Reconhecimento Denso para o Entendimento de Ambiguidades em Ambientes Urbanos de Veículos Autônomos

[1,]Wilbur N. Chiuyari-Veramendi, [1,]Fernando S. Osório

[1]Instituto de Ciências Matemáticas e de Computação, São Carlos, SP, Brazil
E-mails: wilburncv@usp.br, fosorio@icmc.usp.br

***Abstract*—In autonomous navigation of mobile robots and vehicles, dense scene understanding—through semantic segmentation and depth estimation—is essential to ensure safety and adaptability in complex urban environments. Despite recent advances in computer vision, such as vision transformers and foundational models, there are still few studies that unify dense recognition applied to mobile robots and autonomous vehicles. This work proposes a lightweight, integrated architecture that exploits the complementarity between these two tasks by combining representations extracted from foundational models like DINOv2 and Depth Anything. Semantic-geometric integration strategies, including depth map concatenation, weighted spatial attention, and cross-attention, are investigated to enhance segmentation robustness in ambiguous situations, such as reflections, occlusions, and unreal objects. The approach is evaluated on an embedded platform (Jetson AGX Orin), considering quantitative metrics (mIoU, AbsRel, FPS) and qualitative assessments, focusing on computational feasibility. Expected results indicate that, even with a frozen encoder, training limited to the heads combined with efficient strategies can achieve a relevant balance between semantic performance, geometric accuracy, and real-time resource usage.**

***Resumo*—Na navegação autônoma de robôs e veículos móveis, a compreensão de cenas densas — por meio de segmentação semântica e estimativa de profundidade — é essencial para garantir segurança e adaptabilidade em ambientes urbanos complexos. Apesar dos avanços recentes em visão computacional, como transformadores de visão e modelos fundamentais, ainda existem poucos estudos que unificam o reconhecimento denso aplicado a robôs móveis e veículos autônomos. Este trabalho propõe uma arquitetura leve e integrada que explora a complementaridade entre essas duas tarefas, combinando representações extraídas de modelos fundamentais como DINOv2 e Depth Anything. Estratégias de integração semântico-geométrica, incluindo concatenação de mapas de profundidade, atenção espacial ponderada e atenção cruzada, são investigadas para aumentar a robustez da segmentação em situações ambíguas, como reflexões, oclusões e objetos irreais. A abordagem é avaliada em uma plataforma embarcada (Jetson AGX Orin), considerando métricas quantitativas (mIoU, AbsRel, FPS) e avaliações qualitativas, com foco na viabilidade computacional. Os resultados esperados indicam que, mesmo com um codificador congelado, o treinamento limitado às cabeças combinado com estratégias eficientes pode alcançar um equilíbrio relevante entre desempenho semântico, precisão geométrica e uso de recursos em tempo real.**

I. Introdução

O reconhecimento denso da cena, por meio da segmentação semântica e da estimativa de profundidade, é fundamental para a navegação autônoma de robôs móveis e veículos em ambientes urbanos complexos. A segmentação semântica atribui rótulos a cada pixel da imagem, possibilitando a identificação e classificação de objetos [1], enquanto a estimativa de profundidade fornece informações sobre a distância relativa entre o veículo e os objetos, essencial para evitar colisões [2]. Contudo, essas tarefas enfrentam desafios em condições adversas, como reflexos, oclusões e objetos irreais, que geram ambiguidades na percepção [3], [4].

Métodos recentes de estimação de profundidade baseados em aprendizado profundo exploram redes neurais para gerar mapas densos a partir de imagens monoculares, utilizando paradigmas supervisionados, não supervisionados e semi-supervisionados [2], [5]–[7]. Na segmentação, arquiteturas como Convolutional Neural Networks (CNNs) e Vision Transformers (ViT) têm alcançado avanços relevantes [1], [8], embora dependam de grande quantidade de dados anotados e tenham limitações para reconhecer categorias fora do conjunto de treinamento [9].

Modelos fundamentais (*Foundation Models*, FM), como CLIP [10], SAM [11] e Depth Anything [12], vêm ganhando destaque por sua capacidade de generalização e aprendizado auto-supervisionado em larga escala. Em particular, SAM oferece segmentação semântica robusta e adaptável via prompts, enquanto Depth Anything apresenta estimativas de profundidade monocular mais precisas, beneficiando-se de pseudo-rótulos gerados automaticamente.

Este trabalho apresenta uma arquitetura integrada leve que explora a complementaridade entre segmentação semântica e estimativa de profundidade, buscando aumentar a robustez da percepção em cenários ambíguos da navegação autônoma urbana (Figura 1). Investiga-se a combinação de representações extraídas de FM, por meio de estratégias de integração semântico-geométrica, como concatenação de mapas de profundidade e mecanismos de atenção. A avaliação está em andamento em uma plataforma embarcada (Jetson AGX Orin), com métricas quantitativas e qualitativas.

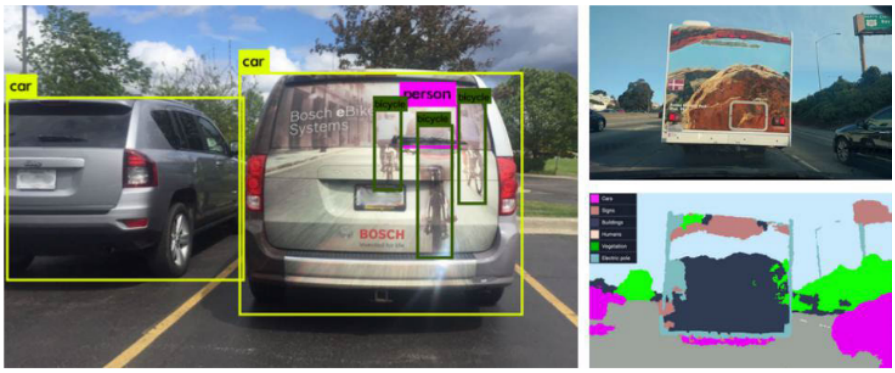

Figura 1. Ambiguidade de percepção 3D com ilustrações sintéticas 2D. Fonte: Bruno *et al*., (2023) [13]

II. Trabalhos Relacionados

O reconhecimento denso, que inclui segmentação semântica e estimação de profundidade, tem sido amplamente explorado nas últimas décadas, com avanços recentes centrados em ViTs e FMs. Caron *et al.* [14] mostraram que o aprendizado auto-supervisionado aplicado a ViTs por meio do método DINO, que usa *knowledge distillation* sem rótulos, pode superar abordagens supervisionadas em tarefas de segmentação e classificação, destacando a importância do aprendizado sem necessidade de anotação manual. Seguindo essa linha, Oquab *et al.* [15] aprimoraram o conceito com o DINOv2, que combina curadoria cuidadosa de dados e treinamento com Masked Autoencoders (MAE), revelando propriedades emergentes capazes de compreender partes e geometria de objetos em múltiplos domínios, melhorando a segmentação e estimativa de profundidade monocular.

A generalização de modelos para segmentação semântica avançou significativamente com o SAM de Kirillov *et al.* [11], que propõe uma arquitetura promptable e zero-shot, treinada em um gigantesco conjunto de dados (SA-1B). SAM consegue lidar com ambiguidades ao gerar múltiplas máscaras para um mesmo prompt, o que é crucial para cenários complexos e variados. Complementando a visão semântica, o modelo *Depth Anything* de Yang *et al.* [12] utiliza uma enorme base de imagens não rotuladas para treinar um estimador monocular de profundidade que supera métodos clássicos, explorando pseudoetiquetas e aprendizado auto-supervisionado para robustez em diferentes ambientes.

Enquanto ViTs e FMs ganham terreno, métodos tradicionais baseados em CNNs ainda apresentam contribuições relevantes para manter a continuidade local e o contexto global na segmentação, como exemplificado pelo Context Aggregation Module (CAM) de QuanTang *et al.* [16], que evita perdas de informação por pooling agressivo, promovendo resultados robustos em ambientes urbanos.

No contexto aplicado à condução autônoma, o projeto CARINA [13] destaca-se por integrar dados 2D e 3D para superar limitações de percepção unidimensional, elevando a precisão da segmentação em elementos urbanos complexos. Paralelamente, modelos linguagem-visão (VLMs), como revisado por Zhou *et al.* [17], introduzem percepções multimodais que combinam visões e comandos textuais para navegação autônoma e detecção, embora dependam de prompts cuidadosamente elaborados.

Estudos recentes também investigam a robustez dos FMs em condições adversas, como o trabalho de Shan *et al.* [18], que demonstrou a queda de desempenho do SAM em cenários climáticos severos, evidenciando desafios práticos para aplicações reais. Em paralelo, Liu *et al.* [19] apresentaram o método Seal, transferindo conhecimento de modelos 2D para segmentação 3D de nuvens de pontos sem rótulos, ampliando o escopo da percepção em veículos autônomos. Outro avanço na estimação métrica de profundidade é o Semi-SD de Xie *et al.* [20], que combina múltiplos sinais espaciais, temporais e semânticos via transformador unificado, aumentando a precisão e eficiência computacional em câmeras de visão surround para carros autônomos.

A Tabela I resume as características desses métodos, destacando sua diversidade e contribuições no reconhecimento denso. Em contraste, nosso trabalho busca integrar esses avanços, focando na robustez em cenários urbanos ambíguos com reflexos, sobreposições e objetos irreais, enquanto considera a viabilidade computacional para sistemas embarcados, um aspecto pouco explorado nas abordagens revisadas.

III. Metodologia

A metodologia proposta neste trabalho tem como objetivo explorar a complementaridade entre segmentação semântica e estimação de profundidade, utilizando representações extraídas de FM. Diferentemente das abordagens tradicionais, que tratam as tarefas de reconhecimento denso de forma isolada, aqui propõe-se uma estratégia de segmentação guiada por profundidade, visando reforçar a compreensão da cena e aumentar a robustez da segmentação em contextos ambíguos de entornos urbanos não estruturados.

A arquitetura base adotada deriva do modelo *Depth Anything* [12], que emprega um *encoder* visual treinado em larga escala com dados rotulados e pseudo-rotulados. A partir deste *encoder* compartilhado, são derivados dois cabeçalhos (*heads*): (i) um para estimação de profundidade (original do *Depth Anything*) e (ii) outro, proposto neste trabalho,

Tabela I
COMPARAÇÃO DE METODOLOGIAS PARA RECONHECIMENTO DENSO

Metodos	Ano	Tarefa Principal	Tipo de Modelo	Contribuição Chave
DINO [14]	2021	Segmentação, Classificação	ViT + Self-Supervised	Treinamento com distilação sem rótulo (DINO)
CAM/CANet [16]	2022	Segmentação	CNN com atenção multi-escala	Preserva continuidade local, sem pooling agressivo
SAM [11]	2023	Segmentação Semântica	FM	Promptable e zero-shot com múltiplas máscaras
DINOv2 [15]	2023	Segmentação, Profundidade	ViT + MAE	Curadoria de dados + propriedades emergentes
Depth Anything [12]	2024	Estimação de Profundidade	FM	62M imagens não rotuladas com pseudoetiquetas
VLMs [17]	2023	Percepção multimodal	VLM (CLIP)	Zero-shot + prompts em navegação e detecção
SAM Robustez [18]	2023	Segmentação adversa	FM	Avaliação sob clima adverso
Seal [19]	2024	Segmentação 3D	FM (2D→3D Transfer)	Segmentação de nuvens de pontos sem rótulo
Semi-SD [20]	2025	Profundidade Métrica	STST (ViT + SAM)	Fusão de múltiplos sinais + perda curva

dedicado à segmentação semântica, inspirado na arquitetura do *DINOv2* [15].

O pipeline geral da arquitetura proposta é ilustrado na Figura 2, evidenciando a estrutura dual com ramificações específicas para segmentação e profundidade, permitindo que a primeira se beneficie das pistas geométricas extraídas pela segunda.

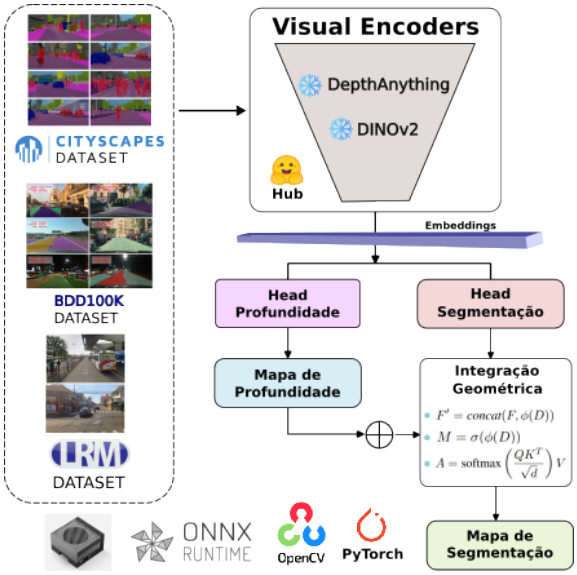

Figura 2. Arquitetura proposta para segmentação guiada por profundidade em sistemas embarcados. Fonte: Elaborada pelo autor.

A. Configuração Experimental

Para a implementação e avaliação da arquitetura proposta, considera-se uma plataforma embarcada de alto desempenho como a *NVIDIA Jetson AGX Orin 64GB*, realista para aplicações em robôs móveis ou veículos autônomos. Essa plataforma oferece até 275 TOPS de poder computacional (modo INT8), 64 GB de memória e suporte completo ao ecossistema *CUDA*, *TensorRT* e *JetPack SDK 5.x*, incluindo *PyCUDA* e exportações padronizadas via *ONNX*, recursos essenciais para execução eficiente em FMs.

Com o objetivo de viabilizar o uso embarcado com eficiência computacional, o treinamento será conduzido apenas nos cabeçalhos (*heads*) das tarefas de segmentação semântica e estimação de profundidade. O encoder visual, baseado no *DINOv2*, permanecerá congelado, atuando exclusivamente como extrator de características multiuso para ambas as tarefas de reconhecimento denso. Essa decisão reduz significativamente o consumo de memória, o tempo de execução e a sobrecarga computacional durante o treinamento.

Nesse sentido, utilizando subconjuntos de datasets como *Cityscapes* [21], *BDD100K* [22], será criada uma coleção específica de imagens reais contendo ambiguidades visuais. Para isso, será desenvolvido um motor de busca e filtragem de *datasets* já existentes, com o objetivo de identificar amostras que incluam elementos representados ou irreais coexistindo com objetos físicos. O resultado será um subconjunto de *LRM Dataset* urbano enriquecido com ambiguidades, permitindo avaliar se as estratégias de integração semântico-geométrica conseguem diferenciar entre objetos representados em 2D e obstáculos físicos no espaço tridimensional (Figura 3). Assim, o conjunto atua como um *stress test* para a arquitetura proposta, complementando a avaliação em cenários reais convencionais.

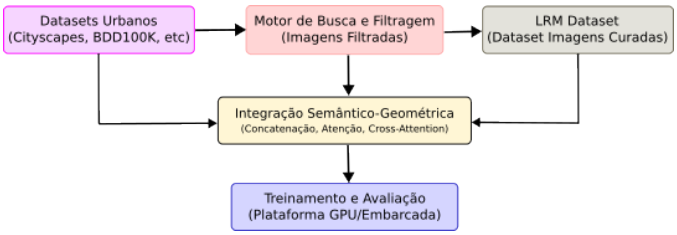

Figura 3. Pipeline dos datasets urbanos considerando ambiguidades e sua integração semântico-geométrica. Fonte: Elaborada pelo autor.

O treinamento dos cabeçalhos será realizado com dados anotados, utilizando *mixed precision* (FP16), *batch size* reduzido e acumulação de gradientes. Para a inferência, serão aplicadas otimizações com *TensorRT*, compressão de tensores e técnicas de quantização, assegurando desempenho em tempo real no ambiente embarcado.

B. Técnicas de reconhecimento denso

Seja $I \in \mathbb{R}^{H\times W\times 3}$ uma imagem de entrada, o pipeline proposto utiliza um *encoder visual compartilhado* $E(\cdot)$, pretreinada em larga escala com supervisão total e pseudo-supervisão.

$$F = E(I) \tag{1}$$

onde $F \in \mathbb{R}^{H'\times W'\times d}$ representa as *características latentes* extraídas da imagem. O encoder é derivado do modelo *Depth Anything*, que aproveita milhões de imagens não rotuladas enriquecidas com pseudo-rótulos, otimizando a generalização geométrica.

A arquitetura é bifurcada em dois cabeçalhos especializados:

- Um *cabeçalho de profundidade* H_{depth}, herdado diretamente da arquitetura original do *Depth Anything*, responsável por estimar um mapa denso de profundidade:

$$D = H_{depth}(F) \in \mathbb{R}^{H\times W} \tag{2}$$

- Um *cabeçalho semântico* H_{seg}, inspirado na arquitetura do *DINOv2*, que utiliza características autorregressivas para gerar uma predição de classes por pixel:

$$S = H_{seg}(F, D) \in \mathbb{R}^{H\times W\times C} \tag{3}$$

Neste contexto, D representa uma representação do mapa de profundidade que será usada como pista auxiliar na inferência semântica. Desta forma, o *DINOv2*, fornecera suas representações robustas, que serão adaptadas ao cabeçalho de segmentação.

C. Estratégias de Integração Semântico-Geométrica

Para conseguir a incorporação de informações geométricas guiadas pela inferência semântica, serão descritas algumas técnicas propostas desde as mais simples (Estrategia 1) até as mais sofisticadas (Estrategia 2 ou Estrategia 3).

1) Concatenação Direta: a primeira abordagem visa em concatenar diretamente o mapa de profundidade às características latentes, conforme é presentado na Equação 4.

$$F' = concat(F, \phi(D)) \tag{4}$$

onde $\phi(D)$ é uma projeção (convolução linear de 1×1) que ajusta as dimensões de D a fim de permitir concatenação, F' é o mapa combinado a ser processado pelas camadas subsequentes do cabeçalho de segmentação.

2) Atenção Espacial Ponderada por Profundidade: nesta abordagem, propõe-se utilizar o mapa de profundidade estimado para modular espacialmente as representações semânticas extraídas pelo encoder. Para isso, uma função de projeção $\phi(D)$ transformaria o mapa de profundidade D em um mapa de atenção M, aplicando uma função de ativação:

$$M = \sigma(\phi(D)) \tag{5}$$

Esse mapa M seria utilizado para ponderar multiplicativamente as feições semânticas S, resultando em uma representação guiada por profundidade $\hat{S}$:

$$\hat{S}(i, j, :) = M(i, j) \cdot S(i, j, :) \tag{6}$$

3) Atenção Cruzada: como variação adicional, explora-se o uso de um módulo de *atenção cruzada* entre as representações de profundidade D e as representações semânticas F. O mecanismo pode ser formalizado como:

$$A = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{7}$$

onde Q são as queries extraídas das características semânticas, A representa a atenção cruzada aplicada às representações semânticas, K e V são as keys e values extraídas das características de profundidade.

Estas abordagems permitiriam que a segmentação seja guiada de forma adaptativa pela estrutura geométrica da cena, sendo útil em regiões com oclusões, reflexos ou artefatos visuais.

IV. Resultados Esperados

Nesta seção, apresentam-se as expectativas relativas ao desempenho da metodologia proposta, tanto em termos quantitativos quanto qualitativos. Considerando o objetivo de integrar informações semânticas e geométricas. As métricas selecionadas para análise contemplam aspectos essenciais de qualidade da segmentação, precisão na estimativa de profundidade e eficiência na inferência, possibilitando uma avaliação completa do impacto da arquitetura em comparação a abordagens baseline.

A. Resultados quantitativos

A métrica primária para segmentação semântica é o *Intersection over Union* médio (mIoU), definido como:

$$\text{mIoU} = \frac{1}{C}\sum_{c=1}^{C}\frac{TP_c}{TP_c + FP_c + FN_c} \tag{8}$$

onde C representa o número total de classes, e TP_c, FP_c, FN_c correspondem, respectivamente, aos verdadeiros positivos, falsos positivos e falsos negativos da classe c. Um aumento no mIoU indicaria que a incorporação da profundidade melhora a delimitação semântica entre categorias visualmente semelhantes.

Além disso, serão consideradas acurácia de pixels (PA - Equação 9) e acurácia média por classe (mPA - Equação 10)

$$\text{PA} = \frac{\sum_c TP_c}{\sum_c (TP_c + FP_c + FN_c)} \tag{9}$$

$$\text{mPA} = \frac{1}{C}\sum_{c=1}^{C}\frac{TP_c}{TP_c + FN_c} \tag{10}$$

Estas métricas avaliam a cobertura geral e a equidade entre categorias, o que é crítico em cenas urbanas onde objetos minoritários (como pedestres) são relevantes.

Em relação a estimação de profundidade, são adotadas o Erro absoluto relativo (AbsRel) e o Erro quadrático médio (RMSE), descritas nas Equações 11 e 12 respectivamente.

$$\text{AbsRel} = \frac{1}{|T|} \sum_{i \in T} \frac{|d_i - \hat{d}_i|}{d_i} \tag{11}$$

$$\text{RMSE} = \sqrt{\frac{1}{|T|} \sum_{i \in T} (d_i - \hat{d}_i)^2} \tag{12}$$

onde d_i e $\hat{d}_i$ são, respectivamente, a profundidade real e a estimada para o pixel i, e T é o conjunto total de pixels válidos. Espera-se que a regressão geométrica de profundidade beneficie a segmentação ao reforçar descontinuidades estruturais e a separação entre planos.

No contexto de viabilidade computacional, especialmente para execução em dispositivos embarcados como a Jetson AGX Orin, será avaliados os Frames por segundo (FPS) e o Uso de memória (MB).

Espera-se que, mesmo com o *encoder* mantido congelado, o treinamento de cabeçalhos especializados, aliado a estratégias como quantização em FP16 e mecanismos de atenção guiada por profundidade, proporcione um equilíbrio eficaz entre desempenho semântico ($\uparrow$ mIoU, $\uparrow$ mPA), precisão geométrica ($\downarrow$ AbsRel, $\downarrow$ RMSE) e custo computacional ($\uparrow$ FPS, $\downarrow$ Memória). Esses indicadores serão comparados sistematicamente com a linha de base (*baseline*) e com as diferentes estratégias de integração semântico-geométrica propostas neste trabalho.

B. Resultados Qualitativos

Além das métricas quantitativas, espera-se que a proposta contribua significativamente para a qualidade visual e a interpretabilidade dos resultados de segmentação em cenários urbanos. A integração entre características semânticas e geométricas tem o potencial de gerar mapas de segmentação semanticamente mais coerentes, com contornos mais precisos e menor ocorrência de falhas em regiões ambíguas.

Visualmente, os resultados esperados incluiram: (i) melhor separação entre planos de profundidade distintos, como calçadas, veículos e fachadas, reduzindo erros de classificação por proximidade de cor ou textura; (ii) maior consistência em regiões com reflexos, sobreposições, sombras e objetos irreais (e.g., adesivos ou pinturas), graças à introdução de informações geométricas oriundas da estimação de profundidade; (iii) redução de ruídos semânticos em bordas e descontinuidades, devido ao reforço espacial guiado pela estrutura da cena.

Para reforçar essa avaliação, pretende-se apresentar comparações visuais lado a lado entre os mapas de segmentação obtidos com a abordagem proposta e os obtidos pelas abordagens de linha de base, destacando melhorias perceptíveis na delimitação semântica e no alinhamento com a estrutura tridimensional da cena.

V. Resultados Preliminares

Nesta etapa, conduzimos experimentos preliminares de segmentação semântica em ambiente não embarcado, com análise qualitativa, comparando uma rede baseada em CNN (BiSeNetV2) e um modelo fundamental (DINOv2), com o objetivo de visualizar as diferenças entre essas abordagens. A Figura 4 ilustra que, a partir de uma imagem de entrada contendo uma pequena região ambígua, a BiSeNetV2 identifica incorretamente essa região como um sinal de trânsito (Figura 4 - centro), além de apresentar classificações equivocadas da classe “grade” sobre o ônibus. Por outro lado, o DINOv2, aplicado à mesma imagem de entrada, consegue resolver essa ambiguidade (Figura 4 - ezquerda), embora apresente uma classificação incorreta ao rotular uma pessoa como “árvore”.

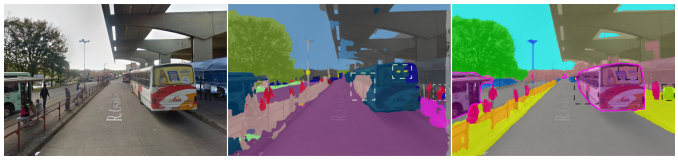

Figura 4. Comparação CNN (centro) e FM (esquerda) a partir de uma imagem fora da distribuição (direita), registrada na rodoviária de São Carlos, SP.

Por outro lado, utilizando a arquitetura *Depth Anything* [12] embarcada na plataforma *NVIDIA Jetson AGX Orin* (DA-JAO). A Figura 5 ilustra exemplos em imagens urbanas do KITTI, onde observamos mapas de profundidade apenas parcialmente consistentes em condições de iluminação variável. Esses resultados reforçam que, embora a execução embarcada seja viável, surgem limitações relevantes: as estimativas de profundidade tendem a se degradar conforme o backbone cresce em complexidade, comprometendo a estabilidade visual em cenários mais desafiadores.

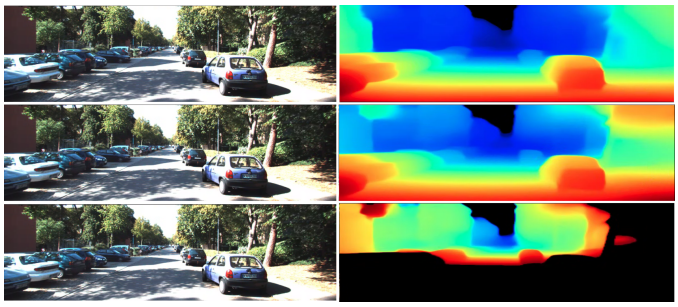

Figura 5. Predições de profundidade com DA-JAO, usando: ViT-S que preserva a estrutura geral (topo), ViT-B introduz ruído e desfoque (meio), enquanto ViT-L apresenta colapso com artefatos severos (base). Fonte: Elaborada pelo autor.

Como próximos passos, planejamos integrar o *DINOv2* como *encoder* visual e explorar as estratégias de fusão semântico-geométrica propostas em plataformas embarcadas baseadas em GPU Jetson. Essa evolução permitirá realizar comparações visuais lado a lado com abordagens de linha de base, ampliando a análise para além da profundidade e incluindo mapas semânticos integrados.

VI. Conclusões e Trabalhos Futuros

Este trabalho propôs uma arquitetura eficiente para reconhecimento denso em ambientes urbanos ambíguos, integrando segmentação semântica e estimação de profundidade a partir de FM, como DINOv2 e Depth Anything. Diferentemente de abordagens tradicionais, explorou-se a complementaridade semântico-geométrica por meio de estratégias como concatenação direta, atenção espacial ponderada, e atenção cruzada. A proposta visa não apenas melhorar a robustez frente a reflexos, sobreposições e objetos irreais, mas também garantir viabilidade em sistemas embarcados como a Jetson AGX Orin. Com treinamento restrito aos cabeçalhos e encoder congelado, espera-se alcançar um equilíbrio entre desempenho, generalização e eficiência, abrindo caminho para futuras aplicações em robótica móvel e carros autonomos.

Embora os experimentos preliminares em plataformas embarcadas (NVIDIA Jetson AGX Orin) mostraram que a qualidade das estimativas de profundidade pode se degradar em cenários mais complexos, especialmente com backbones mais robustos ou superfícies reflexivas e transparentes, reforçando a importância de estratégias de integração semântico-geométrica para lidar com ambiguidades urbanas.

Finalmente para trabalhos futuros, são considerados: (i) integração completa do DINOv2 como encoder visual avaliando a fusão semântico-geométrica em backbones mais complexos, mantendo viabilidade embarcada; (ii) expansão do conjunto de dados com ambiguidades, desenvolvendo o LRM Dataset, incorporando novas cenas urbanas; (iii) otimização para o compromiso de acurácia e tempo real; e (iv) generalização em outros cenários urbanos buscando compreender limitações e ajustes necessários para diferentes aplicações robóticas.

Agradecimentos

Os autores agradecem ao Instituto de Ciências Matemáticas e de Computação (ICMC) da USP e ao Laboratório de Robótica Móvel (LRM) pelo apoio institucional e infraestrutura disponibilizada durante o desenvolvimento deste trabalho, bem como a CAPES e ao programa MOVER/FUNDEP − Projeto SegCom pelo apoio financeiro.

Referências

[1] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[2] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2650–2658.

[3] X. Yu, Y. Zuo, Z. Wang, X. Zhang, J. Zhao, Y. Yang, L. Jiao, R. Peng, X. Wang, J. Zhang, K. Zhang, F. Liu, R. Alcover-Couso, J. C. SanMiguel, M. Escudero-Viñolo, H. Tian, K. Matsui, T. Wang, F. Adan, Z. Gao, X. He, Q. Bouniot, H. Moghaddam, S. N. Rai, F. Cermelli, C. Masone, A. Pilzer, E. Ricci, A. Bursuc, A. Solin, M. Trapp, R. Li, A. Yao, W. Chen, I. Simpson, N. D. F. Campbell, and G. Franchi, "The robust semantic segmentation uncv2023 challenge results," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2023, pp. 4618–4628.

[4] J. Spencer, C. S. Qian, C. Russell, S. Hadfield, E. Graf, W. Adams, A. J. Schofield, J. H. Elder, R. Bowden, H. Cong, S. Mattoccia, M. Poggi, Z. K. Suri, Y. Tang, F. Tosi, H. Wang, Y. Zhang, Y. Zhang, and C. Zhao, "The monocular depth estimation challenge," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, January 2023, pp. 623–632.

[5] C. Zhao, Q. Sun, C. Zhang, Y. Tang, and F. Qian, "Monocular depth estimation based on deep learning: An overview," *Science China Technological Sciences*, vol. 63, no. 9, pp. 1612–1627, 2020.

[6] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1851–1858.

[7] A. Pilzer, S. Lathuiliere, N. Sebe, and E. Ricci, "Refine and distill: Exploiting cycle-inconsistency and knowledge distillation for unsupervised monocular depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9768–9777.

[8] B. Cheng, A. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," *Advances in neural information processing systems*, vol. 34, pp. 17 864–17 875, 2021.

[9] Y. Zhao, L. Wang, X. Yun, C. Chai, Z. Liu, W. Fan, X. Luo, Y. Liu, and X. Qu, "Enhanced scene understanding and situation awareness for autonomous vehicles based on semantic segmentation," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2024.

[10] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[11] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.

[12] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 10 371–10 381.

[13] D. R. Bruno, R. Berri, F. Barbosa, and F. S. Osorio, "Carina project: Visual perception systems applied for autonomous vehicles and advanced driver assistance systems (adas)," *IEEE Access*, 2023.

[14] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," *Advances in neural information processing systems*, vol. 33, pp. 9912–9924, 2020.

[15] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[16] Q. Tang, F. Liu, T. Zhang, J. Jiang, Y. Zhang, B. Zhu, and X. Tang, "Compensating for local ambiguity with encoder-decoder in urban scene segmentation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 10, pp. 19 224–19 235, 2022.

[17] X. Zhou, M. Liu, B. Luka Zagar, E. Yurtsever, and A. C. Knoll, "Vision language models in autonomous driving and intelligent transportation systems," *arXiv e-prints*, pp. arXiv–2310, 2023.

[18] X. Shan and C. Zhang, "Robustness of segment anything model (sam) for autonomous driving in adverse weather conditions," *arXiv preprint arXiv:2306.13290*, 2023.

[19] Y. Liu, L. Kong, J. Cen, R. Chen, W. Zhang, L. Pan, K. Chen, and Z. Liu, "Segment any point cloud sequences by distilling vision foundation models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[20] Y. Xie, Z. Huang, S. Shen, and J. Ma, "Semi-sd: Semi-supervised metric depth estimation via surrounding cameras for autonomous driving," *arXiv preprint arXiv:2503.19713*, 2025.

[21] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[22] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2636–2645.