Cityscapes Challenge

Venkata Pragna Nerella
dept. of Electrical Engineering
Eindhoven University of Technology
Eindhoven, Netherlands
v.p.nerella@student.tue.nl

***Abstract*—Computer vision systems using Neural Networks have significantly advanced applications in urban scene understanding. Semantic segmentation and object detection have become important tasks for various applications like intelligent transportation, autonomous vehicles, and smart city infrastructure. Despite exceptional progress, these technologies face critical challenges in practical, real-world urban environments [1]. These problems primarily include incomplete reliability of models trained on small or biased datasets, computational inefficiencies that make real-time applications difficult and limited generalizability of models because of shifts in illumination conditions, occlusion and complex backgrounds present in cityscapes [1]. This research addresses these challenges by using advanced preprocessing techniques, data augmentation methods like random resizing and flipping and optimised neural network architectures that balance accuracy and computational efficiency.**

***Index Terms*—Object detection, Semantic segmentation, Cityscapes, Deep learning, Neural networks**

I. Introduction

Semantic segmentation and object detection in urban environments is a critical task in computer vision as it enables various applications such as autonomous driving, urban planning and environmental monitoring. This method involves classifying each pixel in an image to pre-defined categories which facilitates a comprehensive understanding of complex urban scenes [2].

Urban scenes contain complexities as they feature a diverse array of objects with varying scales, occlusions and dynamic elements. This complexity poses significant challenges for semantic segmentation models. For instance, the irregularity and non-uniform distribution of elements complicate the application of traditional convolutional neural networks which are generally optimised for regular, grid-like data structures. Developing models capable of handling the spatial irregularities inherent in urban environments remain a critical challenge [3].

Another significant challenge is class imbalance. Common urban features like buildings and roads are often overrepresented in datasets. Less frequent elements like pedestrians and street signs though very critical are underrepresented. This imbalance can bias model training leading to suboptimal performance [3].

Furthermore, the dynamic nature of urban environments which include seasonal variations, daily human activities and construction can cause variability. This challenges the consistency and accuracy of segmentation models. Thus, models that can adjust to the constant changes in urban landscapes without requiring a lot of re-training are required [3].

It is evident from recent studies that a number of solutions have been put out to deal with these challenges. One strategy involves enhancing model architectures to capture both global context and local details better [4]. This can be seen as the integration of attention mechanisms into CNNs which has shown promising improvement in the segmentation of fine-grained details in urban scenes [5]. Another approach focuses on developing lightweight models that maintain accuracy while reducing computational complexity. For instance, the U-Net architecture has been used for cityscape image analysis as it offers a balance between performance and efficiency [6].

Additionally, data augmentation can be used to address the class imbalance challenge. The development of loss functions that can mitigate the effects of imbalance has been explored. These methods aim to ensure a more balanced representation of urban elements during model training which improves performance across both common and rare classes [3].

II. Methodology

A. Dataset Description

The project uses the Cityscapes Dataset [11] which is a well-known benchmark for semantic understanding of urban street scenes. The dataset includes high-resolution (2048×1024 pixels) RGB images captured across various European cities under diverse weather conditions. The dataset comprises of 5000 finely annotated images. Each image pixel is labeled into 30 predefined semantic classes representing common urban elements such as roads, pedestrians and buildings which facilitate detailed semantic segmentation.

B. Baseline Approach

An off-the-shelf semantic segmentation model based on a U-Net architecture is implemented for the baseline approach. U-Net architecture is a widely used convolutional neural network known for its effectiveness in segmentation tasks [6]. The architecture consists of symmetrical encoder-decoder structures with skip connections which helps in capturing both contextual information and fine spatial details effectively [5].

The baseline model used in the project consists of the following key components:

- **Encoder block:** Each encoder block consists of a convolutional block with two consecutive 3×3 convolutional layers. The layers are followed by batch normalization

and ReLU activation. After convolutional operations, each encoder block employs max pooling to preserve the most prominent features (like edges and textures) [2].

- **Bottleneck:** The encoder ends with a bottleneck layer made up of convolutional blocks that gather and combine the most important features of earlier layers. In this phase, the input is compresses into a smaller but more meaningful set of data representations [3].
- **Decoder block:** The decoder works like a reversed version of the encoder. Each decoder block starts with a transpose convolution (also called up-sampling), which increases the image size and decreases the number of channels. Then, it combines this output with matching features from the encoder using skip connections and applies convolutional layers similar to those in the encoder [4].
- **Classifier:** The final layer is a classifier composed of a convolutional layer that maps the features to 19 semantic classes corresponding to the Cityscapes labels. Each pixel in the input image is thus classified independently [6].

C. Data Preprocessing and Postprocessing

All the input images were resized to a uniform dimension $(512 \times 512 pixels)$ to ensure computational feasibility during the pre-processing phase.

Softmax function was applied across all the class dimensions to the network's raw output, identifying the class with the maximum probability for each pixel during the post-processing phase. The predicted segmentation masks were resized back to their original dimensions using nearest neighbour interpolation to ensure accurate evaluation against ground truth labels.

D. Peak Performance metric

Key enhancements were made to the the baseline U-Net model architecture to improve the segmentation accuracy and robustness for the Cityscapes dataset.

- **ASPP (Atrous Spatial Pyramid Pooling) Module for Multi-Scale Context Aggregation:** The model integrates an ASPP module which applies dilated convolutions with rates 6 and 12 to the bottleneck feature map allowing the network to extract features at multiple spatial resolutions. This helps the model to recognise both fine details (e.g., traffic signs) and larger structures (e.g., buildings) in urban scenes.
- **Global Context Integration:** To add a holistic view of the entire scene, the $useglobal = True$ flag in ASPP module is applied to enable a global average pooling branch. This enhancement improves the overall semantic consistency.
- **Improved Skip Connections:** The decoder uses enhanced skip connections with precise spatial alignment. It's achieved by padding the upsampled feature maps to match the encoder dimensions ensuring that the concatenated features are properly aligned and preserve spatial accuracy.
- **Transposed Convolutions for Upsampling:** Instead of simple interpolation, the decoder uses learnable transposed convolutions to perform upsampling. This allows the model to reconstruct higher-resolution features more accurately by learning the optimal upsampling patterns during training.
- **Channel Reduction for Efficiency:** A channel reduction strategy is used where the number of channels is reduced after the concatenation with skip connections. This keeps the parameter count low while retaining performance making the model faster and more memory efficient.

E. Robustness

To ensure that the model performs reliably across varying input conditions and generalizes well to unseen urban scenes, modifications were introduced to improve robustness:

- **Dropout Regularization:** A dropout layer with a rate of 0.1 was added after each convolution in the $DoubleConv$ blocks to prevent overfitting. This helps in making the model less sensitive to noise and specific patterns in the training data.
- **Batch Normalisation:** $BatchNorm$ was applied after every convolutional layer to stabilise and accelerate training by normalizing layer inputs which helps improve generalization.
- **Dilated Convolutions:** Within the **ASPP module**, dilated convolutions with varying rates were used. These increase the receptive field of filters without increasing the number of learnable parameters. This allows the model to learn from a broader spatial context.
- **Data Augmentations for real-world resilience:** Different data augmentations were applied during the training phase to improve the model's robustness.
  - Brightness and Contrast Variations
  - Motion Blur / Camera Shake
  - Rain and Fog Overlays
  - JPEG compression Artifacts
  - Random Shadows and Occlusions

III. Results

To evaluate the performance and effectiveness of the enhanced segmentation model, the following key metrics were monitored over the course of the training:

- **Training Loss:** Measures how well the model fits the training data.
- **Validation Loss:** Indicates how well the model generalizes to unseen data.
- **Validation Dice Coefficient:** A segmentation accuracy metric that measure overlap between predicted and ground truth masks.

A. Training Loss

The training loss graph (Fig. 1) shows a steady and consistent decrease over the 10,000+ training steps. Initially, the model started with a high loss (> 3.0), which dropped sharply within the first few thousand steps. After this, the loss

continued to decline gradually and stabilized around $0.2-0.3$, indicating successful convergence and effective learning of the training data.

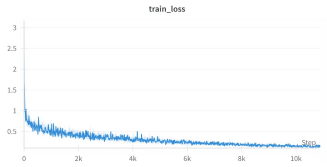

Fig. 1. Training Loss

B. Validation Dice Score

The validation dice score coefficient (from Fig. 2) improved steadily throughout the training. Stating around 0.27, it reached nearly 0.60 by the end of the training. This upward trend indicates that the model's segmentation accuracy on unseen validation data improved consistently with training. This suggests that the applied enhancements (ASPP, dropout, batch normalisation and better upsampling strategies) contributed positively to generalisation performance.

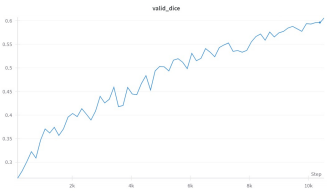

Fig. 2. Validation Dice

C. Validation Loss

The Validation loss (as seen in Fig. 3) also showed a notable decline from around 0.7 to below 0.3, further reinforcing that the model avoided overfitting and maintained good generalisation ability. Occasional small spikes are visible, which is expected due to the natural variation in validation batches. However, the overall trend remains downward.

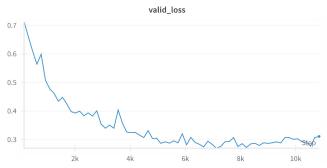

Fig. 3. Validation Loss

TABLE I
FINAL PERFORMANCE METRICS AFTER TRAINING

Metric	Value
Training Loss	∼0.20
Validation Loss	∼0.28
Validation Dice	**0.598**

The drop in training and validation loss demonstrates that the model learned the segmentation task effectively. The rise in the Dice score confirms improved segmentation quality, validating the architectural choices such as ASPP, transposed convolutional upsampling, and regularisation techniques.The smooth curves across all metrics indicate stable and well-tuned training with no signs of overfitting.

Fig. 4. Ground truth

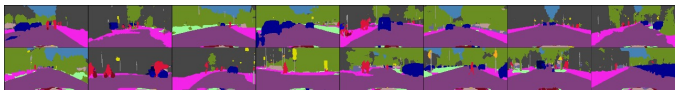

Fig. 5. Model prediction

Figure 4 and Figure 5 provides a comparison between the ground truth segmentation and the model's predicted output. The ground truth image accurately outlines the semantic regions, including roads, vehicles, and building structures, as manually annotated in the Cityscapes dataset. The predicted model image closely matches the ground truth, demonstrating the model's ability to identify key urban components with spatial precision. While the predictions are not perfect, especially along object boundaries, the results indicate strong overall segmentation accuracy and structural consistency.

IV. DISCUSSION

The performance metrics and training curves demonstrate that the enhanced semantic segmentation model performs effectively on Cityscapes dataset. The integration of ASPP for multi-scale context aggregation, global average pooling, and dropout regularisation significantly improved both the segmentation accuracy and generalisation capabilities.

The steady rise in validation Dice score, reaching nearly 0.60 suggests that the model captures fine-grained details and scene context with high fidelity. This makes the models suitable for real-world applications such as autonomous driving, smart traffic monitoring, and urban scene understanding.

The robustness of this model can be highlighted through its resilience to noise and visual distortions achieved through data augmentations and regularisation. The spatial accuracy was maintained by using transposed convolutions for upsampling and with the help of carefully aligned skip connections. This proved critical in detecting object boundaries such as lane markings or pedestrian outlines.

Despite the encouraging results, there are several limitations that should be considered:

- While the Dice coefficient steadily increased, it plateaued near 0.60, indicating that further gains may require dataset-level or architectural improvements.

- The current results do not break down performance by class which makes it hard to diagnose where the errors persist.
- While weather and noise augmentations were applied, their synthetic nature may not fully match the distribution of real-world artifacts leading to affect the field performance.

To further improve the model's performance and practicality, the following directions can be explored:

- **Class-wise performance tuning:** Analyzing per-class accuracy to guide loss weighting or focal loss for underrepresented classes [7].
- **Model pruning and quantisation:** To reduce inference time and memory usage for deployment on resource-constrained devices [10].
- **Attention Mechanisms:** Adding spatial or channel-wise attention blocks could help the model focus more on relevant features [8].
- **Semi-Supervised Learning:** Using consistency regularisation or pseudo-labels to improve training on unlabelled urban data [9].

ACKNOWLEDGMENT

I would like to express my sincere gratitude to the course instructors for providing this valuable opportunity to work on a real-world problem through the Cityscape Challenge. I'm especially thankful to the teaching assistants for their guidance and support throughout the project, particularly during the initial setup sessions which laid the groundwork for the entire implementation. I would also like to extend my heartfelt thanks to my friend Bharadwaj Palakurthy (palakurthybharadwaj012@gmail.com), who works as an Associate AI-ML engineer at Edwisely. Their insights and technical assistance right from conceptualizing the approach to fine-tuning the implementation played a crucial role in this project.

REFERENCES

[1] Cordts, Marius, et al. "The cityscapes dataset for semantic urban scene understanding." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[2] Wang, Libo, et al. "Transformer meets convolution: A bilateral awareness network for semantic segmentation of very fine resolution urban scene images." Remote Sensing 13.16 (2021): 3065.

[3] Yan, Hailun, Albert Lau, and Hongchao Fan. "Evaluating Deep Learning Advances for Point Cloud Semantic Segmentation in Urban Environments." KN-Journal of Cartography and Geographic Information (2025): 1-20.

[4] Bi, Qi, Shaodi You, and Theo Gevers. "Learning content-enhanced mask transformer for domain generalized urban-scene segmentation." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 38. No. 2. 2024.

[5] Liu, Jingyi, et al. "Semantic Segmentation of Urban Remote Sensing Images Based on Deep Learning." Applied Sciences (2076-3417) 14.17 (2024).

[6] Arulananth, T. S., et al. "Semantic segmentation of urban environments: Leveraging U-Net deep learning model for cityscape image analysis." Plos one 19.4 (2024): e0300767.

[7] Lin, Tsung-Yi, et al. "Focal loss for dense object detection." Proceedings of the IEEE international conference on computer vision. 2017.

[8] Wang, Xiaolong, et al. "Non-local neural networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[9] Chen, Xiaokang, et al. "Semi-supervised semantic segmentation with cross pseudo supervision." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.

[10] Han, Song, Huizi Mao, and William J. Dally. "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding." arXiv preprint arXiv:1510.00149 (2015).

[11] Cityscapes Dataset. [Online]. Available: https://www.cityscapes-dataset.com/dataset-overview/