

Segmentação Semântica de Cenas Urbanas usando Arquitetura UNET

João Pedro Vicente Ramalho GRR20224169

Luan Marko Kujavski GRR20221236

Heloisa Benedet Mendes GRR20221248

Departamento de Informática

Universidade Federal do Paraná – UFPR

Curitiba, Brasil

{jpvr22, lmk22, hbm22}@inf.ufpr.br

I. INTRODUÇÃO

A segmentação semântica de imagens é uma tarefa fundamental na visão computacional, com aplicações em sistemas de navegação autônoma, monitoramento urbano e análise geográfica. Este projeto implementa uma rede neural convolucional baseada na arquitetura UNET para segmentação de cenas urbanas do dataset Cityscapes.

O principal desafio abordado é a classificação pixel a pixel em imagens de alta resolução (1024x2048 pixels) contendo múltiplos objetos e classes desbalanceadas. A arquitetura UNET foi escolhida por sua eficiência em tarefas de segmentação e capacidade de preservar informações espaciais através de conexões residuais (*skip connections*).

II. FUNDAMENTAÇÃO TEÓRICA: ARQUITETURA UNET

A UNET é uma arquitetura de rede neural convolucional especializada em tarefas de segmentação semântica, originalmente desenvolvida para segmentação de imagens biomédicas. Sua estrutura combina um caminho de contração para capturar contexto e um caminho simétrico de expansão para permitir localização precisa, resolvendo o desafio de combinar informações contextuais com detalhes espaciais (Figura 1).

A. Arquitetura

A arquitetura opera através de três componentes principais:

- 1) **Caminho de Contração (Encoder):** Consiste em uma série de blocos convolucionais intercalados com operações de *downsampling*. Cada estágio aplica duas convoluções 3x3 não-preenchidas seguidas de funções de ativação ReLU. Em seguida, uma operação de *max pooling* 2x2 reduz as dimensões espaciais pela metade. O número de filtros dobra a cada estágio, seguindo a progressão de 64 para 128, depois para 256 e, por fim, para 512. O objetivo desse caminho é extrair características hierárquicas.

- 2) **Gargalo (Bottleneck):** Representa a camada intermediária da rede, caracterizada por uma maior profundidade, com 1024 filtros. Sua função é capturar características abstratas de alto nível, atuando como uma ponte entre os caminhos de contração e expansão.

- 3) **Caminho de Expansão (Decoder):** Consiste em operações simétricas inversas às do caminho de contração. Em cada estágio, realiza-se um *upsampling* 2x2 seguido de uma convolução 2x2. Em seguida, os *feature maps* resultantes são concatenados com os correspondentes do caminho de contração, permitindo a recuperação de detalhes espaciais. Duas convoluções 3x3 são aplicadas para refinar as características combinadas. O número de filtros é reduzido pela metade a cada estágio, seguindo a sequência de 512 para 256, depois 128 e, finalmente, 64.

B. Mecanismo de Conexões Residuais

A U-Net utiliza conexões residuais (*skip connections*) entre os blocos correspondentes do encoder e do decoder. Essas conexões transferem diretamente os *feature maps* do caminho de contração para o de expansão, preservando informações espaciais de alta resolução que poderiam ser perdidas durante o *downsampling*. Elas permitem a combinação de características contextuais extraídas em camadas mais profundas com detalhes locais presentes nas camadas mais rasas.

Além disso, contribuem para a mitigação do problema de *vanishing gradients* durante o treinamento, isto é, a diminuição progressiva dos gradientes à medida que se propagam pelas camadas iniciais da rede, o que pode dificultar o ajuste dos pesos nessas camadas. Essas conexões estão representadas na Figura 1 pelas setas horizontais em cinza que ligam diretamente os blocos do encoder aos seus correspondentes no decoder.

C. Principais Operações

- **Blocos Convolucionais Duplos:** Cada unidade básica aplica duas convoluções 2D sequenciais. Após cada convolução, são aplicadas normalização em lote (*Batch Normalization*) e função de ativação ReLU. A saída da primeira sequência serve como entrada para a segunda, formando um bloco que ajuda a estabilizar o treinamento.
- **Upsampling:** Implementado via convolução transposta:
$$\text{ConvTranspose2d}(C_{\text{in}}, C_{\text{out}}, \text{kernel_size} = 2, \text{stride} = 2)$$
 (1)

Onde:

- C_{in} : número de canais de entrada (provenientes do decoder);
 - C_{out} : número de canais desejado na saída (geralmente igual ao do bloco correspondente do encoder);
 - $kernel_size = 2$: define o tamanho da janela da convolução transposta (amplia as dimensões espaciais);
 - $stride = 2$: controla o fator de upsampling, dobrando altura e largura da imagem.
- **Camada Final:** Convolução 1×1 mapeia para $N_{classes}$ canais:

$$\text{Output} = \text{Conv2d}(64, N_{classes}, 1) \quad (2)$$

Onde:

- 64: número de canais na entrada da camada final (saída do último bloco do decoder);
- $N_{classes}$: número de classes da tarefa de segmentação;
- 1: tamanho do kernel 1×1 , que realiza um mapeamento ponto a ponto para cada pixel.

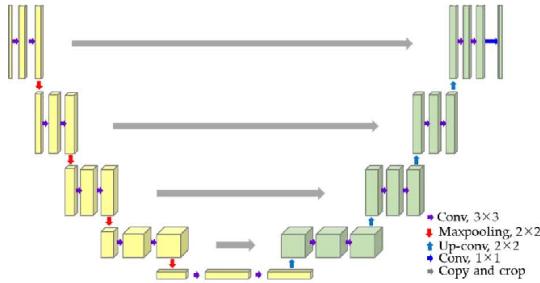


Figura 1: Diagrama da arquitetura UNET.

D. Vantagens para Segmentação Semântica

A eficácia da UNET deriva de:

- **Preservação de Informação Espacial:** Conexões residuais mantêm detalhes finos;
- **Campo Receptivo Ampliado:** Combina contexto global com detalhes locais;
- **Precisão em Fronteiras:** Recuperação progressiva de resolução espacial.

Na implementação deste trabalho, adaptamos a arquitetura original para processar imagens urbanas de alta resolução (1024×2048 pixels) com 34 classes, mantendo os princípios fundamentais da UNET enquanto otimizamos operações para eficiência computacional.

III. METODOLOGIA

A. Conjunto de Dados

O dataset Cityscapes contém:

- 5.000 imagens de cenas urbanas com alta diversidade;
- Anotações pixel-level para 34 classes semânticas;
- Imagens com resolução de 1024×2048 pixels;
- Divisão oficial: treino (2975), validação (500) e teste (1525).

B. Preparação dos Dados

Como o conjunto de teste não está publicamente disponível, os dados de treino e validação foram unificados. Em seguida, essa junção foi dividida em 70% para treinamento, 15% para validação e 15% para teste.

C. Parâmetros de Treinamento

- Função de perda: *CrossEntropyLoss*;
- Otimizador: Adam (learning rate=0.001);
- Batch size: 1 (devido a restrições de memória);
- Épocas: 20.

IV. RESULTADOS

A. Métricas Quantitativas

O modelo foi avaliado usando:

- **Acurácia de Pixels:** Proporção de pixels classificados corretamente;
- **IoU Médio:** Média das interseções sobre união por classe;
- **IoU por Classe:** Desempenho individual por categoria.

Os resultados obtidos mostram uma Pixel Accuracy de 0,8424, já o Mean IoU foi de 0,3269. Observando a Tabela I, que apresenta o desempenho por classe para as cinco melhores categorias, vemos que as classes "Out of ROI", "Ego Vehicle" e "Rectification Border" atingiram IoUs muito altas, próximas de 0,9 ou superiores, indicando que o modelo segmenta essas regiões com bastante precisão. As classes "Road" e "Sky" também apresentam bons resultados, com IoUs acima de 0,84, o que é esperado, pois são categorias visualmente bem definidas e frequentes no dataset.

No entanto, o valor relativamente baixo do Mean IoU geral indica que as outras classes do conjunto tiveram um desempenho significativamente pior, o que impacta a média. Isso pode ter ocorrido devido à complexidade da segmentação em classes menos representadas ou mais difíceis de identificar, o que é comum em tarefas de segmentação semântica com múltiplas categorias.

Tabela I: Melhores desempenhos por classe.

Classe	IoU	Categoria
3	0.9995	Out of ROI
1	0.9029	Ego Vehicle
2	0.9111	Rectification Border
7	0.8529	Road
23	0.8439	Sky

B. Análise Qualitativa

As visualizações geradas mostram:

- Boa segmentação de classes dominantes (estradas, edifícios, céu);
- Dificuldades em distinguir pedestres (classe *person*) de ciclistas (classe *rider*);
- Erros frequentes em fronteiras entre classes semelhantes e/ou sobrepostas;
- Confusão entre classes visualmente similares (ex: caminhões/ônibus).

V. RESULTADOS VISUAIS

A seguir, apresentamos quatro exemplos ilustrativos dos resultados obtidos pelo modelo.

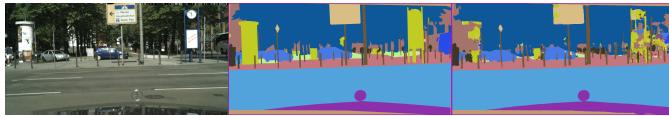


Figura 2: Exemplo 1.

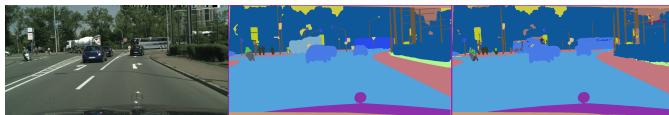


Figura 3: Exemplo 2.



Figura 4: Exemplo 3.

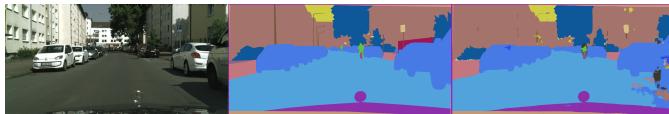


Figura 5: Exemplo 4.



Figura 6: Exemplo 5.

VI. DISCUSSÃO

Os resultados quantitativos indicam que o modelo UNet adaptado para cenas urbanas apresenta bom desempenho em classes visualmente distintas e frequentes, como “Out of ROI” ($\text{IoU}=0,9995$), “Rectification Border” ($\text{IoU}=0,9111$), “Ego Vehicle” ($\text{IoU}=0,9029$), “Road” ($\text{IoU}=0,8529$) e “Sky” ($\text{IoU}=0,8439$). Esses valores sugerem que a combinação do caminho de contração, gargalo e expansão, aliada às conexões residuais, é eficaz para recuperar detalhes espaciais e segmentar regiões homogêneas de grande área. A acurácia geral de pixels de 84,24% reforça essa capacidade de classificação global, mas o Mean IoU de 0,3269 sinaliza que muitas classes têm desempenho significativamente abaixo das melhores, puxando a média para baixo.

A discrepância entre a alta acurácia de pixels e o baixo Mean IoU evidencia o desbalanceamento em segmentação semântica multiclass: classes majoritárias e com aparências

bem definidas bem segmentadas, enquanto categorias menos representadas ou visualmente semelhantes — por exemplo, pedestres *versus* ciclistas (Figura 6) ou veículos de grande porte como caminhões e ônibus (Figura 3) — sofrem de alta taxa de falsos positivos e negativos. Esse comportamento foi confirmado na análise qualitativa, que mostrou confusões nas fronteiras entre objetos e falhas na separação de instâncias próximas ou parcialmente ocluídas.

Para mitigar essas limitações, futuras melhorias podem incluir estratégias de balanceamento de classes durante o treinamento (como ponderação de perda ou *oversampling* de classes minoritárias), o uso de módulos de atenção para focar em regiões de difícil distinção e a experimentação com arquiteturas híbridas que combinem UNet com blocos de agrupamento mais profundos (por exemplo, ResNet).

VII. CONCLUSÃO

Este projeto implementou um *pipeline* completo para segmentação semântica usando arquitetura UNET no dataset Cityscapes. Os principais resultados são:

- Sistema funcional de preparação, treinamento e avaliação;
- Modelo capaz de segmentar 34 classes com acurácia global de 84.24%;
- Boa performance em classes dominantes (estrada, edificações, céu);
- Desafios em classes minoritárias e objetos pequenos.

O código-fonte completo está disponível em: https://github.com/joaop-vr/computer_vision