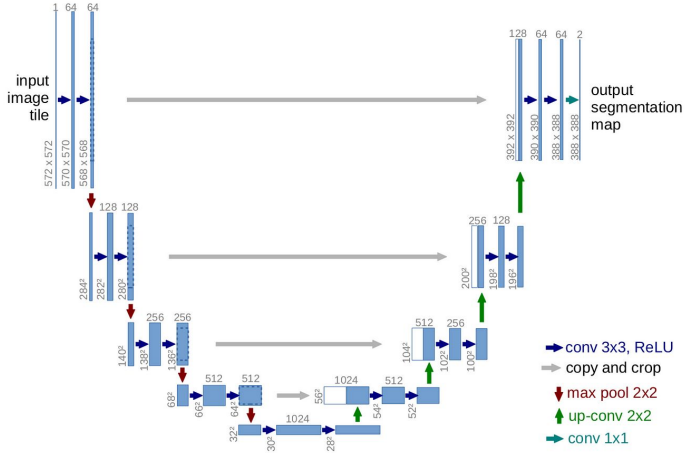
Dataset - CamVid

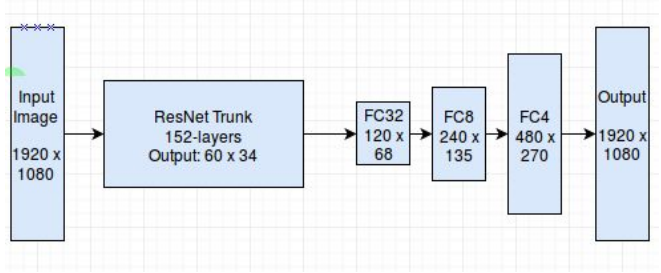
- CamVid (Cambridge-Driving Labeled Video Database)
- provides ground truth labels that associate each pixel with one of 32 semantic classes
- split into training (367), validation (101) and test (233)

Models & Implementation

- Preprocessing: Normalize & Resize to 512x512
- UNet variations (most common)
  - different model sizes/ depths
  - ReLU/ LeakyReLU, with/ without DropOut, BS: 32, 150
- Fine-tuning pretrained FCN_ResNet_50
  - default and with augmentation (RandomHorizontalFlip)
  - BS: 4, 80 Epochs
- Hyperparameters
  - AdamW, CrossEntropyLoss



https://towardsdatascience.com/unet-line-by-line-explanation-9b191c76baf5



https://www.researchgate.net/figure/ResNet-FCN-architecture_fig1_314071170

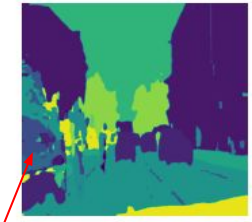
Evaluation – Metrics

	UNet	small UNet	XL UNet	XL UNet + Dropout	XXL UNet	XXL UNet + Dropout	FCNResNet50 default	FCNResNet50 w/ flip
Train Loss	0.131	***0.079***	1.135	0.437	0.848	0.340	0.121	0.099
Val Loss	0.451	0.692	1.087	0.494	0.833	0.403	***0.384***	0.440
Val mIoU	0.528	0.499	0.357	0.480	0.392	***0.570***	0.285	0.289

Loss = CrossEntropyLoss
mIoU = mean Intersection over Union/ Jaccard Index

Evaluation – Visualization

FCN ResNet 50

Input	Label	Pred	IoU
			0.388
			0.355
			0.195

Best UNet

Lessons Learned & Outlook

Learned

- Applying augmentations needs to be done carefully (Flip bug) & test more before long runs
- Add more/ better checkpoints, early stopping
- Putting pieces together: Dataset, Modelling, Tuning & Evaluating

Outlook/ Future Work

- Mix of CNN and Transformer (Yuan, F., et. al.)
  - CNN encoder and Transformer encoder in parallel.
  - Sent to feature fusion.
  - Decoded with transformer decoder
  - Feature fusion
  - Catch smaller areas/segments