UNIVERSITÀ DI PARMA

DEPARTMENT OF ENGINEERING AND ARCHITECTURE
COMPUTER ENGINEERING

3D Perception & Learning Based Data Fusion

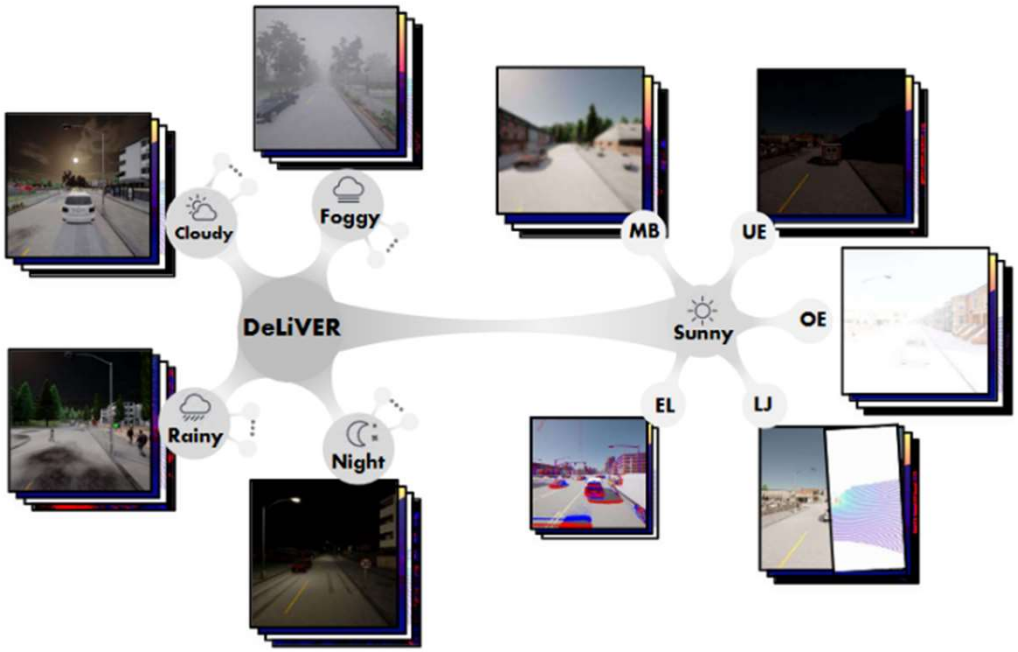
Multi-Modal Semantic Segmentation with Deep Learning based Data-Fusion

Giuseppe Trimigno

DATASET

- **DeLiVER** is a really complex dataset, with 25 classes.
- Although the full dataset consists of **47310** samples for each sensor modality, only **7885** samples are publicly available.
- Split constists of 5988 training and 1897 testing samples.

- Contains images coming from **four different sensors**: RGB, Depth and Event cameras, plus LiDAR points projected onto 2D plane.
- Contains **four severe weather conditions** (cloudy, foggy, rainy and night), as well as **five sensor failures** (motion blur, over-exposure, under-exposure, LiDAR jitter and event low-resolution).



MODEL

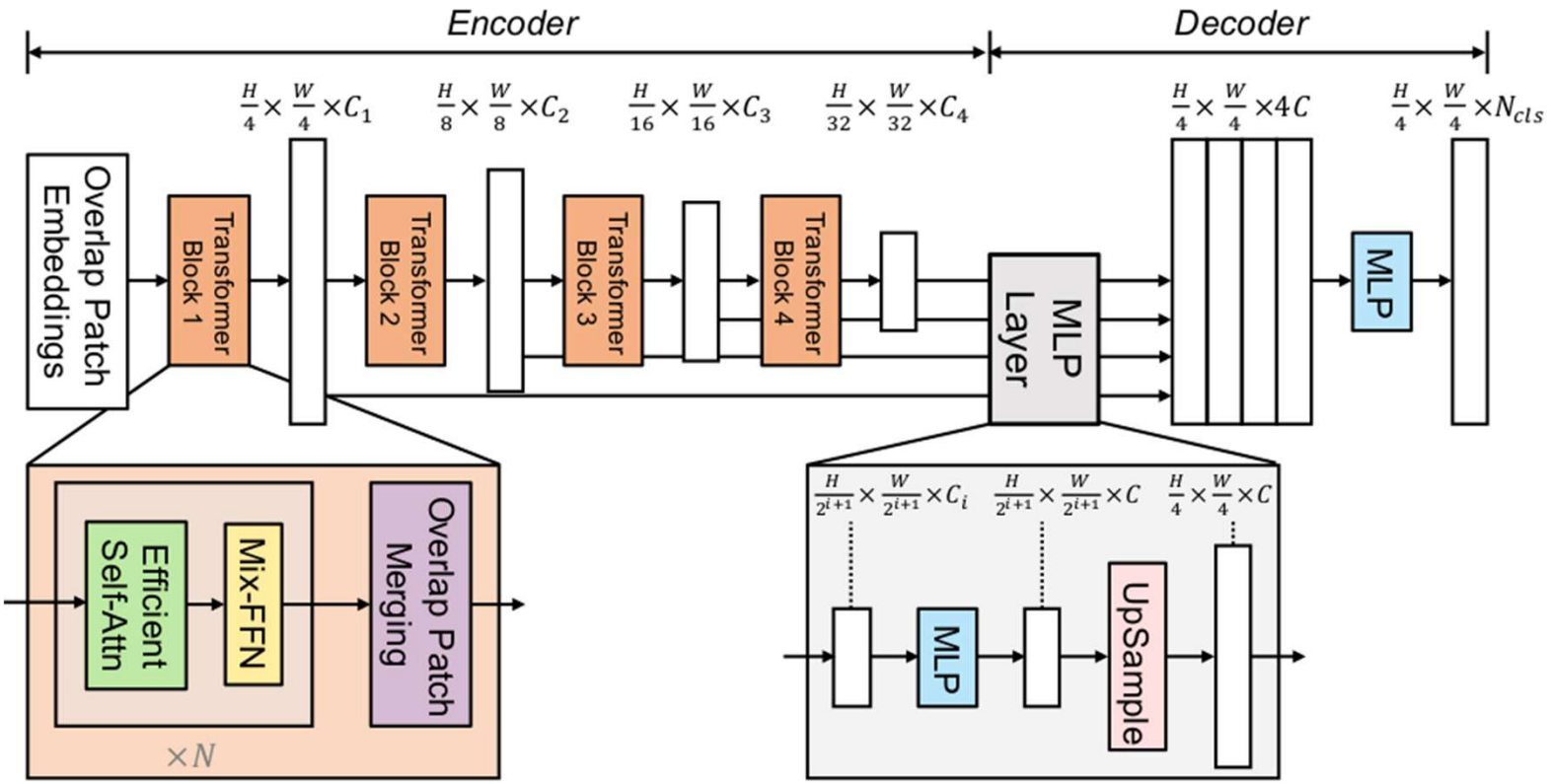
SegFormer

- Powerful semantic segmentation framework which unifies Transformers with lightweight MLP decoder.

It has two main appealing features:

- A novel hierarchically structured Transformer encoder which outputs multiscale features. → Does not need positional encoding, helpful when testing resolutions are different from training ones.
- Avoids complex decoders, using a simple MLP decoder, which aggregates information from different layers. → Allows to combine both local and global attention.

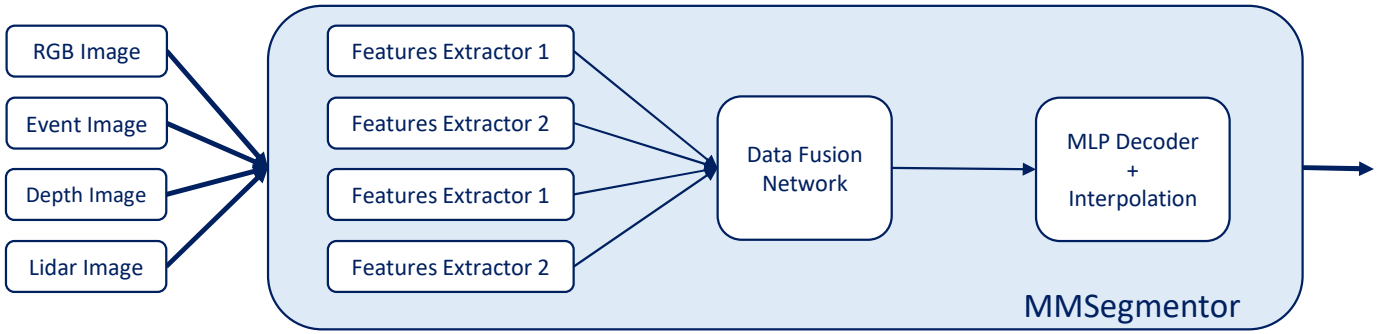
MODEL


Encoder
Decoder
$\frac{H}{4} \times \frac{W}{4} \times C_1$
$\frac{H}{8} \times \frac{W}{8} \times C_2$
$\frac{H}{16} \times \frac{W}{16} \times C_3$
$\frac{H}{32} \times \frac{W}{32} \times C_4$
$\frac{H}{4} \times \frac{W}{4} \times 4C$
$\frac{H}{4} \times \frac{W}{4} \times N_{cls}$
Overlap Patch Embeddings
Transformer Block 1
Transformer Block 2
Transformer Block 3
Transformer Block 4
MLP Layer
MLP
Efficient Self-Attn
Mix-FFN
Overlap Patch Merging
$\times N$
$\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times C_i$
$\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times C$
$\frac{H}{4} \times \frac{W}{4} \times C$
MLP
UpSample

MODEL

Problem

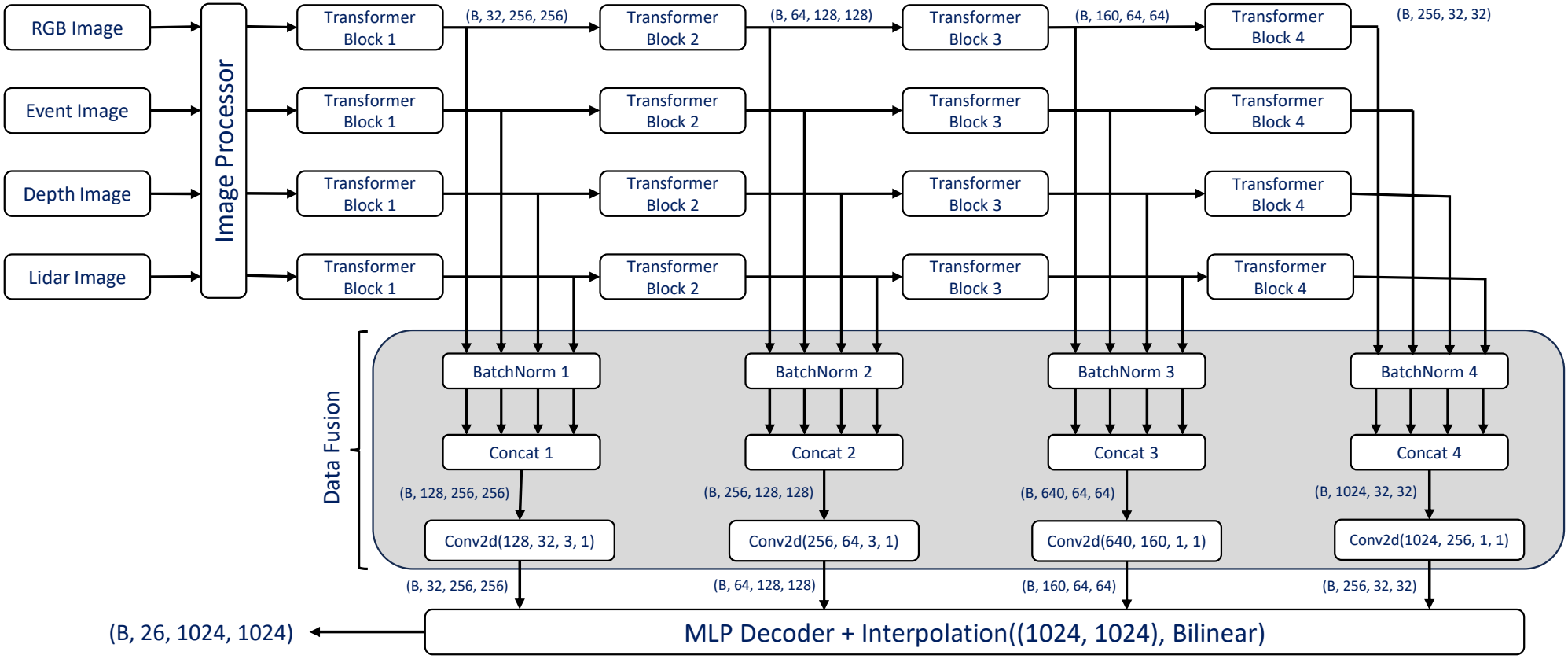
- SegFormer does not support multimodality.
- We need to overload its architeture.



MMSegmentor

- The basic idea is to insert a deep learning data fusion architecture between the encoder and decoder.
- The final model is composed by four components:
  1. **Feature Extractors**: a total of four encoders, one for each modality;
  2. **Data-Fusion network**: four parallel combinations of batch normalization and 2D convolutional layers which acts on hidden states to fuse extracted features;
  3. **MLP Decoder:** takes fused hidden states as input and produces logits as output;
  4. **Interpolator**: a simple bilinear interpolation layer which allows to upsample logits back to labels' required shape.

MODEL

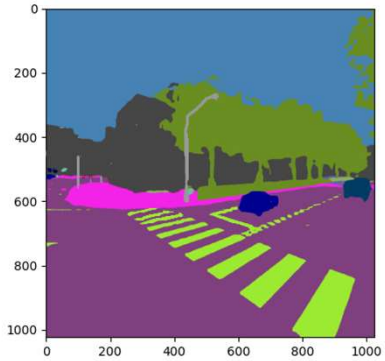

RGB Image
Event Image
Depth Image
Lidar Image
Image Processor
Transformer Block 1
Transformer Block 2
Transformer Block 3
Transformer Block 4
(B, 32, 256, 256)
(B, 64, 128, 128)
(B, 160, 64, 64)
(B, 256, 32, 32)
Data Fusion
BatchNorm 1
BatchNorm 2
BatchNorm 3
BatchNorm 4
Concat 1
Concat 2
Concat 3
Concat 4
(B, 128, 256, 256)
(B, 256, 128, 128)
(B, 640, 64, 64)
(B, 1024, 32, 32)
Conv2d(128, 32, 3, 1)
Conv2d(256, 64, 3, 1)
Conv2d(640, 160, 1, 1)
Conv2d(1024, 256, 1, 1)
(B, 32, 256, 256)
(B, 64, 128, 128)
(B, 160, 64, 64)
(B, 256, 32, 32)
MLP Decoder + Interpolation((1024, 1024), Bilinear)
(B, 26, 1024, 1024)

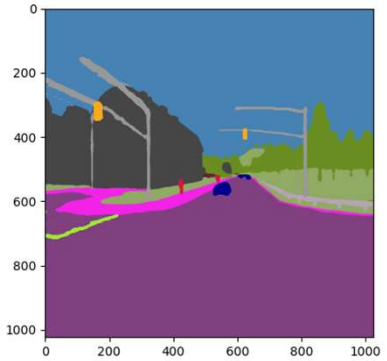
RESULTS

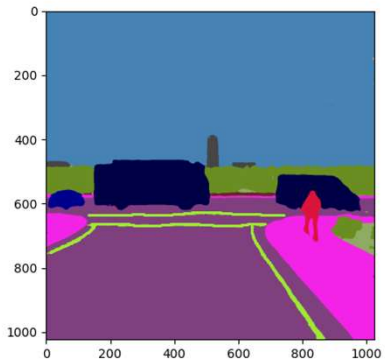
Model	mIoU (macro)	mIoU (weighted)	Accuracy (macro)	Accuracy (weighted)
DeepLabV3+ResNet50 NoAux PCA (Early-Fusion)	0.328	0.811	0.391	0.871
DeepLabV3+ResNet101 NoAux PCA (Early-Fusion)	0.34	0.85	0.403	0.905
MMSegmentor-B0	**0.429**	**0.907**	**0.49**	**0.917**
MMSegmentor-B3	**0.442**	**0.914**	**0.522**	**0.926**

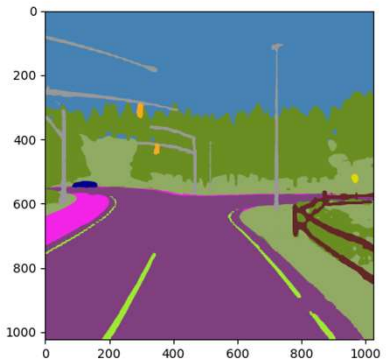
Model	Road		Sky		Building		Vegetation		RoadLine		Cars		SideWalk		Pedestrian		Pole		Truck	
	mIoU	Acc	mIoU	Acc	mIoU	Acc	mIoU	Acc	mIoU	Acc	mIoU	Acc	mIoU	Acc	mIoU	Acc	mIoU	Acc	mIoU	Acc
DeepLabV3 ResNet101 NoAux RGB-Only	0.93	0.961	0.867	0.944	0.601	0.725	0.556	0.66	0.628	0.68	0.538	0.694	0.522	0.637	0.3	0.411	0.25	0.323	0.25	0.3
DeepLabV3 ResNet50 NoAux PCA Early-Fusion	0.8997	0.9362	0.8992	0.965	0.6114	0.7297	0.4972	0.551	0.5955	0.6864	0.5424	0.616	0.5202	0.6757	0.3241	0.4152	0.2929	0.4485	0.2113	0.2754
DeepLabV3 ResNet101 NoAux PCA Early-Fusion	0.9389	0.9704	0.9259	0.9668	0.6386	0.7364	0.6159	0.7083	0.6097	0.6801	0.5688	0.7277	0.5518	0.7099	0.3434	0.4614	0.3018	0.4239	0.2431	0.2851
MMSegmentor-B0	**0.9375**	**0.9651**	**0.9488**	**0.9826**	**0.6995**	**0.7797**	**0.6704**	**0.7736**	**0.5958**	**0.661**	**0.5707**	**0.737**	**0.5322**	**0.6664**	**0.4101**	**0.5134**	**0.2918**	**0.419**	**0.2279**	**0.2684**
MMSegmento-B3	**0.9556**	**0.975**	**0.9509**	**0.9833**	**0.7289**	**0.801**	**0.6979**	**0.7872**	**0.6012**	**0.672**	**0.5708**	**0.7377**	**0.5512**	**0.7003**	**0.4308**	**0.5192**	**0.3027**	**0.448**	**0.2339**	**0.271**

RESULTS

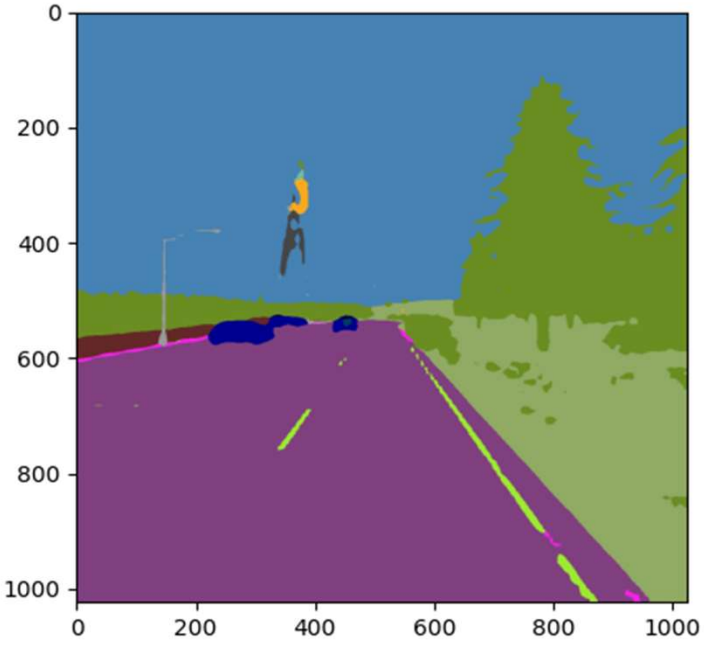

0
200
400
600
800
1000
0
200
400
600
800
1000

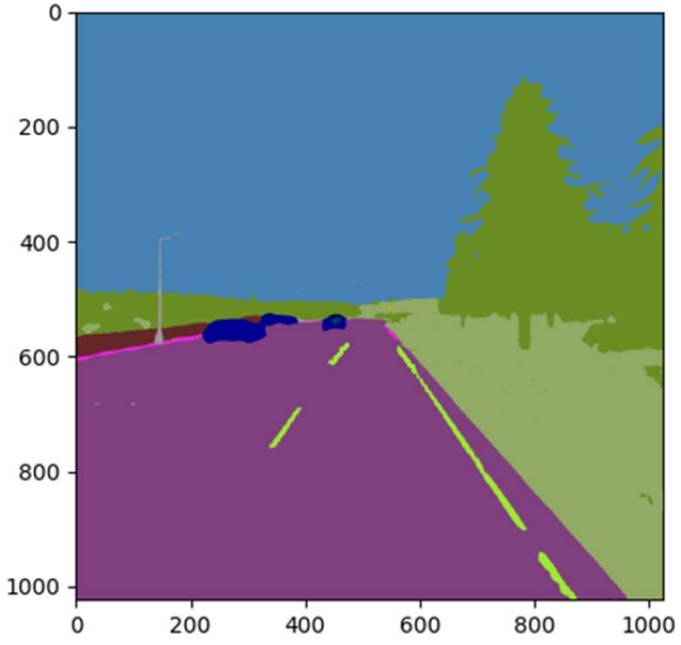

0
200
400
600
800
1000
0
200
400
600
800
1000


0
200
400
600
800
1000
0
200
400
600
800
1000


0
200
400
600
800
1000
0
200
400
600
800
1000

RESULTS


0
200
400
600
800
1000
0
200
400
600
800
1000


0
200
400
600
800
1000
0
200
400
600
800
1000

QUESTIONS

Thanks for the attention!