University

Problem Domain

Current depth sensing technologies are flawed.

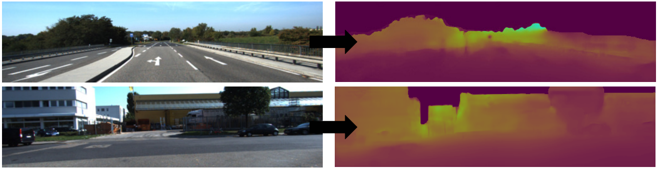
- Can we obtain the scene depth from a single RGB image by learning about content and context of the scene?



Problem Domain

Current depth sensing technologies are flawed.

- By learning about the contents and context of the scene, monocular depth estimation can lead to improved scene depth.



Proposed Approach

Proposed Approach

Scene depth from urban driving scenarios is estimated from a single RGB image by semantically understanding scene components.

Proposed Approach

Scene depth from urban driving scenarios is estimated from a single RGB image by semantically understanding scene components.

- The scene is decomposed into four object groups (segments):

Durham University

Proposed Approach

Scene depth from urban driving scenarios is estimated from a single RGB image by semantically understanding scene components.

- The scene is decomposed into four object groups (segments):
  1. small and narrow foreground objects (e.g. pedestrians, signs)
  2. flat surfaces (e.g. roads, buildings)
  3. vegetation (e.g. trees, bushes)
  4. background objects (unlabelled irrelevant components, e.g. bench)

Proposed Approach

Scene depth from urban driving scenarios is estimated from a single RGB image by semantically understanding scene components.

- The scene is decomposed into four object groups (segments):
  1. small and narrow foreground objects (e.g. pedestrians, signs)
  2. flat surfaces (e.g. roads, buildings)
  3. vegetation (e.g. trees, bushes)
  4. background objects (unlabelled irrelevant components, e.g. bench)
- Each group is segmented using a separate segmentation network.

Durham University

Proposed Approach

Scene depth from urban driving scenarios is estimated from a single RGB image by semantically understanding scene components.

- The scene is decomposed into four object groups (segments):
  1. small and narrow foreground objects (e.g. pedestrians, signs)
  2. flat surfaces (e.g. roads, buildings)
  3. vegetation (e.g. trees, bushes)
  4. background objects (unlabelled irrelevant components, e.g. bench)
- Each group is segmented using a separate segmentation network.
- Segmented object groups are used to choose sections of the RGB image passed as inputs to depth generators.

Durham University

Proposed Approach

Scene depth from urban driving scenarios is estimated from a single RGB image by semantically understanding scene components.

- The scene is decomposed into four object groups (segments):
  1. small and narrow foreground objects (e.g. pedestrians, signs)
  2. flat surfaces (e.g. roads, buildings)
  3. vegetation (e.g. trees, bushes)
  4. background objects (unlabelled irrelevant components, e.g. bench)
- Each group is segmented using a separate segmentation network.
- Segmented object groups are used to choose sections of the RGB image passed as inputs to depth generators.
- Generated segment-wise depth images are fused via summation.

Durham University

Proposed Approach

Scene depth from urban driving scenarios is estimated from a single RGB image by semantically understanding scene components.

- The scene is decomposed into four object groups (segments):
  1. small and narrow foreground objects (e.g. pedestrians, signs)
  2. flat surfaces (e.g. roads, buildings)
  3. vegetation (e.g. trees, bushes)
  4. background objects (unlabelled irrelevant components, e.g. bench)
- Each group is segmented using a separate segmentation network.
- Segmented object groups are used to choose sections of the RGB image passed as inputs to depth generators.
- Generated segment-wise depth images are fused via summation.
- The overall consistency is controlled via adversarial training.

Durham University

Proposed Approach

Scene depth from urban driving scenarios is estimated from a single RGB image by semantically understanding scene components.

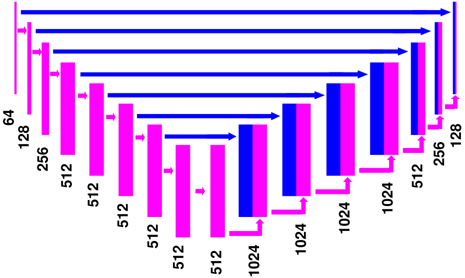
- The scene is decomposed into four object groups (segments):
  1. small and narrow foreground objects (e.g. pedestrians, signs)
  2. flat surfaces (e.g. roads, buildings)
  3. vegetation (e.g. trees, bushes)
  4. background objects (unlabelled irrelevant components, e.g. bench)
- Each group is segmented using a separate segmentation network.
- Segmented object groups are used to choose sections of the RGB image passed as inputs to depth generators.
- Generated segment-wise depth images are fused via summation.
- The overall consistency is controlled via adversarial training.
- Synthetic dataset with depth and segmentation labels.

[Ros et al., 2016]

Proposed Approach

Networks

- Our approach comprises three types of networks, with similar architectural elements for consistency.



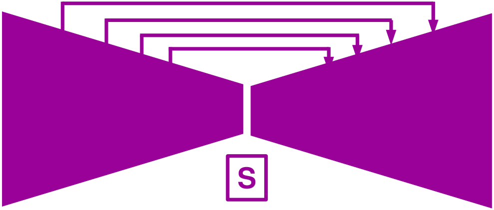
Network Architecture

Durham University

Proposed Approach

Networks

- *Segmentation Networks* - One for each object group, with an RGB input and a binary output denoting where segment pixels exist within the scene.



Segmentation Networks

Proposed Approach

Networks

- *Segmentation Networks* - Each trained using a binary cross-entropy loss:

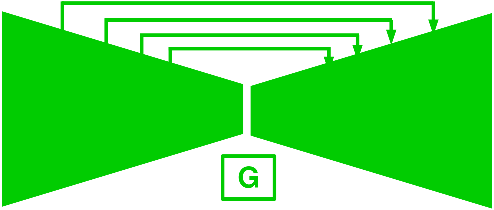
$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log(p_i) + (1 - y_i)\log(1 - p_i))$$

Durham University

Proposed Approach

Networks

- *Generator Networks* - One for each object group, with a region of the RGB selected based on the segmentation output is used as the input and depth is produced.



Generator Networks

Proposed Approach

Networks

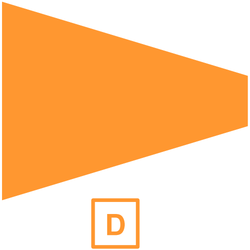
- *Generator Networks* - The networks are trained by minimising the Euclidean distance between the output and the ground truth depth.

$$\mathcal{L}_{\text{rec}} = ||G_1(S_1(x) \times x) - (S_1(x) \times y)||_1$$

Proposed Approach

Networks

- *Discriminator Network* - One in the overall model, responsible for removing artefacts (stitching, bleeding, blurring, etc.) from the final output.

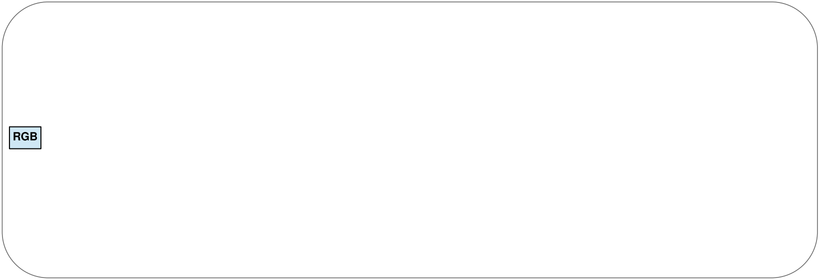


Discriminator Network

Proposed Approach

Training Procedure

- An RGB image is used as the input. The overall model will generate the complete scene depth based on this RGB image.

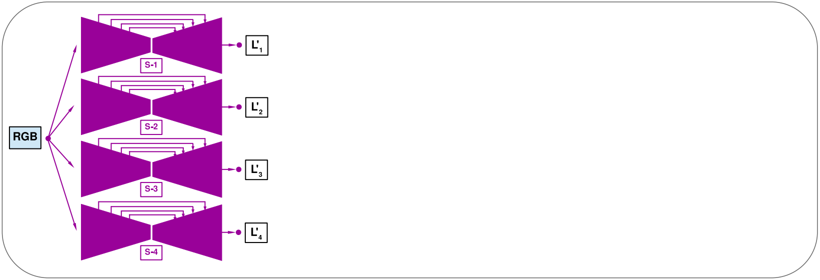


Durham University

Proposed Approach

Training Procedure

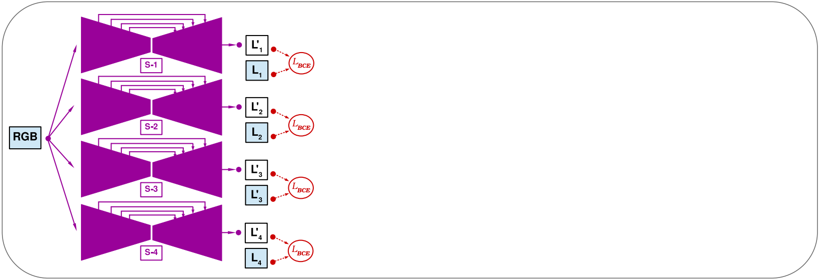
- The input is passed through the *Segmentation* networks, which identify the object groups (segments) within the scene.



Proposed Approach

Training Procedure

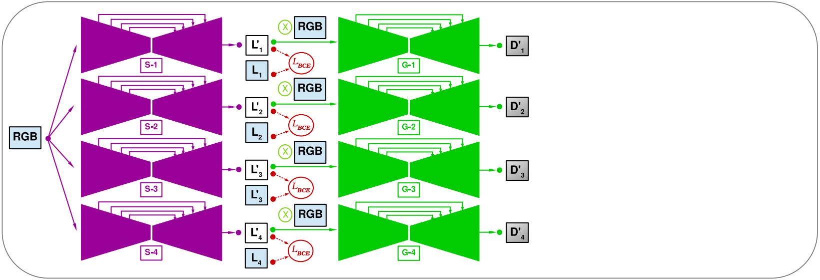
- Gradients are calculated with respect to a binary cross-entropy loss to train the *Segmentation* networks.



Proposed Approach

Training Procedure

- The outputs are multiplied by the RGB input and passed through the *Generator* networks to produce the depth for each segment.

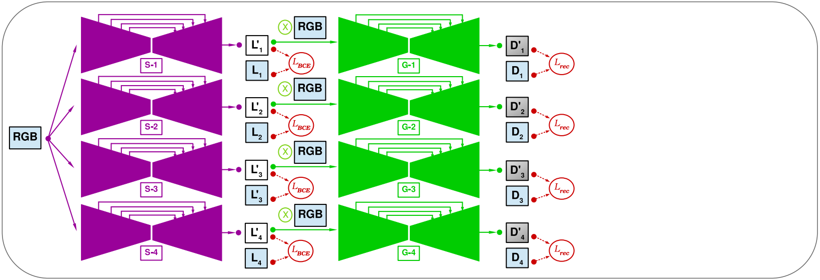


Durham University

Proposed Approach

Training Procedure

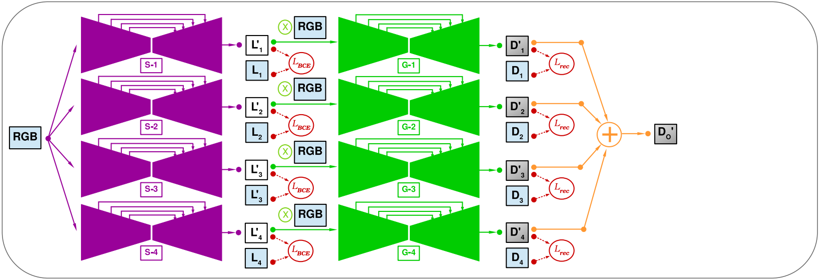
- Gradients are calculated with respect to an L_1 loss to train the *Generator* networks.



Proposed Approach

Training Procedure

- The outputs of the *Generator* networks are simply summed up to produce the overall scene depth.

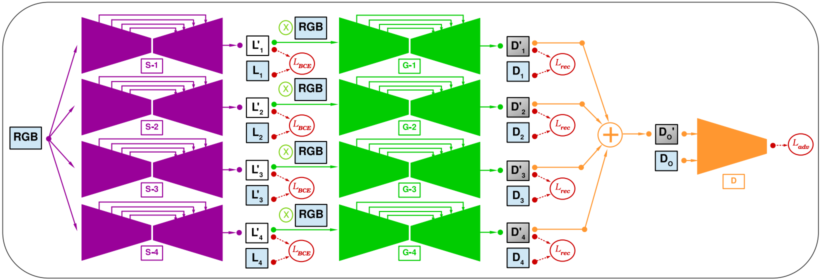


Durham University

Proposed Approach

Training Procedure

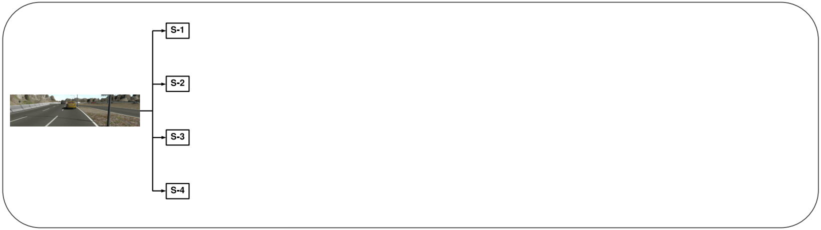
- The result is passed through the *Discriminator*, which ensures no artefacts exist as a result of the summation operation.



Proposed Approach

Inference Process

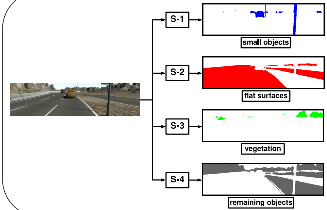
- During inference, the RGB image if first passed through the *Segmentation* networks.



Proposed Approach

Inference Process

- The *Segmentation* networks produce class labels for their corresponding object groups.

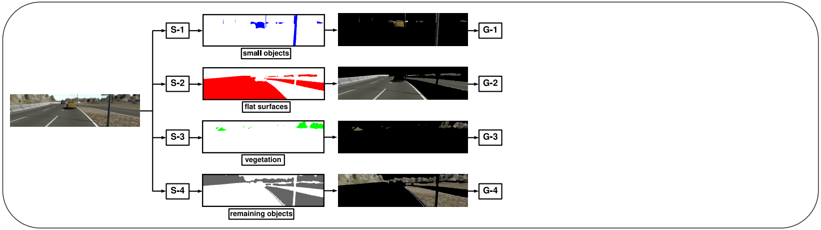


Durham University

Proposed Approach

Inference Process

- The generated class labels are multiplied by the RGB image to create the depth generation inputs.

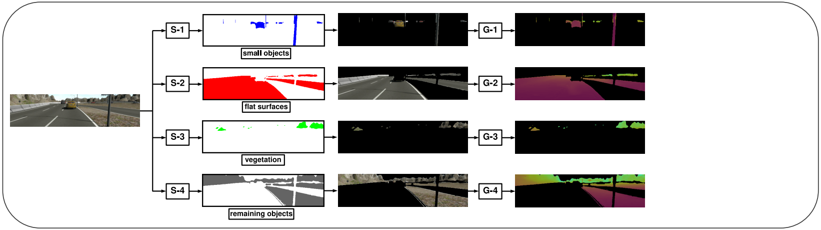


Durham University

Proposed Approach

Inference Process

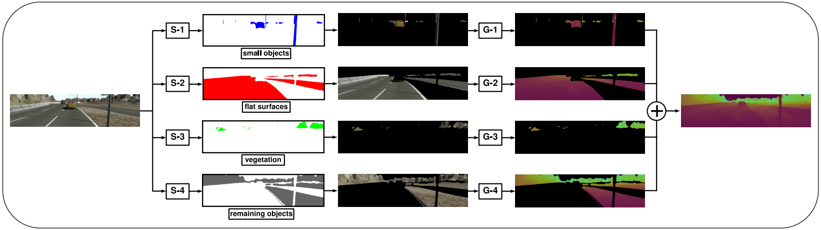
- The *Generator* networks produce partial depth outputs for their corresponding object groups.



Proposed Approach

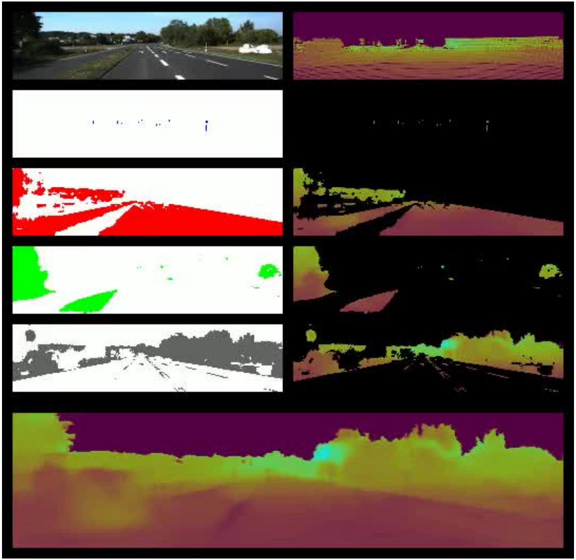
Inference Process

- The generated depth outputs are summed to produce the final depth image.



Proposed Approach

Video: Scene from the KITTI Dataset



Experimental Results

Experimental Results

Ablation Studies

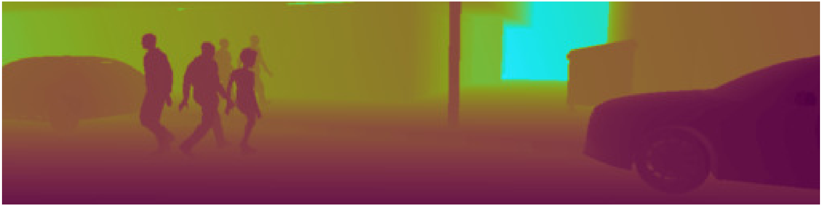
- Through ablation studies, we demonstrate the importance of the components of our complex approach.

RGB Image

Experimental Results

Ablation Studies

- We train three separate models to evaluate the necessity of the components of the approach.



Ground Truth Depth

Durham University

Experimental Results

Ablation Studies

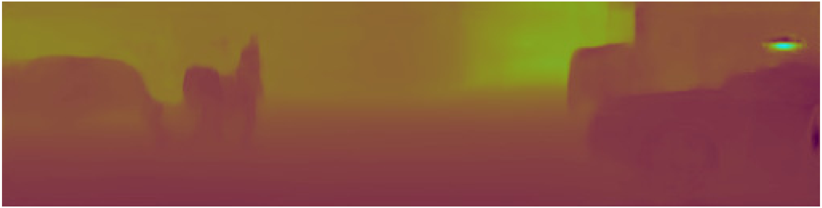
- We train a *direct* model to carry out global depth estimation without any segmentation.

Direct Model

Experimental Results

Ablation Studies

- In our *implicit* model, we train the depth generator networks to implicitly perform segment-wise depth estimation.



Implicit Model

Experimental Results

Ablation Studies

- The *implicit* model produces promising results and can offer faster run-times (61 *ms*).

Implicit Model

Experimental Results

Ablation Studies

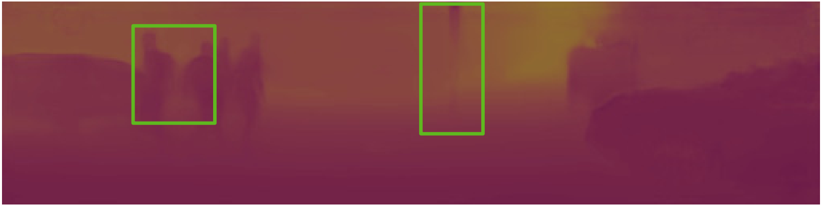
- The full *explicit* approach is more complex and inefficient (112 *ms*) than the *implicit* model.

Explicit Model

Experimental Results

Ablation Studies

- The *explicit* model provides the best results especially for small and narrow objects in the scene.



Explicit Model

Experimental Results

Comparisons against other approaches

- Our approach is not trained on real-world data.

RGB Image

Experimental Results

Comparisons against other approaches

- No domain adaptation is used in our model.

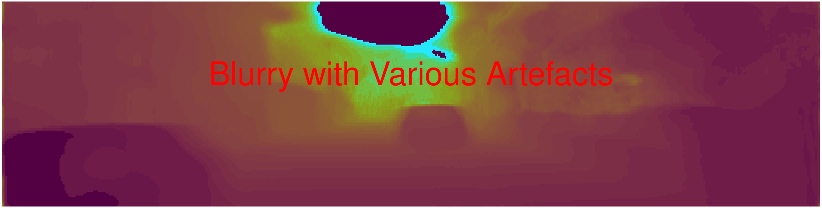
[Zhou et al., CVPR, 2017]

Durham University

Experimental Results

Comparisons against other approaches

- No temporal continuity is enforced in our approach.



[Godard et al., CVPR, 2017]

Durham University

Experimental Results

Comparisons against other approaches

- Our approach visually outperforms comparators.

Our Result

Experimental Results

Comparisons against other approaches

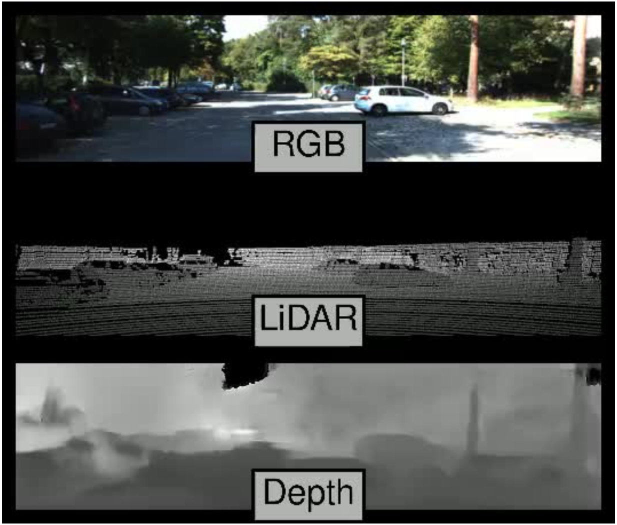
- Numerical evaluation demonstrates the efficacy of our approach.

Method		Error				Accuracy
		Abs. Rel.	Sq. Rel.	RMSE	RMSE log	$\sigma < 1.25^3$
Eigen et al.	[NeurIPS, 2014]	0.203	1.548	6.307	0.308	0.958
Liu et al.	[TPAMI, 2016]	0.202	1.614	6.523	0.308	0.965
Zhou et al.	[CVPR, 2017]	0.208	1.768	6.856	0.283	0.957
Godard et al.	[CVPR, 2017]	**0.148**	1.344	5.927	**0.247**	0.964
Our Approach		0.168	**1.338**	**5.702**	0.252	**0.968**

Durham University

Experimental Results

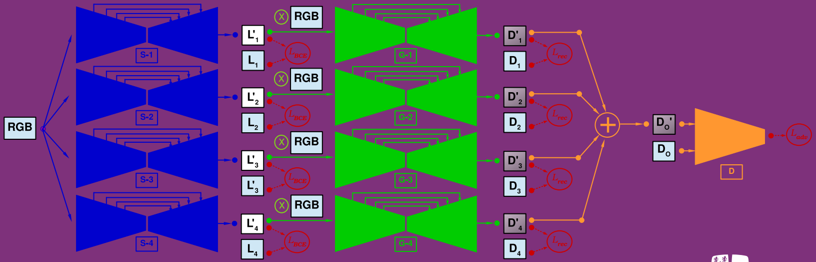
Video: Scene from the KITTI Dataset



Amir Atapour-Abarghouei

and Toby P. Breckon

amir.atapour-abarghouei@durham.ac.uk



Durham University