Semantic Segmentation on Cityscapes
UNet–ResNet50 with 3-Fold Cross-Validation

Dominik Drabik THU ID: 2025403033

June 7, 2025

1 Introduction

Implemented a UNet with a ResNet-50 encoder for semantic segmentation on the Cityscapes dataset. Performance is measured by mean Intersection over Union (mIoU) on both 512×1024 and 1024×2048 resolutions, using 3-fold cross-validation.

1.1 Architecture

The segmentation network follows a U-Net design with a ResNet-50 backbone:

- **Encoder**: A pretrained ResNet-50 truncated after its fourth residual stage.
  - Initial conv–BN–ReLU produces 64 channels at 1/2 resolution, followed by max-pooling.
  - Residual stages produce feature maps of size 1/4, 1/8, 1/16 and 1/32, with 256, 512, 1024 and 2048 channels respectively.
- **Bottleneck**: A dual ConvBlock on the 2048-channel, 1/32-scale tensor. Each ConvBlock comprises two consecutive 3×3 convolutions (Conv–BN–ReLU) and an optional Dropout2d layer for regularization.
- **Decoder**: Four upsampling stages mirror the encoder's downsampling.
  1. A 2× transposed convolution halves the channel count (e.g. 2048→1024) and doubles spatial resolution.
  2. Bilinear interpolation enforces exact alignment with the corresponding encoder output.
  3. Concatenation with the encoder feature map provides skip connections.
  4. A ConvBlock (two 3×3 Conv–BN–ReLU) fuses high- and low-level features.

  This yields decoder outputs at 1/16, 1/8, 1/4 and 1/2 scales, with 1024, 512, 256 and 64 channels respectively.
- **Output Head**: A final 2× transposed convolution raises the 64-channel map to full resolution with 32 channels, followed by a 1×1 convolution to produce 19 class logits. A last bilinear upsample ensures the output exactly matches the input image size.

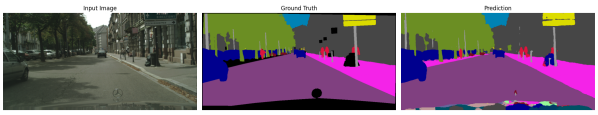

Figure 1: Prediction Fold 1

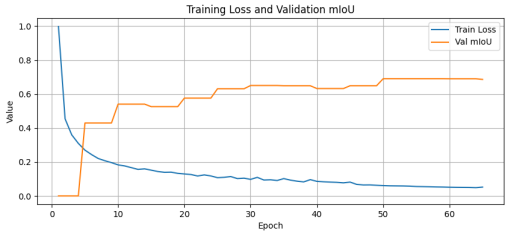

Figure 2: Training history Fold 1

2 Experimental Setup

- **Data:** Cityscapes "fine" split; 3-fold on the training set.
- **Augmentations:**
  - Train: random scale (±20%), pad, resize to 512×1024, horizontal flip, color jitter, normalize.
  - Val: resize to 512×1024, normalize.
- **Training:** 65 epochs per fold, batch size 2, Adam lr=1e-4, ReduceLROnPlateau (factor=0.5, patience=2).
- **Loss:** CrossEntropy (ignore_index=255).
- **Metric:** Multiclass Jaccard Index (19 classes).

3 Results

Best mIoU: 0.6719, Mean mIoU: 0.6552.
Table with the results: 1

3.1 Fold 1:

- Prediction: 1
- Training history: 2

3.2 Fold 2:

- Prediction: 3
- Training history: 4

Figure 3: Prediction Fold 2

Figure 4: Training history Fold 2

3.3 Fold 3:

- Prediction: 5
- Training history: 6

4 Discussion

The model achieves an average full-resolution mIoU of 65.52%, surpassing a vanilla baseline (60%). Low across-fold variance (std 1%) indicates stable training. The minor drop from low-res to full-res suggests our upsampling is effective, though multi-scale fusion or CRF could boost accuracy further.

5 Conclusion

A UNet–ResNet50 with simple augmentations produces strong segmentation results on Cityscapes. Future work could integrate attention modules or deeper decoder blocks for additional gains.

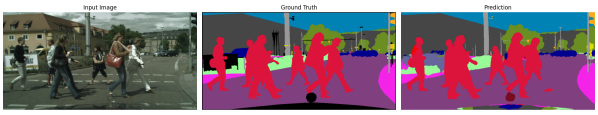

Figure 5: Prediction Fold 3

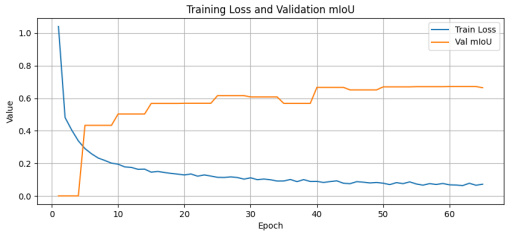

Figure 6: Training history Fold 3

Table 1: Cross-Validation mIoU (**512×1024** / **1024×2048**)

Fold	Low-Res mIoU	Full-Res mIoU
1	0.6905	0.6719
2	0.6696	0.6495
3	0.6717	0.6444
Mean	**0.6773**	**0.6552**
Std	0.0094	0.0120