

UNIVERSIDADE FEDERAL DO ACRE

Gustavo Moreira Oliveira de Castro

Análise comparativa de estimadores monoculares de profundidade relativa

**RIO BRANCO
2024**

UNIVERSIDADE FEDERAL DO ACRE

Gustavo Moreira Oliveira de Castro

Análise comparativa de estimadores monoculares de profundidade relativa

Proposta de dissertação de mestrado submetida ao Programa de Pós-Graduação em Ciência da Computação na Universidade Federal do Acre como requisito parcial para obtenção do título de mestre em Ciência da Computação. Linha de Pesquisa: Sistemas Computacionais Inteligentes

Orientador:

Prof. Dr. Roger Fredy Larico Chavez

RIO BRANCO

2024

Gustavo Moreira Oliveira de Castro

Análise comparativa de estimadores monoculares de profundidade relativa

Proposta de dissertação de mestrado submetida ao Programa de Pós-Graduação em Ciência da Computação na Universidade Federal do Acre como requisito parcial para obtenção do título de mestre em Ciência da Computação. Linha de Pesquisa: Sistemas Computacionais Inteligentes.

Approved in <MONTH> of <YEAR>.

Prof. Dr. Roger Fredy Larico Chavez

Universidade Federal do Acre

Prof. Dr. ...

Universidade Federal do Acre

Prof. Dr. ...

Universidade Federal do Acre

RIO BRANCO

2024

dfsaas

-

Agradecimentos

...

Resumo

Análise comparativa de estimadores monoculares de profundidade relativa

Informação de profundidade possui grande importância em diversas aplicações que exigem informação geométrica, como veículos autônomos, navegação robótica, realidade aumentada e geração de conteúdo por inteligência artificial. Mapas de profundidade podem ser adquiridos com sensores ou através de estimação por métodos passivos. O desenvolvimento de técnicas de aprendizado profundo propiciou a viabilidade da tarefa de estimação monocular de profundidade, em que é utilizada uma única imagem RGB como entrada de um sistema de rede neural. O presente trabalho se propõe a realizar uma análise comparativa dos estimadores monoculares do estado da arte em termos de métricas presentes na literatura, transferência de domínio de relativo para métrico e o emprego em aplicações práticas. Os resultados preliminares demonstraram o funcionamento do método de transformação de intensidades para transferência de domínio, no entanto, ajustes ainda precisam ser feitos no algoritmo para tratamento de funções geradas que não são monotônicas.

Palavras-chave: Mapa de profundidade; Aprendizado Profundo; Detecção de Objetos 3D; Estimação Monocular de Profundidade.

Abstract

Depth information has great importance in various applications that require geometric information, such as autonomous vehicles, robotic navigation, augmented reality, and artificial intelligence content generation. Depth maps can be acquired with sensors or through estimation by passive methods. The development of deep learning techniques has made monocular depth estimation a feasible task, where a single RGB image is used as input to a neural network system. This work aims to perform a comparative analysis of state-of-the-art monocular depth estimators in terms of metrics found in the literature, domain transfer from relative to metric, and application in practical scenarios. Preliminary results have demonstrated the effectiveness of the intensity transform method for domain transfer, however, adjustments still need to be made to the algorithm to address generated functions that are not monotonic.

Keywords: Depth Map; Deep Learning; 3D Object Detection; Monocular Depth Estimation.

Listas de Figuras

1.1	Exemplo de Mapa de Profundidade.	2
3.1	Etapas do processamento digital de imagens que vai desde a aquisição de imagens até a identificação e descrição de objetos presente nelas.	12
3.2	Funções para ajuste da intensidade em imagens. (a) Método de ampliação de contraste, que aumenta a diferença entre os níveis de intensidade, destacando áreas mais escuras e mais claras. (b) Método de binarização, que converte a imagem em dois níveis distintos de intensidade, geralmente preto e branco, com base em um limiar definido.	14
3.3	Transformações de intensidade identidade, negativo, logarítmica, exponencial e por partes.	15
3.4	Diagrama de um perceptron, ilustrando uma abordagem simplificada baseada na estrutura e função de um neurônio biológico.	17
3.5	Exemplo clássico de uma arquitetura de CNN, mostrando a imagem de entrada, camadas convolucionais, camadas de <i>pooling</i> , camadas densas e saída.	18
3.6	Rede neural convolucional LeNet-5	20
3.7	Exemplo de Mapa de Profundidade.	21
3.8	Princípio de funcionamento de câmeras ToF.	22
3.9	Câmera LiDAR Intel RealSense L515.	22
3.10	Microsoft Kinect.	23
3.11	Um sistema de correspondência estéreo genérico.	24
3.12	Um sistema de estimativa monocular de profundidade por aprendizado profundo.	24
3.13	Um exemplo de ilusão de ótica	25

4.1	Exemplo do <i>dataset</i> NYU Depth v2	27
4.2	Exemplo do <i>dataset</i> KITTI de estimativa de profundidade	28
4.3	Exemplo do <i>dataset</i> Sintel	29
4.4	Exemplo do <i>dataset</i> ETH3D	30
4.5	Exemplo do <i>dataset</i> DIODE	31
4.6	Diagrama do método de transferência de domínio	33
4.7	Arquitetura da rede D4LCN.	35
5.1	Exemplo 1. (a) Imagem RGB, (b) Resultado relativo do modelo, (c) Transformação de intensidades de transferência de domínio relativo para métrico, (d) Em ordem: Mapa de profundidade do <i>dataset</i> , mapa de profundidade do <i>dataset</i> corrigido e mapa de profundidade estimado e corrigido com transferência de domínio.	38
5.2	Exemplo 2. (a) Imagem RGB, (b) Resultado relativo do modelo, (c) Transformação de intensidades de transferência de domínio relativo para métrico, (d) Em ordem: Mapa de profundidade do <i>dataset</i> , mapa de profundidade do <i>dataset</i> corrigido e mapa de profundidade estimado e corrigido com transferência de domínio.	39
5.3	Exemplo 3. (a) Imagem RGB, (b) Resultado relativo do modelo, (c) Transformação de intensidades de transferência de domínio relativo para métrico, (d) Em ordem: Mapa de profundidade do <i>dataset</i> , mapa de profundidade do <i>dataset</i> corrigido e mapa de profundidade estimado e corrigido com transferência de domínio.	40
5.4	Exemplo 4. (a) Imagem RGB, (b) Resultado relativo do modelo, (c) Transformação de intensidades de transferência de domínio relativo para métrico, (d) Em ordem: Mapa de profundidade do <i>dataset</i> , mapa de profundidade do <i>dataset</i> corrigido e mapa de profundidade estimado e corrigido com transferência de domínio.	41
A.1	Protocolo de ajuste fino do Marigold.	49
A.2	Processo de inferência do Marigold.	50

Lista de Tabelas

4.1	Características dos datasets utilizados no trabalho	27
6.1	Cronograma com as atividades realizadas para o desenvolvimento da pesquisa do ano de 2023	42
6.2	Cronograma com as atividades realizadas e pretendidas para o desenvolvimento da pesquisa do ano de 2024 e Janeiro de 2025.	43

Sumário

1	Introdução	1
1.1	Contextualização	1
1.2	Motivação e Justificativa	3
1.3	Objetivos	5
1.3.1	Objetivo Geral	5
1.3.2	Objetivos Específicos	5
1.4	Resultados Esperados	5
2	Trabalhos Relacionados	7
3	Fundamentação Teórica	10
3.1	Processamento Digital de Imagens	10
3.1.1	Transformação de Intensidade	13
3.2	Aprendizado Profundo	15
3.2.1	Redes Neurais Convolucionais	17
3.3	Estimação de Profundidade	19
4	Materiais e Métodos	26
4.1	Datasets	26
4.1.1	NYUv2	27
4.1.2	KITTI	28
4.1.3	SINTEL	29
4.1.4	ETH3D	30

Sumário

4.1.5	DIODE	30
4.2	Protocolo de Avaliação	31
4.3	Método de Transformação de Intensidades (pós-processamento)	32
4.4	Análise com Aplicação	33
4.4.1	Aplicação: Detecção de objetos 3D	34
4.5	Considerações Metodológicas	35
5	Resultados Preliminares e Discussões	37
6	Cronograma	42
Referências		44
Apêndices A - Apêndice A		49
A.1	Modelos Estudados	49
A.1.1	Marigold	49

Capítulo 1

Introdução

1.1 Contextualização

Informação de profundidade é uma das representações mais úteis para o entendimento de ambientes físicos (LASINGER et al., 2019) (ZHOU; KRÄHENBÜHL; KOLTUN, 2019). São também uma parte importante da caracterização de relações geométricas de uma determinada cena. As imagens de profundidades (ou mapas de profundidade) desempenham um papel importante em uma série de aplicações que envolvem visão computacional (EIGEN; PUHRSCH; FERGUS, 2014). Entre elas, podemos citar: compreensão de cenas (JARITZ et al., 2018), veículos autônomos (SONG et al., 2021), navegação de robôs (MA et al., 2019) navegação de VANTs, (PADHY et al., 2023) fazendas inteligentes (FARKHANI et al., 2019), e realidade aumentada (DU et al., 2020).

Os mapas de profundidade representam as distâncias de cada ponto (ou pixel) numa cena física em relação ao eixo do dispositivo de captura. Podem ser representados por imagens em escala de cinza, com as cores dos pixels sendo proporcionais à distância, com cinzas mais claros para objetos mais próximos e tons mais escuros para pontos mais afastados (e vice-versa) (DOURADO; PEDRINO, 2020). A Figura 1.1 mostra um exemplo com uma imagem RGB de uma cena, um mapa de profundidade em escala de cinza e o mesmo mapa colorizado para visualização.

Para capturar tais imagens geralmente são empregadas câmeras RGB-D, que podem prover tanto informação de profundidade quanto imagens coloridas da cena. Entre suas tecnologias mais comuns, são encontrados diversos tipos de aquisição que podem ser baseados em visão estereoscópica, que trabalha com múltiplos ângulos de visão, sensores *Time-of-Flight* (ToF) que emprega projeção de lasers infravermelhos (IR) estruturados e

Figura 1.1: Exemplo de Mapa de Profundidade.



Fonte: Dataset NYU Depth V2, do trabalho de Silberman et al. (2012)

técnicas mais precisas como o LiDAR (*Light Detection and Ranging*) (CASTELLANO; TERRERAN; GHIDONI, 2023).

Problemática

Garantir a correta representação dos mapas em escala de pixel é de considerável importância para as tarefas que dependem de profundidade e que requerem um alto grau de segurança e confiabilidade dos dados, como veículos autônomos ou navegação de drones por exemplo. A tecnologia LiDAR é a alternativa com implementação mais confiável entre as que foram citadas, no entanto, ressalta-se que nem o LiDAR e nem câmeras RGB-D convencionais produzem mapas completos e densos. No caso do LiDAR, são produzidos mapas esparsos (até 95% de esparsidade). Mapas esparsos são imagens de profundidade em que uma grande quantidade de pixels espalhados não possuem leitura válida. Além disso, sensores LiDAR possuem alto custo financeiro. No caso de câmeras RGB-D ou sensores ToF, são produzidos mapas com partes faltantes em determinadas superfícies ou bordas (HU et al., 2012).

Considerando as limitações impostas por métodos ativos de aquisição de profundidade, surge a possibilidade de inferir um mapa de profundidade denso e completo de uma cena a partir de uma ou mais imagens RGB, processo conhecido como estimativa de profundidade (*Depth Estimation - DE*) (RAJAPAKSHA et al., 2024). Quando duas imagens de câmeras diferentes são utilizadas para obter-se a informação de profundidade, denomina-se *Stereo Matching (SM)*. No entanto, métodos baseados em imagens estéreo requerem processos complexos de calibração¹ e alinhamento (DONG et al., 2022).

Uma das formas de gerar um mapa de profundidade de forma prática é utilizando imagens geradas por câmeras monoculares. O problema da estimativa monocular de pro-

¹São parâmetros de calibração referentes à câmera estéreo: Distorção de lente, translação, rotação, inclinação e guinada (DING et al., 2011)

fundidude (*Monocular Depth Estimation - MDE*) tem por objetivo inferir o mapa de profundidade através de uma única imagem RGB. Esse problema é considerado mal-posto devido à ausência de informação geométrica na projeção da cena 3D para a imagem 2D. No entanto, os avanços nas tecnologias de *Deep Learning - DL* e visão computacional tornaram factível e conveniente o uso de MDE para estimar mapas de profundidade densos e completos (SPENCER et al., 2024) (RAJAPAKSHA et al., 2024).

Ao longo dos anos, houveram diversas pesquisas científicas abordando o tema de estimação monocular de profundidade utilizando um vasto número de técnicas e metodologias dentro do universo do aprendizado profundo, empregando desde redes neurais convolucionais (KOPF; RONG; HUANG, 2021), estruturas *encoder-decoder* (GODARD et al., 2019), mistura de bases de dados em grande escala em modos diferentes (LASINGER et al., 2019), transformadores de visão (BIRKL; WOFK; MÜLLER, 2023), modelos de difusão (KE et al., 2024), e treinamento utilizando dados reais pseudo-rotulados em larga escala (YANG et al., 2024b).

Neste cenário, este trabalho propõe uma análise comparativa entre os diversos modelos de estimação monocular de profundidade relativa baseados em aprendizado profundo através da abordagem quantitativa, utilizando métricas e *benchmarks* presentes na literatura, abordagem qualitativa e através de uma aplicação.

1.2 Motivação e Justificativa

Os recentes avanços na área de MDE propiciaram a aquisição de informação de profundidade de maneira mais precisa e rápida, dessa forma, também favorecem indiretamente as aplicações que dependem desse tipo de dado, como reconstrução 3D, navegação e veículos autônomos. Além disso, devido à facilidade de implementação dessas técnicas, também podemos citar melhoramentos em aplicações mais modernas como conteúdo gerado por Inteligência Artificial (YANG et al., 2024b).

Reconstruir estruturas 3D a partir de imagens e informação geométrica prévia é um dos tópicos amplamente investigados pela ciência nos últimos anos (ZHAO et al., 2020). A técnica *Simultaneous Localization and Mapping* (SLAM) consiste planejar e controlar os movimentos de um robô por meio da construção de um mapa espacial do ambiente ao seu redor e obter a sua localização relativa, relacionando a área da visão computacional e a robótica através reconstrução de ambientes 3D e sensores de imagem (PLACED et al., 2023) (STACHNISS; LEONARD; THRUN, 2016). Métodos de SLAM que objetivam

a fusão de característica de mapas de profundidade obtidos através de sensores em movimento tiveram um aumento de popularidade em tempos recentes, visto que podem ser empregados para navegação e mapeamento de diversos tipos de dispositivos autônomos, como drones e robôs, além das aplicações em realidade aumentada e computação gráfica (TATENO et al., 2017).

Detectores de objetos com imagens tem sido aplicados em diversas áreas, como veículos autônomos e visão robótica, em que os sistemas necessitam estimar a localização de pedestres, veículos ou outros obstáculos. Devido à ausência informação de profundidade em imagens 2D, algumas aplicações exigem que a detecção seja feita no espaço tridimensional. O problema da detecção 3D consiste em estimar os vértices das caixas tridimensionais que contenham determinados objetos (HU et al., 2022a). Uma das formas de adquirir a informação de profundidade necessária é através de LiDAR, entretanto, segundo Wu et al. (2022), métodos de detecção 3D baseados somente em informação de LiDAR sofrem com a esparsidade dos dados, além do alto custo financeiro do equipamento. De acordo com Ding et al. (2020) uma alternativa mais desejável é o uso de câmeras monoculares. Mapas de profundidade podem ser empregados de duas formas: Transformando os mapas em representação pseudo-LiDAR, ou em sistemas multi-modais em conjunto com a informação RGB. A performance dos métodos de detecção 3D que utilizam informação de profundidade dependem da qualidade e densidade dos mesmos, portanto, as novas tecnologias de estimação monocular profundidade podem contribuir significativamente para este fim.

Segundo Khan, Salahuddin e Javidnia (2020), é evidente o potencial da estimação monocular para problemas de aplicações que envolvam informação de profundidade. Para que as diversas aplicações apresentadas possam funcionar de forma eficaz, é necessário garantir a correta representação dos mapas que são utilizados como entrada dos sistemas. Dado os recentes avanços nos modelos de MDE, tornou-se possível gerar mapas de profundidade de alta qualidade, com fineza de detalhes e de forma rápida, o que cobre as desvantagens de outros tipos de aquisição.

Devido aos melhoramentos possíveis nas aplicações que utilizam informação de profundidade, faz-se necessária a avaliação dos diversos modelos e técnicas de estimação monocular de profundidade do estado da arte, analisando tanto a sua capacidade de gerar mapas densos e precisos quanto à sua empregabilidade em aplicações práticas que envolvam informação de profundidade, que é a proposta do presente trabalho.

1.3 Objetivos

1.3.1 Objetivo Geral

Este trabalho possui como objetivo geral a análise comparativa de estimadores monoculares de profundidade capazes de produzir informação de profundidade densos para qualquer imagem.

1.3.2 Objetivos Específicos

- Estudar e escolher *datasets* que contenham imagens apropriadas para teste considerando variedade de ambientes e métodos de captura.
- Estudar modelos de estimação monocular de profundidade relativa baseados em aprendizado profundo do estado da arte.
- Analisar e escolher entre os modelos estudados para implementação e testes visando eficiência, rapidez, robustez ou outros parâmetros a escolher.
- Avaliar o desempenho perante métricas utilizadas na literatura para comparação entre os modelos no espaço relativo e métrico.
- Implementar método de pós-processamento para transferência do domínio relativo para métrico baseado em transformação de intensidade.
- Avaliar qualitativamente os resultados.
- Implementar e avaliar aplicações com os mapas de profundidade gerados a partir dos estimadores do estado da arte.

1.4 Resultados Esperados

Este trabalho pretende realizar uma comparação quantitativa e qualitativa entre mapas gerados por estimadores monoculares de profundidade, portanto, almeja-se evidenciar as qualidades dos modelos através das métricas de avaliação, expondo suas vantagens em relação às outras técnicas de estimação e captura de informação de profundidade.

Além disso, espera-se explorar as técnicas de transformação de intensidades para transferir mapas relativos inferidos pelos modelos do domínio relativo para o métrico, mensurando sua aplicabilidade por meio da comparação com estimadores métricos.

Dado que os modelos estimadores de profundidade relativa do estado da arte fornecem mapas de profundidades precisos e com objetos bem definidos, espera-se que as aplicações que utilizam informação de profundidade tenham o seu desempenho melhorado.

Os resultados esperados deste trabalho servirão como base para futuras pesquisas aplicadas do laboratório PAVIC (Pesquisa Aplicada em Visão e Inteligência Computacional) nas temáticas relacionadas, por exemplo, navegação robótica, realidade aumentada, metaverso e outras técnicas que utilizem informação tridimensional.

Capítulo 2

Trabalhos Relacionados

No passado, a tarefa de Estimação Monocular de Profundidade não era abordada de forma direta. Um exemplo deste cenário é o trabalho de Hoiem, Efros e Hebert (2005), em que o objetivo é reconstruir uma cena 3D em um ambiente virtual através de uma única imagem RGB. Apesar da finalidade não ser a construção de um mapa de profundidade, a reconstrução 3D de uma cena é diretamente ligada à informação de profundidade, portanto, esse trabalho é creditado em revisões bibliográficas do tema (MERTAN; DUFF; UNAL, 2022). É considerado que um ambiente externo consista de elementos fixos, o céu, um plano de chão e objetos verticais saindo deste plano. É realizada uma classificação de superpixels nas classes através de características pré-selecionadas manualmente, e os objetos são colocados em 3D através das mesmas.

Ainda nos primórdios da MDE, um dos primeiros trabalhos a se propor a estimar um mapa de profundidade métrico de uma única imagem RGB é o de Saxena, Chung e Ng (2005). Filtros manualmente projetados são aplicados em pequenos pedaços de uma imagem de entrada para extrair características. Para cada parte, um valor de distância é estimado. Os filtros são então aplicados em múltiplas escalas para levar em consideração as pistas visuais globais e de partes adjacentes. Pesos maiores são atribuídos às características dos pedaços que ficam nas mesmas colunas, baseado na premissa de que as estruturas dos objetos observados são em sua maioria, verticais. Além disso, um modelo baseado em Campos Aleatórios de Markov (*Markov Random Field* - MRF) é treinado de maneira supervisionada para estimar a profundidade a partir das características.

Algum tempo depois, outro trabalho publicado por Saxena, Sun e Ng (2008), adicionou um pressuposto pertinente ao estado da arte de MDE, que uma cena consiste de várias pequenas superfícies planas e a orientação e localização 3D dessa superfície podem ser utilizadas para calcular sua profundidade. Esse pressuposto é utilizado até hoje em

motores gráficos que criam modelos de objetos complexos através de malhas triangulares. Novamente, é utilizado um modelo baseado em MRF treinado de maneira supervisionada. As características são obtidas através de filtros manualmente projetados e a contextualização global é considerada através de superpixels adjacentes.

Considerando o desenvolvimento do aprendizado profundo à época, Eigen, Puhrsich e Fergus (2014) introduziu o uso de redes neurais convolucionais para a tarefa de MDE, superando as técnicas anteriores. O problema foi formulado como um método de regressão com aprendizado supervisionado de um conjunto de duas redes neurais. A primeira é responsável por uma estimativa grosseira do mapa de profundidade. Sendo composta por camadas convolucionais totalmente conectadas, possui a imagem toda como campo receptivo, utilizando melhor o contexto global, a custo de um grande custo computacional. A segunda rede neural é totalmente convolucional e possui como entrada o mapa da rede anterior, e tem como finalidade o ajuste fino do mapa de profundidade, operando através de filtros locais. Além disso foi utilizada uma função de perda com invariância em escala no espaço logarítmico.

O modelo de estimativa de profundidade denominado MiDaS foi apresentado por Ranftl et al. (2020), a pesquisa possui como principal contribuição o desenvolvimento de protocolos de mesclagem de conjuntos de dados de profundidade mesmo que suas anotações não sejam compatíveis. O núcleo dessa abordagem consiste em uma função que é invariante em escala e alcance em um processo de aprendizado multi-objetivo combinando dados de diferentes fontes. A arquitetura da rede consiste em uma estrutura baseada em ResNet em multi escala. Outra contribuição foi o emprego de filmes 3D para composição da base de dados de treinamento em larga escala, apesar de não apresentar anotação de profundidade, foi utilizado *stereo matching* para rotulagem do conjunto de treino.

Em (BIRKL; WOFK; MÜLLER, 2023), foi dado sequência ao trabalho anterior e apresentou-se novos modelos com diferentes codificadores, sendo a maioria baseado em transformadores de visão, fundamentando-se nos recentes avanços e resultados desta tecnologia. Os autores realizam uma comparação entre os transformadores BEiT, Swin, SwinV2, Next-ViT e LeViT considerando performance versus tempo de execução. Os resultados alcançados foram melhores que a versão puramente convolucional do modelo para maior parte dos modelos.

Em (KE et al., 2024) foi apresentado um protocolo de *fine tuning* de modelos de difusão latente pré-treinados para estimativa relativa de profundidade sob qualquer circunstância. O protocolo, chamado de Marigold, contribui com o estado da arte sendo um

dos trabalhos que investigou o uso de bases de dados de imagens sintéticas para treinamento, dado que estas não estariam propensas a erros de captura. Utilizou-se um modelo de difusão estável pré-treinado, e o ajuste do modelo é realizado utilizando uma função objetivo calculada no espaço latente entre a saída da U-Net e o ruído inicial. Outra contribuição do trabalho foi a aplicação de ruído retificado em multi-resolução para o processo de difusão.

Na temática do uso de dados não rotulados em larga escala, os autores Yang et al. (2024a) construíram um modelo fundacional para MDE capaz de produzir imagens de profundidade em alta qualidade sob quaisquer circunstâncias, chamado de *Depth Anything*. O uso de dados não rotulados é fundamentado em três principais vantagens: fácil e barata aquisição, diversidade de cenas e fácil rotulagem. O protocolo apresentado consiste em dois modelos. O primeiro, chamado de "professor", é treinado em um conjunto menor de imagens de profundidade com anotações. Em seguida, o modelo professor é usado para anotar os 62 milhões de imagens não rotuladas adquiridas de oito datasets públicos. O conjunto de dados final é por fim, utilizado para treinamento do modelo final. A arquitetura é baseada no modelo DINov2 com pré-treinamento para segmentação semântica, o que auxilia ainda mais no processo de MDE.

Prosseguindo o trabalho anterior, em (YANG et al., 2024b) é apresentado uma versão aprimorada do sistema *Depth Anything*, que foca não somente em melhorar as métricas, mas em produzir previsões mais robustas e com alto grau de detalhes. Esses objetivos são alcançados através do escalonamento do modelo professor e o uso de imagens exclusivamente sintéticas no seu processo de treinamento.

As recentes pesquisas da área de estimativa monocular de profundidade apresentaram uma variedade de técnicas robustas baseadas em aprendizado profundo. Neste cenário, este trabalho se propõe a realizar comparações entre os modelos do estado da arte por meio de métricas e aplicações.

Capítulo 3

Fundamentação Teórica

3.1 Processamento Digital de Imagens

Processamento digital de imagens segundo Jain (1989), consiste na manipulação de imagens em formato digital utilizando algoritmos para extrair informações, melhorar a qualidade ou transformar dados visuais. Esta área engloba diversas técnicas que permitem o tratamento de imagens capturadas por dispositivos digitais, como câmeras e scanners, visando a otimização de aspectos como contraste, nitidez e remoção de ruído (GONZALEZ; WOODS, 2010). Além de aprimorar a percepção visual, essas técnicas são essenciais para a análise automática de imagens, facilitando a identificação e classificação de objetos, a medição de propriedades geométricas e a extração de padrões (RUSS, 2006).

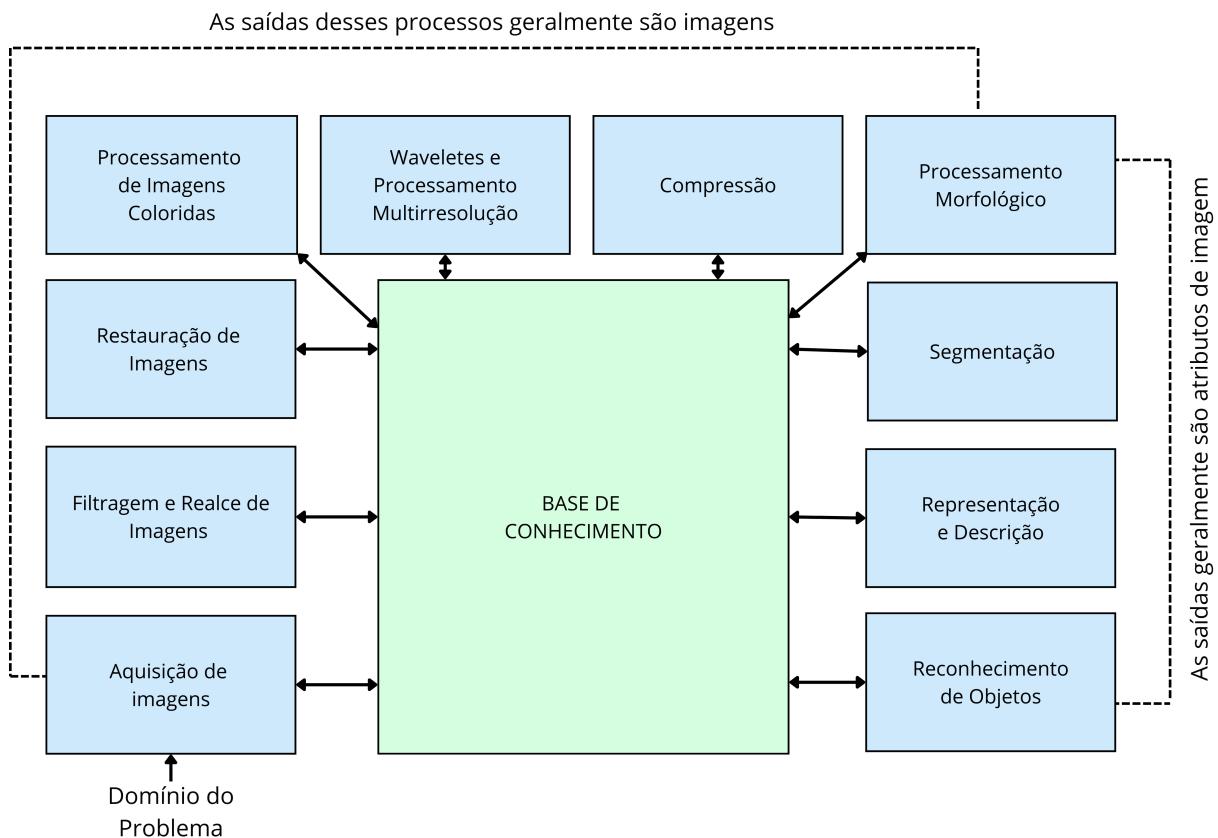
No processamento digital de imagens, os processos geralmente são abordados em três níveis distintos, cada um com um papel específico na análise e interpretação das imagens. O nível baixo trata de manipulações mais primitivas, realizando operações fundamentais como filtragem, aprimoramento de contraste e remoção de ruído. O nível médio foca na segmentação e extração de características, onde a imagem é dividida em regiões de interesse e características relevantes são identificadas e descritas. Finalmente, o nível alto envolve a interpretação e reconhecimento dos dados processados, onde algoritmos são aplicados para classificar objetos, reconhecer padrões e realizar decisões baseadas em informações extraídas dos níveis anteriores. Cada um desses níveis do processamento de imagens contribui de maneira distinta para a análise visual, formando uma cadeia de processamento que vai desde a manipulação básica até a interpretação complexa (GONZALEZ; WOODS, 2010).

Ainda segundo Gonzalez e Woods (2010) o processamento digital de imagens envolve

uma série de passos que vão desde a aquisição até interpretação das imagens. Como pode ser visto na Figura 3.1, essas etapas incluem:

1. **Aquisição de Imagem:** Este estágio trata da obtenção de imagens, que podem ser capturadas por dispositivos digitais ou provenientes de arquivos digitais já existentes. Neste processo, podem ser realizados ajustes iniciais, como modificar o tamanho da imagem.
2. **Filtragem e realce de imagens:** Nesta fase, a imagem é manipulada para adequá-la a uma aplicação específica. Em alguns casos, como em imagens médicas, a melhoria pode não ser a abordagem mais apropriada.
3. **Restauração de imagens:** Este passo busca aprimorar a qualidade visual das imagens através de métodos baseados em modelos matemáticos ou probabilísticos para compensar a degradação da imagem.
4. **Processamento de Imagens em Cores:** Com o aumento do uso de imagens digitais na web, o processamento de imagens coloridas tornou-se fundamental. Trabalhar com cores facilita a extração de características relevantes de uma imagem.
5. **Processamento em Multirresolução:** Este método envolve a representação de uma imagem em diferentes níveis de resolução, permitindo uma análise detalhada em várias escalas.
6. **Compressão de Imagem:** Nesta fase, são aplicadas técnicas para armazenar imagens de maneira mais eficiente ou reduzir a largura de banda necessária para sua transmissão.
7. **Processamento Morfológico:** O processamento morfológico foca na extração e análise dos componentes da imagem para descrever suas formas.
8. **Segmentação:** A segmentação é um dos aspectos mais desafiadores no processamento digital de imagens. Ela consiste em dividir a imagem em partes ou objetos distintos.
9. **Representação e Descrição de Características:** Este processo, também conhecido como seleção de atributos, visa extrair características que forneçam informações quantitativas ou qualitativas para distinguir entre diferentes tipos de objetos.

Figura 3.1: Etapas do processamento digital de imagens que vai desde a aquisição de imagens até a identificação e descrição de objetos presentes nelas.



Adaptado de Gonzalez e Woods (2018)

10. **Identificação de Objetos:** O reconhecimento de objetos é o processo de atribuir rótulos a elementos presentes em uma imagem, como classificar um objeto como um "carro".
11. **Base de conhecimento:** Envolve o entendimento do domínio do problema, incluindo informações detalhadas sobre áreas específicas na imagem onde se espera encontrar dados relevantes.

O processamento digital de imagens é amplamente aplicado em áreas como a medicina, para a análise de imagens de diagnóstico, na indústria, para o controle de qualidade de produtos, na segurança, para reconhecimento facial e monitoramento e entre outros (GONZALEZ; WOODS, 2010).

3.1.1 Transformação de Intensidade

As transformações de intensidade são técnicas no processamento digital de imagens focadas na manipulação direta dos valores de intensidade dos pixels. Elas operam individualmente em cada pixel, possibilitando ajustes de contraste, brilho e outros atributos de uma imagem. O objetivo principal dessas transformações é modificar a aparência visual da imagem para realçar características específicas ou preparar a imagem para análises posteriores (GONZALEZ; WOODS, 2010).

As principais funções de transformação de intensidade incluem:

Transformações Lineares: Essas funções incluem transformações como o negativo e a identidade. A transformação de negativo inverte os valores de intensidade, enquanto a identidade mantém os valores de intensidade inalterados.

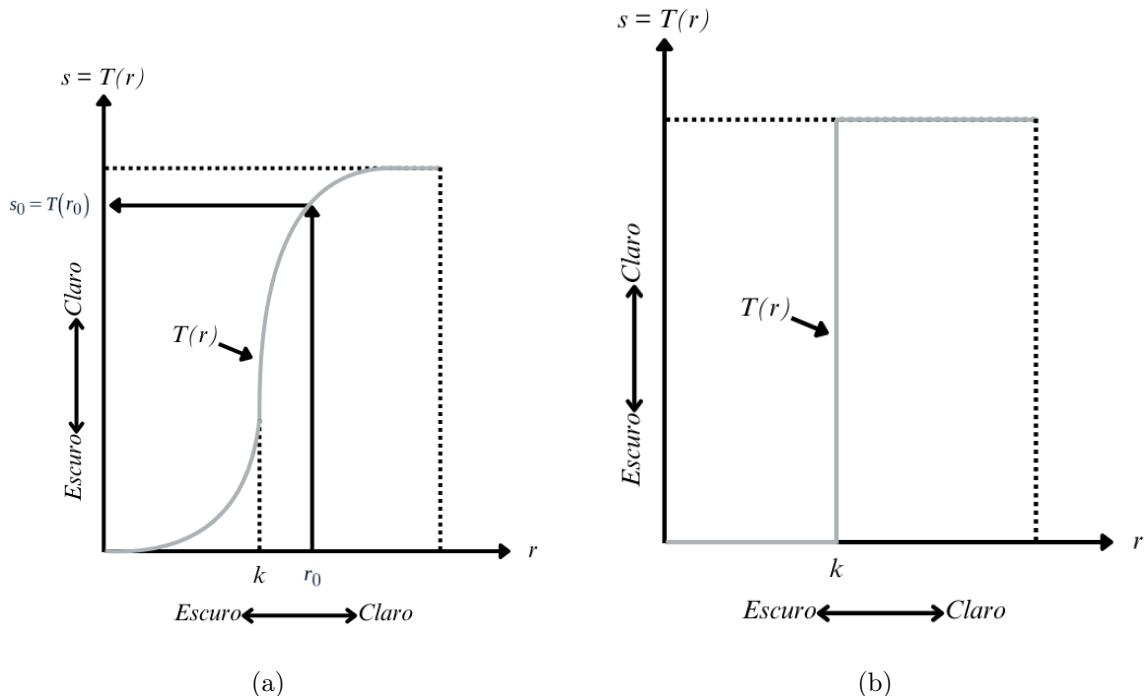
Transformações Logarítmicas: Essas funções utilizam operações logarítmicas para alterar a distribuição de intensidade. Transformações como o logaritmo e o logaritmo inverso são utilizadas para melhorar detalhes em áreas escuras da imagem ou para estender o alcance dinâmico.

Transformações de Potência: Utilizam funções de potência e raiz para ajustar o contraste e a gama da imagem. Transformações de n-ésima potência e n-ésima raiz são exemplos de como essas funções podem ajustar as características da imagem.

Assim, ao aplicar transformações de intensidade e técnicas de filtragem, é possível ajustar e melhorar imagens de maneira significativa, seja para visualização aprimorada ou para análises mais precisas. Considerando a transformação de intensidade ilustrada na Figura 3.3(a). Como mostra Gonzalez e Woods (2010), calculando essa transformação a cada pixel da imagem original f , gera-se uma nova imagem g com maior contraste. Nesta transformação, os valores de intensidade abaixo de um ponto k são escurecidos, enquanto os valores acima de k são clareados. Isso resulta em uma imagem com contraste ampliado, onde áreas escuras são mais densas e áreas claras são mais evidentes.

Na Figura 3.3(b), a transformação $T(r)$ resulta em uma imagem binária, onde os pixels são convertidos em apenas dois níveis de intensidade, dependendo de um limiar específico. Essa técnica, conhecida como limiarização, simplifica a imagem original em uma forma mais clara e destacada, com apenas duas cores, tipicamente preto e branco.

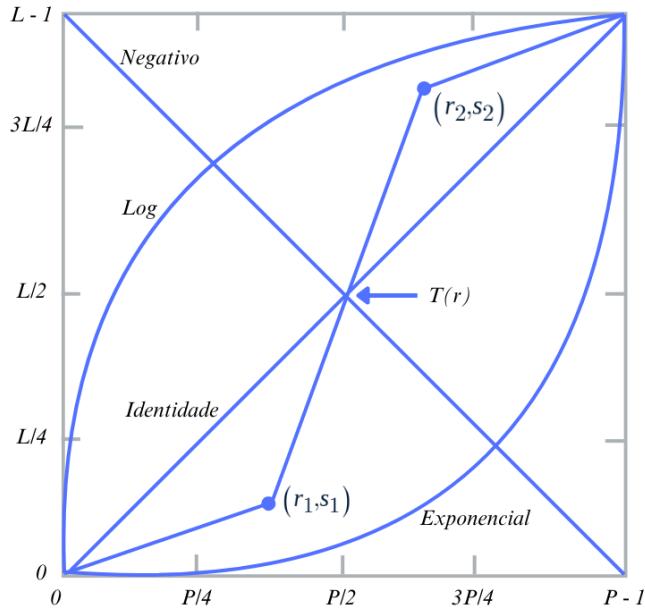
Figura 3.2: Funções para ajuste da intensidade em imagens. (a) Método de ampliação de contraste, que aumenta a diferença entre os níveis de intensidade, destacando áreas mais escuras e mais claras. (b) Método de binarização, que converte a imagem em dois níveis distintos de intensidade, geralmente preto e branco, com base em um limiar definido.



Adaptado de Gonzalez e Woods (2018)

Essas transformações são exemplos de como ajustes nos valores de intensidade podem alterar significativamente a aparência e a utilidade da imagem, dependendo do objetivo do processamento. A Figura 3.3 ilustra diversas transformadas de intensidades encontradas na literatura, como a logarítmica e exponencial, identidade, negativo e uma transformação linear por partes genérica $T(r)$.

Figura 3.3: Transformações de intensidade identidade, negativo, logarítmica, exponencial e por partes.



Adaptado de Gonzalez e Woods (2018)

3.2 Aprendizado Profundo

Deep Learning, ou Aprendizado Profundo, é uma subárea do aprendizado de máquina que se concentra em redes neurais artificiais com muitas camadas. Estas redes neurais são projetadas para simular o funcionamento do cérebro humano e aprender representações de dados em múltiplos níveis de abstração (GOODFELLOW; BENGIO; COURVILLE, 2016). As redes neurais profundas, são compostas por múltiplas camadas de neurônios artificiais. Cada camada da rede realiza uma transformação não linear sobre os dados, permitindo à rede aprender representações complexas e hierárquicas dos dados de entrada (HAYKIN, 2001).

O campo das redes neurais evoluiu significativamente desde suas primeiras iterações. Um marco importante nesse desenvolvimento foi o trabalho de Rosenblatt (1958), que introduziu o perceptron na década de 1960. Essa técnica inicial mostrou que, sob certas condições, um perceptron poderia aprender a classificar dados linearmente separáveis, mas enfrentava limitações com problemas mais complexos. A real revolução no campo

das redes neurais veio com o desenvolvimento de redes neurais de múltiplas camadas. Em 1986, Rumelhart, Hinton e Williams (1988) introduziram o algoritmo de retropropagação, também conhecido como a regra delta generalizada. Este método permitiu o treinamento eficaz de redes neurais com várias camadas, superando as limitações dos percepitrons simples e proporcionando um avanço significativo no desempenho e na aplicabilidade das redes neurais.

O perceptron é uma das estruturas fundamentais das redes neurais, representando o modelo mais simples de um neurônio artificial. Introduzido por Rosenblatt (1958), o perceptron é capaz de realizar classificações binárias, distinguindo entre duas classes distintas com base em dados de entrada.

Um perceptron funciona através de uma série de operações matemáticas simples. Inicialmente, cada entrada é multiplicada por um peso, que é um valor numérico associado a essa entrada. Esses pesos são ajustados durante o processo de treinamento para que o perceptron aprenda a mapear as entradas para as saídas desejadas. A soma ponderada dessas entradas é então calculada e passada por uma função de ativação, que decide pela ativação do neurônio. Tradicionalmente, a função de ativação usada em um perceptron simples é a função degrau, que retorna um valor binário de 0 ou 1 (HERTZ, 2018). Esse processo pode ser formalmente expresso pela Equação 3.1:

$$y = f \left(\sum_{i=1}^n w_i x_i + b \right) \quad (3.1)$$

Onde y é a saída do perceptron, x_i são as entradas, w_i são os pesos associados, b é o bias (um termo adicional que permite que o modelo ajuste a função de ativação, mesmo com todas as entradas em zero), e f é a função de ativação.

A Figura 3.4 ilustra visualmente a estrutura de um perceptron, mostrando as entradas, pesos, soma ponderada, bias, função de ativação e uma saída.

O perceptron aprende a partir de um processo de ajuste de pesos, conhecido como regra de aprendizagem do perceptron. Durante o treinamento, o perceptron ajusta os pesos com base no erro da saída prevista em relação à saída desejada. Esse ajuste é feito através de um processo iterativo, onde o erro é propagado de volta através da rede e os pesos são atualizados para minimizar o erro, seguindo as equações 3.2 e 3.3:

$$w_i = w_i + \Delta w_i \quad (3.2)$$

$$\Delta w_i = \eta(d - y)x_i \quad (3.3)$$

onde Δw_i é o ajuste do peso, η é a taxa de aprendizado, d é a saída desejada, e y é a saída calculada pelo perceptron. A taxa de aprendizado é um parâmetro importante que controla o quanto rapidamente o modelo se adapta aos dados.

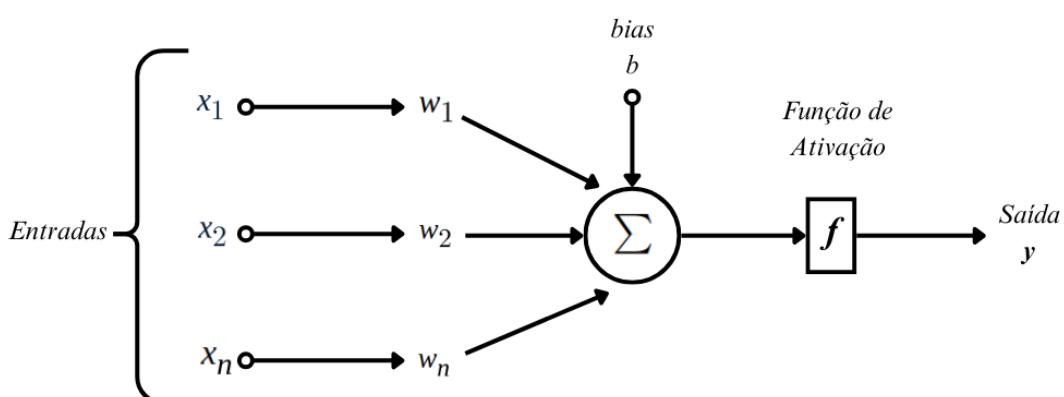
Embora o perceptron tenha sido um avanço significativo na época de sua introdução, possui limitações, especialmente em relação a problemas de classificação não linear. Por exemplo, não consegue resolver problemas representados pela porta lógica XOR, onde as classes não são linearmente separáveis. Essa limitação levou ao desenvolvimento de redes neurais mais complexas, como os perceptrons multicamadas (do inglês, *Multilayer Perceptron* ou MLP), que usam várias camadas de neurônios para aprender representações mais complexas dos dados (BISHOP; NASRABADI, 2006).

Segundo Hertz (2018) o perceptron permanece um conceito central na teoria das redes neurais, servindo como base para a compreensão de modelos mais sofisticados e sendo uma ferramenta educacional importante para introduzir os conceitos de aprendizado supervisionado.

3.2.1 Redes Neurais Convolucionais

As *Convolutional Neural Networks* (CNNs), ou Redes Neurais Convolucionais, são um tipo específico de rede neural artificial projetada para processar e analisar dados que possuem uma estrutura matricial, como imagens (O'SHEA; NASH, 2015). As CNNs têm

Figura 3.4: Diagrama de um perceptron, ilustrando uma abordagem simplificada baseada na estrutura e função de um neurônio biológico.



Adaptado de Haykin (2009)

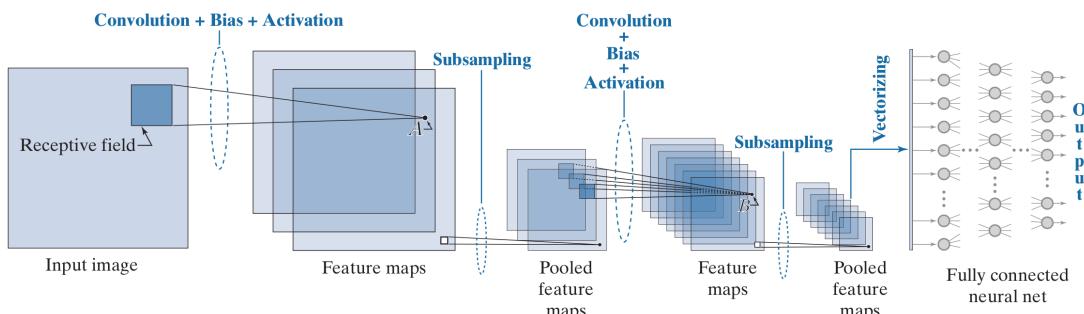
se mostrado altamente eficazes em tarefas de visão computacional, como reconhecimento de imagem, detecção de objetos e segmentação de imagem. Ao contrário das redes neurais tradicionais, que tratam os dados de entrada como um vetor unidimensional, as CNNs exploram as suas propriedades espaciais, permitindo que a rede aprenda características locais e complexas (NIELSEN, 2015).

A arquitetura das CNNs é composta por várias camadas de diferentes tipos, cada uma desempenhando um papel diferente no processamento e na extração de características dos dados de entrada. As principais camadas das CNNs incluem camadas convolucionais, camadas de *pooling* (ou subamostragem), e camadas totalmente conectadas (ou densas). Um exemplo clássico de uma arquitetura de CNN pode ser visualizado na Figura 3.5, que mostra a imagem de entrada passando por camadas convolucionais, *pooling*, e densas antes de chegar à saída (O'SHEA; NASH, 2015).

O funcionamento das CNNs pode ser explicado em várias etapas. Primeiramente, a camada convolucional aplica filtros (ou kernels) aos dados de entrada. Esses filtros são pequenas matrizes que percorrem a imagem, realizando operações de convolução. Cada filtro é treinado para detectar diferentes características, como bordas, texturas e padrões específicos. A saída dessa operação é um mapa de características, que representa a presença dessas características na imagem original. O objetivo principal das camadas convolucionais é aprender a identificar e abstrair características relevantes dos dados de entrada, preservando as relações espaciais entre os pixels (AGGARWAL et al., 2018).

Após a convolução, a saída passa por uma camada de pooling, que reduz a dimensionalidade dos mapas de características, mantendo as informações mais importantes. A camada de pooling mais comum é o max pooling, que seleciona o valor máximo de uma região específica do mapa de características. Esse processo ajuda a reduzir a quantidade

Figura 3.5: Exemplo clássico de uma arquitetura de CNN, mostrando a imagem de entrada, camadas convolucionais, camadas de *pooling*, camadas densas e saída.



Fonte: Gonzalez e Woods (2018)

de parâmetros e a computação na rede, além de tornar a detecção de características mais robusta a variações e deslocamentos na imagem (AGGARWAL et al., 2018).

Finalmente, as camadas densas recebem a saída das camadas convolucionais e de pooling e produzem a saída final da rede. Nessas camadas, cada neurônio está conectado a todos os neurônios da camada anterior, permitindo a combinação e a interpretação das características extraídas. As camadas densas são responsáveis por realizar a classificação final ou outras tarefas de saída (NIELSEN, 2015).

O treinamento das CNNs é realizado usando algoritmos de aprendizado supervisionado, como a retropropagação, em conjunto com otimizações como o gradiente descendente. Durante o treinamento, a rede ajusta os pesos dos filtros e das conexões entre os neurônios para minimizar a diferença entre a saída prevista e a saída desejada. A função de perda, que mede essa diferença, é calculada para cada exemplo do conjunto de treinamento, e os pesos são atualizados iterativamente para melhorar a precisão da rede (AGGARWAL et al., 2018).

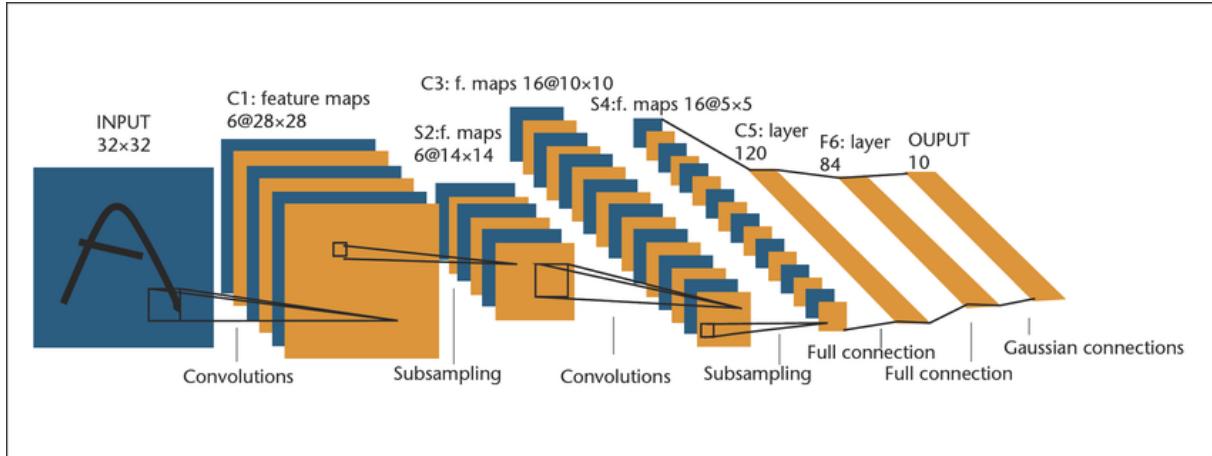
As CNNs são especialmente poderosas devido à sua capacidade de capturar hierarquias de características, desde características de baixo nível, como bordas e texturas, até características de alto nível, como formas e objetos inteiros. Isso é conseguido através do empilhamento de múltiplas camadas convolucionais e de pooling, permitindo que a rede aprenda representações cada vez mais complexas dos dados de entrada (GOODFELLOW; Bengio; COURVILLE, 2016).

Um exemplo clássico de CNN é a arquitetura LeNet-5, desenvolvida por Yann LeCun na década de 1990 para reconhecimento de dígitos manuscritos. Como pode ser visto na Figura 3.6, a LeNet-5 consiste em três camadas convolucionais seguidas por camadas de pooling, e termina com duas camadas totalmente conectadas. Essa arquitetura foi pioneira no uso de CNNs e demonstrou o potencial dessas redes para tarefas de visão computacional (LECUN et al., 1998).

3.3 Estimação de Profundidade

De acordo com Khan, Salahuddin e Javidnia (2020), o conceito de estimação de profundidade refere-se ao processo de preservar a informação tridimensional de uma cena a partir de uma imagem bidimensional capturada por câmeras. Mapas de profundidade são imagens que caracterizam esse conceito, e podem ser representadas por imagens em escala de cinza onde os tons dos pixels são definidos de acordo com sua distância correspondente na

Figura 3.6: Rede neural convolucional LeNet-5



Fonte: LeCun et al. (1998)

cena original (DOURADO; PEDRINO, 2020). A Figura 3.7 exibe uma imagem RGB e dois mapas de profundidade, um em escala de cinza e outro colorizado para visualização.

De acordo com Foley (2019), as distâncias aos objetos em uma cena são percebidas pelo sistema visual humano por meio de pistas visuais monoculares. Entre elas, podemos citar :

- **Oclusão**: Quando um objeto é parcialmente coberto por outro, então é considerado mais distante.
- **Perspectiva**: As linhas paralelas que se encontram no horizonte ajudam a perceber as distâncias dos objetos.
- **Tamanho**: Objetos maiores são percebidos como mais próximos, enquanto objetos menores são percebidos mais distantes.
- **Gradiente de textura**: A textura de uma superfície fica mais densa de acordo com o aumento da distância ao observador.
- **Perspectiva atmosférica**: Refere-se à percepção de desfoque em objetos mais distantes.
- **Luz e sombra**: Percepção de profundidade a partir das projeções das sombras dos objetos.
- **Altura**: Objetos próximos à linha do horizonte são percebidos mais distantes.

Sensores de profundidade estão cada vez mais embarcados em equipamentos amplamente difundidos como dispositivos de realidade aumentada e até mesmo em smartphones (DU et al., 2020), principalmente as câmeras ToF, pois são capazes de desempenhar de maneira satisfatória mesmo com baixa potência (BRANSCOMBE, 2018). De acordo com (XIE et al., 2021), a adoção de sensores de profundidade em smartphones tende a aumentar nos próximos anos, com diversas aplicações como tradução de linguagem de sinais (PARK; LEE; KO, 2021) e sistemas de navegação mobile para pessoas com deficiência visual (SEE; SASING; ADVINCULA, 2022).

Figura 3.7: Exemplo de Mapa de Profundidade.



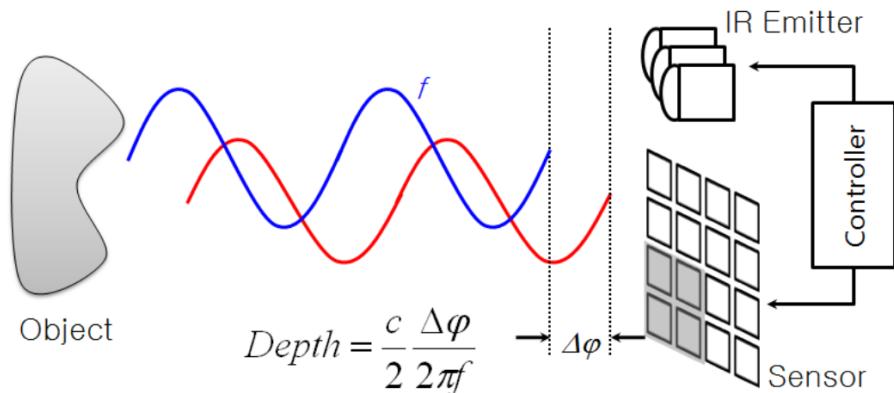
Fonte: Dataset NYU Depth V2, do trabalho de Silberman et al. (2012)

Diversos tipos de técnicas são atualmente empregadas para adquirir informação de profundidade. Os métodos de aquisição de imagens de profundidade são divididos em duas categorias. Métodos ativos de estimação de profundidade envolvem calcular as distâncias interagindo fisicamente com os objetos do ambiente. Ultrassom e sensores *Time of Flight* (ToF) por exemplo, utilizam da emissão de ondas com velocidade conhecida e medem o tempo de retorno a um receptor. Métodos passivos, por outro lado, envolvem a extração da informação através de processamento de imagens como visão estéreo e estimação monocular de profundidade (KHAN; SALAHUDDIN; JAVIDNIA, 2020).

De acordo com Hansard et al. (2012), o funcionamento de câmeras ToF consiste em um emissor e um receptor de pulso de luz, a distância percorrida pode ser calculada medindo o atraso de detecção ou o desvio de fase do pulso recebido, produzindo um mapa de profundidade, como exemplificado na Figura 3.8.

As câmeras ToF se dividem em dois tipos de acordo com o método de medição de distância. O primeiro, chamado de ToF ponto-a-ponto, utiliza um mecanismo de inclinação panorâmica para obter uma sequência temporal de medições do tempo de retorno de um laser infravermelho, essa técnica é conhecida como *Light Detection and Ranging* (LiDAR). Esse tipo de dispositivo é empregado com maior frequência em ambientes ex-

Figura 3.8: Princípio de funcionamento de câmeras ToF.



Fonte: (HANSARD et al., 2012)

ternos, por exemplo, sensores embarcados em veículos. Além disso, entre as vantagens desse tipo de tecnologia podemos citar a sua boa performance mesmo em cenários de baixa luminosidade ou em movimento e sua alta resolução e acurácia da informação adquirida (ZOLLHÖFER, 2019). Um exemplo de câmera LiDAR é o Intel RealSense L515, mostrado na Figura 3.9.

Figura 3.9: Câmera LiDAR Intel RealSense L515.



Fonte: Castellano, Terreran e Ghidoni (2023)

O segundo tipo é chamado de ToF modulado, em que uma câmera utiliza um pulso contínuo modulado em uma determinada frequência. A onda recebida é processada por um detector de desvio de fase. Um exemplo de dispositivo que usa desse mecanismo é o Microsoft Kinect (Figura 3.10) (ZOLLHÖFER, 2019).

De acordo com Szeliski (2022), o termo correspondência estéreo (do inglês, *Stereo*

Figura 3.10: Microsoft Kinect.

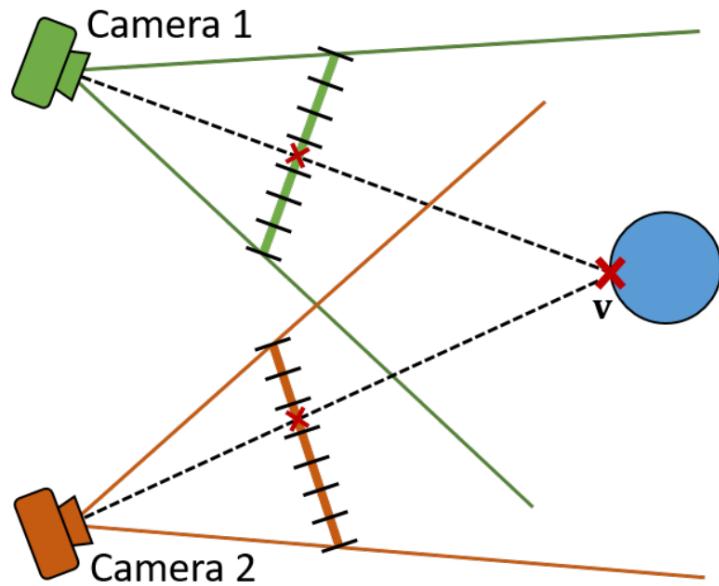


Fonte: Amos (2014)

Matching) refere-se ao processo de construir um modelo 3D de uma cena por meio da busca de pixels correspondentes em pelo menos duas imagens 2D. Técnicas estereoscópicas de estimação de profundidade usam câmeras com duas lentes (ou duas câmeras separadas) em perspectivas diferentes, resultando em uma imagem de disparidade, que pode ser processada em um mapa de profundidade (CASTELLANO; TERRERAN; GHIDONI, 2023). A técnica baseia-se na premissa de que o sistema visual humano percebe as profundidades através das diferenças entre as imagens capturadas pelos olhos esquerdo e direito. O esquema de configuração básica de um sistema de correspondência estéreo é ilustrado na figura 3.11. Segundo Lahiri, Ren e Lin (2024), devido ao desenvolvimento da área do aprendizado profundo, atualmente, a técnica de estimação de profundidade por meio de visão estéreo é formulada por uma tarefa de aprendizado não-supervisionado.

Segundo (CASTELLANO; TERRERAN; GHIDONI, 2023), cada uma das técnicas de aquisição de imagens de profundidade possui lados negativos que podem impactar o seu emprego em aplicações práticas. Por exemplo, as câmeras ToF podem sofrer com invalidação de pixels em cantos ou bordas dos objetos devido a saturação do sinal infravermelho, outro tipo de área afetada são superfícies reflexivas ou metálicas (ZOLLMÖFER, 2019). De acordo com Zhang et al. (2022) quando capturadas em ambientes internos, tais imagens podem conter até 50% de dados faltantes. Dados de profundidade adquiridos por LiDAR não são densos, podendo chegar até 95% de esparsidade, além do equipamento apresentar custo financeiro e energético elevado (KHAN; SALAHUDDIN; JAVIDNIA, 2020) (HU et al., 2022b). Uma das desvantagens presentes nos sistemas estéreo é a presença de ruído devido a má triangulação de objetos com características pouco distinguíveis (CASTELLANO; TERRERAN; GHIDONI, 2023).

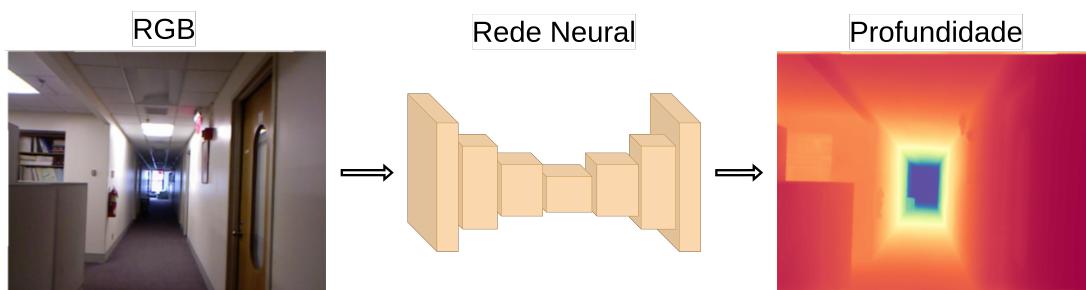
Figura 3.11: Um sistema de correspondência estéreo genérico.



Fonte: Zollhöfer (2019)

Compondo a lista de métodos passivos, a estimação de profundidade monocular refere-se ao processo de regredir um mapa de profundidade denso a partir de uma única imagem de câmera monocular (BIRKL; WOFK; MÜLLER, 2023). Segundo Ranftl et al. (2020), este ainda é um problema desafiador, visto que envolve a compreensão de pistas visuais, contexto de cenas e conhecimento prévio dos objetos para deduzir as relações geométricas da cena, o que implica na necessidade de técnicas baseadas em aprendizado. Com o desenvolvimento da área da inteligência artificial, técnicas mais robustas começaram a ser empregadas para estimação de profundidade.

Figura 3.12: Um sistema de estimação monocular de profundidade por aprendizado profundo.



Elaborado pelo autor.

Segundo Mertan, Duff e Unal (2022) um dos desafios que são intrínsecos ao problema de estimação monocular de profundidade é a ambiguidade. A partir de uma única imagem

RGB podem haver inúmeros possíveis resultados 3D, ou seja, é caracterizado um problema de mapeamento onde cada entrada pode gerar muitas saídas. Por isso é considerado um problema mal-posto, ou do inglês, *ill posed problem*. O problema da ambiguidade pode ser ilustrado pela Figura 3.13, em que a cadeira é percebida na primeira imagem como mais próxima do que realmente é, e quando posta em comparação com o tamanho médio de uma pessoa, sua distância real fica clara. A percepção humana de profundidade baseia-se principalmente no conhecimento prévio dos tamanhos dos objetos. Ao observar uma imagem, o sistema visual humano estima a cena 3D mais provável entre inúmeras possibilidades geométricas.

Figura 3.13: Um exemplo de ilusão de ótica



Fonte: Mertan, Duff e Unal (2022)

De acordo com Ke et al. (2024) um grau elevado de compreensão espacial e contextual é necessário para identificar apropriadamente a geometria de cenas, portanto, é evidente que os avanços na temática da inteligência artificial tenha acompanhado o salto de performance dessa tarefa. Atualmente, a tarefa de estimação monocular de profundidade é apresentada como um problema de tradução imagem para imagem baseado em redes neurais com aprendizado de maneira supervisionada por meio de coleções de pares de imagem RGB e mapas de profundidade devidamente alinhados.

Capítulo 4

Materiais e Métodos

4.1 Datasets

Bases de dados para treinamento ou teste de algoritmos de estimativa de profundidade consistem em imagens RGB de uma cena e sua anotação correspondente em profundidade. Ao longo do tempo, diversos *datasets* foram propostos para este fim com variações em formatos de anotações, tipos de cena (interior ou exterior), métodos de captura, qualidade, resolução e tamanho.

Geralmente são empregados sensores e outras tecnologias como *Stereo Matching* e *Structure from Motion* para criar os *datasets* de profundidade, porém, são abordagens muito complexas, custosas, ou inviáveis em algumas situações particulares, por exemplo, obter mapas de profundidades densos a partir de veículos em movimento (YANG et al., 2024a). Cada *dataset* possui suas próprias características, problemas e vieses. Dados com informação de profundidade e em alta qualidade são complexos de adquirir, sendo que os melhores conjuntos são utilizados no treinamento dos modelos presentes na literatura (RANFTL et al., 2020).

Dessa forma, para escolha das bases de dados a serem utilizadas para teste, temos os critérios: i) não ter composto o conjunto de treinamento dos modelos escolhidos para comparação, ii) conter dados válidos para avaliação considerando anotações precisas de profundidade, ou caso sejam esparsas, possuam máscara para indicar os pixels válidos, iii) ser uma base de dados conceituada na literatura. Os *datasets* escolhidos e suas características podem ser visualizados na Tabela 4.1.

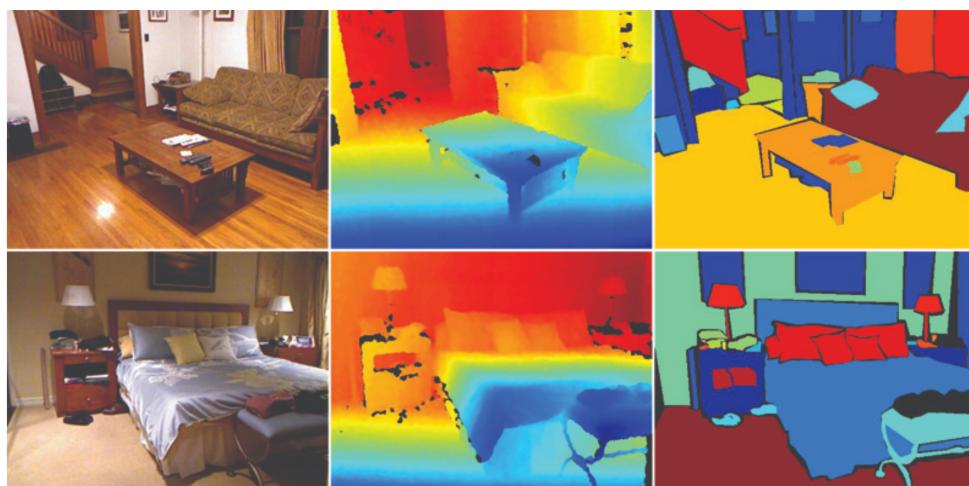
Tabela 4.1: Características dos datasets utilizados no trabalho

Dataset	Sensor	Anotação	Tipo	Cenário	Num. Imagens	Resolução
KITTI	LiDAR	Esparsa	Real	Outdoor	44 K	1024 × 320
Nyu-V2	Kinect V1	Densa	Real	Indoor	1449	640 × 480
DIODE	Laser Scanner	Densa	Real	Indoor/Outdoor	25,5 K	768 × 1024
SINTEL	-	Densa	Sintético	Indoor/Outdoor	1064	1024 × 436
ETH3D	Laser Scanner	Densa	Real	Indoor/Outdoor	454	6048 × 4032

4.1.1 NYUv2

O dataset NYUv2 é um dos mais utilizados em tarefas de visão computacional que envolvam estimativa de profundidade, segmentação de cenas e reconhecimento de objetos. Possui 1449 pares de imagens RGB e mapas de profundidade densos em diversas cenas *indoor* divididos em 795 para treinamento e 654 para teste (SILBERMAN et al., 2012). A resolução das imagens é de 640 × 480 pixels. O equipamento de aquisição foi o Microsoft Kinect que utiliza a técnica de ToF ponto a ponto, que produz resultados precisos de informação de profundidade. Além dos pares RGB-D, também são disponibilizados os dados brutos de leitura dos sensores em que é possível encontrar aproximadamente 70% de pixels com informação válida de profundidade, no entanto, as imagens finais foram processadas utilizando um método de correção, resultando em um mapa denso. Além das informações de profundidade, a base de dados provém rótulos de segmentação de objetos e relações de suporte (LAHIRI; REN; LIN, 2024), como observado na Figura 4.1 que ilustra um exemplo. Entre as cenas observadas, podemos citar cômodos comuns de casas como quartos, cozinhas, sala de aula, banheiro e etc (SILBERMAN et al., 2012).

Figura 4.1: Exemplo do dataset NYU Depth v2



Fonte: Dataset NYU Depth V2, do trabalho de Silberman et al. (2012)

4.1.2 KITTI

Com o objetivo de impulsionar o desenvolvimento da área de veículos autônomos, o trabalho de Geiger, Lenz e Urtasun (2012) apresenta o *benchmark* KITTI. Este *benchmark* possui bases de dados para diversos tipos de tarefas de visão computacional relacionadas à navegação de veículos, tais como, visão estéreo, fluxo ótico, correção de mapas de profundidade, estimativa de profundidade, odometria, detecção de objetos 2D e 3D, rastreamento de veículos e pessoas, detecção de pavimentação urbana e segmentação semântica.

Os dados foram capturados por meio de um veículo equipado com diversos tipos de sensores, por exemplo, câmeras RGB, scanner a laser, GPS (*Global Positioning System*) de alta precisão e IMU (*Inertia Measurement Unit*). Além disso, as imagens RGB utilizando um par de câmeras em estereoscopia e os dados de profundidade foram adquiridos com um sensor Velodyne HDL-643 que utiliza tecnologia LiDAR. Os rótulos providos por esse sensor são esparsos, apresentando aproximadamente 30% de pixels válidos em cada imagem. A resolução das imagens da base de dados de estimativa de profundidade possuem resolução de 1216×352 (LAHIRI; REN; LIN, 2024). Um dos exemplos do *dataset* pode ser visualizado na Figura 4.2, em que podemos observar um mapa de profundidade esparso e a imagem RGB correspondente.

Figura 4.2: Exemplo do *dataset* KITTI de estimativa de profundidade



Fonte: *Dataset KITTI*, do trabalho de Geiger, Lenz e Urtasun (2012)

4.1.3 SINTEL

O Sintel é uma base de dados sintéticos criada a partir de uma projeto aberto de animação em CGI (*Computer-Generated Imagery*) chamado *Durian Open Source Movie Project*. Foi elaborado no *software* Blender e contém os arquivos fontes utilizados na criação do filme animado. Em (BUTLER et al., 2012), os autores adaptaram os arquivos brutos em uma base de dados primariamente utilizada para o problema de visão computacional de estimativa de fluxo ótico. Apesar do seu uso primário ser em outra tarefa, ele também contém anotações de profundidade das imagens. A Figura 4.3 mostra um quadro renderizado e seu mapa de profundidade calculado.

Figura 4.3: Exemplo do *dataset* Sintel



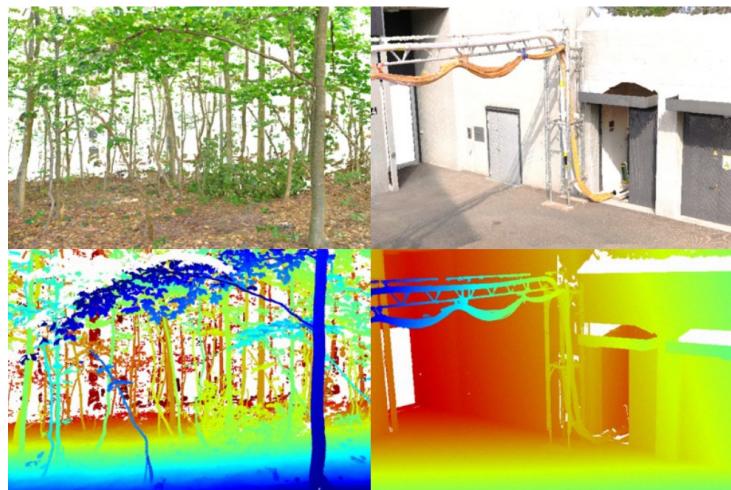
Fonte: *Dataset* Sintel, do trabalho de Butler et al. (2012)

A base de dados contém um total de 35 cenas, cada uma possuindo entre 20 a 50 quadros, totalizando 1628 imagens. A resolução de renderização é de 1024×436 em 24 FPS (*Frames Per Second*). As cenas possuem como características presentes efeitos atmosféricos, efeitos de movimento, desfoque em razão de velocidade e alta variabilidade de ambientes, personagens e ações. É dividido em 1064 quadros para o conjunto de treinamento e 564 para teste, sendo anotadas apenas as imagens de treinamento (WULFF et al., 2012).

4.1.4 ETH3D

O ETH3D é uma base de dados geralmente utilizada para reconstrução em vistas múltiplas e *stereo matching*. Contém dados de treinamento com imagens RGB *multiview*, capturadas com câmeras DSLR (*Digital Single Lens Reflex*) e anotações de profundidade capturadas utilizando um scanner a laser Faro Focus X 330. Oferece três versões de imagens de profundidade, uma correspondente à leitura bruta do sensor (*raw*), outra com valores extremos removidos por trabalho manual e uma ferramenta automática (*clean*) e uma com valores extremos e pontos observados por uma única imagem RGB removidos. A partição de teste não contém anotações. A base de dados é associada à um desafio aberto ao público. Inclui cenas tanto internas quanto externas, oferecendo um protocolo de avaliação bem variado (LAHIRI; REN; LIN, 2024) (SCHOPS; SATTLER; POLLEFEYS, 2019).

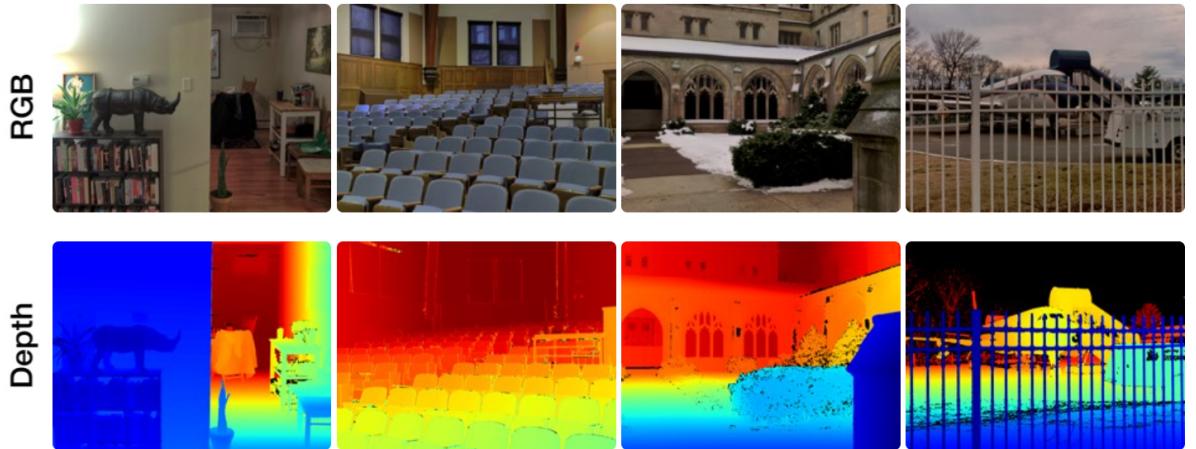
Figura 4.4: Exemplo do *dataset* ETH3D



Fonte: *Dataset* ETH3D, do trabalho de Schöps et al. (2017)

4.1.5 DIODE

O *dataset* DIODE (*Dense Indoor and Outdoor Depth Dataset*), é uma base de dados para estimativa monocular de profundidade e possui cerca de 8.000 imagens de ambientes internos e 16.000 de ambientes externos para treinamento e teste. Possui resolução de 768 × 1024 com faixa de distâncias entre 50m e 300m para os ambientes internos e externos respectivamente. O equipamento de aquisição é o scanner a laser Faro Focus S350. Alguns exemplos do *dataset* podem ser visualizados na Figura 4.5 (VASILJEVIC et al., 2019).

Figura 4.5: Exemplo do *dataset* DIODE

Fonte: *Dataset* DIODE, do trabalho de Vasiljevic et al. (2019)

4.2 Protocolo de Avaliação

Para avaliar os modelos de estimação de profundidade, será utilizado o protocolo de *zero-shot cross-dataset transfer*, i.e. realizar os testes e métricas em bases de dados que não compuseram os conjuntos de treinamentos dos modelos analisados. Por exemplo, se uma partição de treinamento de uma base de dados específica for empregada para o treinamento de um modelo, nem mesmo a partição destinada aos testes dessa base será utilizada para a sua avaliação. A performance em *cross-dataset* é considerada uma aproximação mais fiel da performance em mundo real em uma aplicação, pois os conjuntos de testes relativos aos conjuntos utilizados no treinamento podem refletir os mesmos vieses e situações (RANFTL et al., 2020).

Há dois tipos de métricas usadas para avaliação de mapas de profundidade, a acurácia de profundidade e o erro métrico de profundidade. Apesar do problema de MDE ser caracterizado como uma tarefa de regressão, é possível calcular a sua acurácia através da porcentagem de pixels cujo erro excede um determinado limite. Essa métrica, também chamada de acurácia relativa a um limiar (*Relative Accuracy Threshold* - δ) é definida pela equação 4.1.

$$\delta_t = \max \left(\frac{d_i}{\hat{d}_i}, \frac{\hat{d}_i}{d_i} \right) < 1.25^t \quad \text{com } t \in (1, 2, 3) \quad (4.1)$$

As métricas relacionadas a erro métrico de profundidade buscam calcular o desvio numérico da profundidade estimada em relação ao esperado, sendo padrão em proble-

mas de regressão. Nesse tipo, as métricas mais presentes na literatura de estimadores monoculares de profundidade são o Erro Absoluto Relativo (*Absolute Relative Error - AbsRel*) e a raiz do erro médio quadrático (*Root Mean Squared Error - RMSE*), definidos respectivamente pelas equações 4.2 e 4.3.

$$\text{AbsRel} = \frac{1}{N} \sum_{i \in N} \left(\frac{|d_i - \hat{d}_i|}{\hat{d}_i} \right) \quad (4.2)$$

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i \in N} (|d_i - \hat{d}_i|)^2} \quad (4.3)$$

Outro fator importante no protocolo de avaliação de mapas de profundidade é que os modelos de MDE geralmente inferem um mapa relativo, ou seja, os números presentes em cada pixel são correlacionados com a distância real por um fator desconhecido. Dessa forma, seguindo o protocolo com invariância afim presente em (KE et al., 2024), o mapa de profundidade predito \hat{d} que será comparado com o mapa verdadeiro d precisa ser alinhado por um fator de escala s e deslocamento t , seguindo a equação 4.4.

$$\mathbf{a} = \mathbf{m} \times s + t \quad (4.4)$$

O alinhamento do mapa de profundidade será feito utilizando o método dos mínimos quadrados.

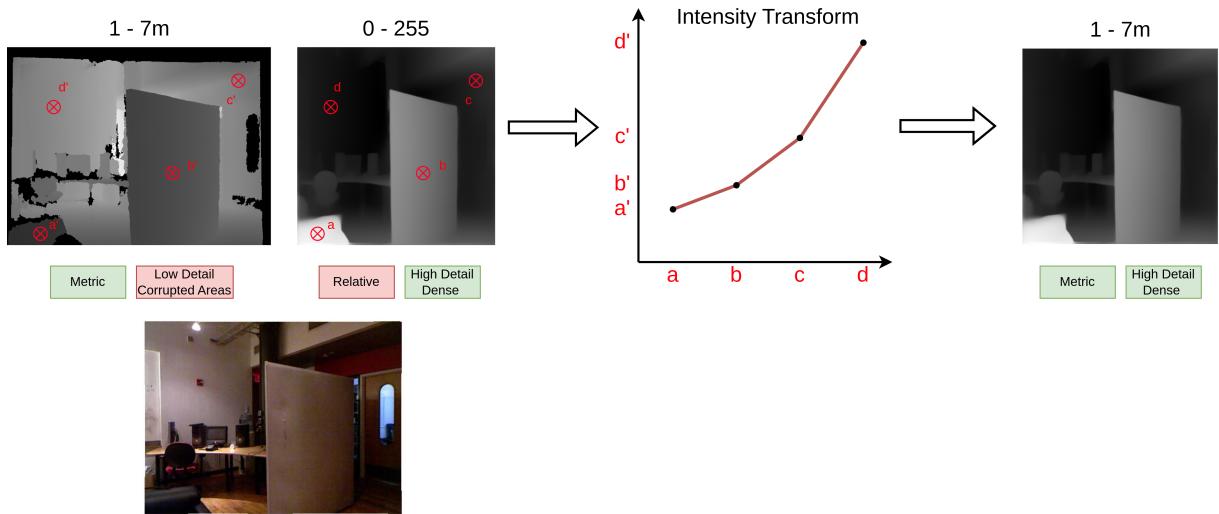
4.3 Método de Transformação de Intensidades (pós-processamento)

Um mapa de profundidade inferido por um método de estimativa de profundidade possui a característica de ser denso, pois todos os pixels possuem um valor predito associado, preciso, bem detalhado, de acordo com os últimos trabalhos do estado da arte porém é relativo, i.e. o valor de cada pixel é apenas correlacionado com a medição de distância real por um fator desconhecido. Já um mapa de profundidade adquirido com um sensor físico consegue representar as grandezas de forma métrica (em metros, centímetros ou até milímetros), mas pode ter características negativas associadas a depender do dispositivo de aquisição, podendo conter áreas falhas que não possuem medição associada, ou um elevado grau de esparsidate. O método de transformação de intensidades para transfe-

rência de domínio almeja como resultado uma imagem de profundidade que possuam as características positivas dos dois casos anteriormente citados.

O método proposto por este trabalho consiste em uma transformação de intensidades que é projetada para cada imagem de um conjunto de dados utilizando pontos correspondentes em ambas e associando uma transformação linear para cada ponto, como visualizado na Figura 4.6.

Figura 4.6: Diagrama do método de transferência de domínio



Elaborado pelo autor.

O método proposto diferencia-se do tradicional baseado em fator de escala e deslocamento por mínimos quadrados pois é associada uma função linear para cada região na quantização da imagem, o que propicia uma correção adaptada para cada proporção de distância. Ressalta-se que o método não será utilizado no protocolo de avaliação dos modelos de estimativa de profundidade.

Será realizada a comparação do resultado da técnica de pós-processamento e o resultado de estimadores métricos de profundidade com os conjuntos de dados que possuem leituras métricas de sensores.

4.4 Análise com Aplicação

Para avaliar os mapas de profundidade gerado pelos modelos do estado da arte, será utilizada também uma ou mais aplicações práticas. O processo de escolha da aplicação a ser implementada e avaliada passa por três principais critérios, sendo eles:

- Empregar diretamente como entrada do sistema um mapa de profundidade denso.
- Possuir código de testes público.
- Ter o seu desempenho mensurável através de métricas.

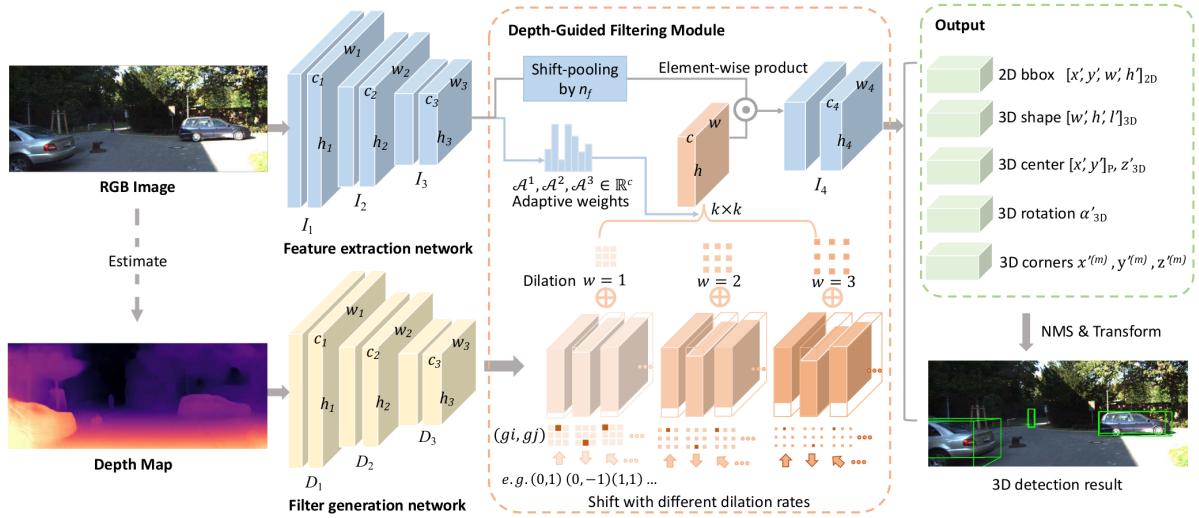
4.4.1 Aplicação: Detecção de objetos 3D

Uma das aplicações escolhidas foi a detecção de objetos 3D, que é uma tarefa fundamental da área de veículos autônomos. A percepção automática do ambiente ao redor do veículo ajuda-o a entender a situação para tomada de decisões e navegação. De acordo com Ding et al. (2020), um dos fatores limitantes da performance de métodos de detecção 3D é a acurácia dos mapas de profundidade. Dessa forma, o método apresentado pelo autor será utilizado como *benchmark* para avaliação dos mapas gerados pelos estimadores monoculares do estado da arte.

Em Ding et al. (2020) é apresentado uma rede neural *Depth-guided Dynamic Depthwise-Dilated Local Convolutional Network* (D4LCN). A proposta do trabalho é gerar filtros convolucionais (kernel) dinâmicos e dilatados, com o processo sendo guiado a partir de mapas de profundidade. Esses filtros são aplicados nas imagens RGB em escala de pixel e de canal, fazendo assim a conexão entre as características bidimensionais e tridimensionais.

Para performar a detecção 3D guiada profundidade, o autor apresentou a arquitetura da Figura 4.7, que consiste em uma rede de dois ramos. O primeiro ramo é responsável pela extração de características da imagem RGB e possui como *backbone* a rede ResNet-50 sem sua camada final completamente convolucional e sem camadas de *pooling* pré-treinada na base de dados ImageNet. O segundo ramo realiza a geração dos kernels convolucionais com dilatação dinâmica a partir do mapa de profundidade para aprender as especificidades da geometria da imagem. As convoluções são aplicadas de forma a distinguir o que é objeto e o que é região de fundo para cada pixel, enquanto as dilatações dinâmicas fazem com que filtros diferentes possuam campos receptivos diferentes para objetos em diferentes escalas. As saídas de cada uma dos ramos são fusionadas por meio de um módulo de filtragem guiada por profundidade, que executa convoluções locais com diferentes kernels para cada pixel e diferentes dilatações para cada canal.

Figura 4.7: Arquitetura da rede D4LCN.



Fonte: Ding et al. (2020)

Como métrica de avaliação, seguindo o trabalho de Ding et al. (2020), serão utilizadas as curvas de precisão-recall com limiar de IoU de 0,7. A base de dados KITTI dispõe as métricas de precisão média interpolada de 11 pontos $AP|R_{11}$.

4.5 Considerações Metodológicas

A metodologia do presente trabalho consiste em três pilares. O primeiro é a comparação entre estimadores de profundidade através de métricas presentes na literatura com o fim de compreender e explorar as capacidades do estado da arte da área e evidenciar suas vantagens e desvantagens em relação a métodos tradicionais. Apesar dos modelos a serem utilizados ainda não terem sido escolhidos, o Apêndice A contém um breve estudo de um modelo estimador baseado em difusão latente, proposto por Ke et al. (2024).

O segundo pilar consiste na aplicação de uma metodologia baseada em processamento digital de imagens para transferir o resultado de estimadores relativos para o espaço métrico, ou seja, com os valores numéricos em unidade de metros. O objetivo é mostrar que com dados auxiliares provenientes de sensores é possível obter distâncias em unidades métricas sem a necessidade de ajuste fino dos modelos, pois esse processo geralmente causa uma adaptação à domínios específicos.

Por fim, almeja-se comparar os mapas gerados por métodos do estado da arte por meio do desempenho de aplicações práticas, desse modo, mensurando seu comportamento em

situações de mundo real. Até o presente momento, a seleção de aplicações segundo os critérios estabelecidos não foi concluída, havendo somente o estudo de métodos de detecção 3D aplicados a veículos autônomos.

Capítulo 5

Resultados Preliminares e Discussões

Como resultados preliminares deste trabalho até o presente momento há o uso do modelo *Depth Anything v1* (YANG et al., 2024a) para estimar a profundidade em alguns exemplos do *dataset* Nyu Depth V2, exibidos nas Figuras 5.1, 5.2, 5.3 e 5.4. Como se trata de um experimento preliminar, foi utilizada resolução de 256×256 .

Podemos observar que o mapa inferido pelo modelo de estimação de profundidade não possui as tonalidades da escala de cinza de acordo com o verdadeiro proveniente da base de dados. No entanto, após a aplicação da transformação de intensidade, é obtido um mapa visualmente próximo do esperado. Um fato a ser observado é também o fato das transformações de intensidades não serem aproximadamente lineares, havendo a necessidade de correção de proporção em determinadas regiões na quantização dos mapas. Um caso particular ocorre nas Figuras 5.4(c) e 5.5(c), em que a função obtida não possui crescimento monotônico, portanto, são necessário ajustes no algoritmo desenvolvido. Por fim, também é visualizado que os mapas de profundidade corrigidos com o método proposto possuem característica de acurácia métrica e definição dos objetos em termos visuais.

Figura 5.1: **Exemplo 1.** (a) Imagem RGB, (b) Resultado relativo do modelo, (c) Transformação de intensidades de transferência de domínio relativo para métrico, (d) Em ordem: Mapa de profundidade do *dataset*, mapa de profundidade do *dataset* corrigido e mapa de profundidade estimado e corrigido com transferência de domínio.

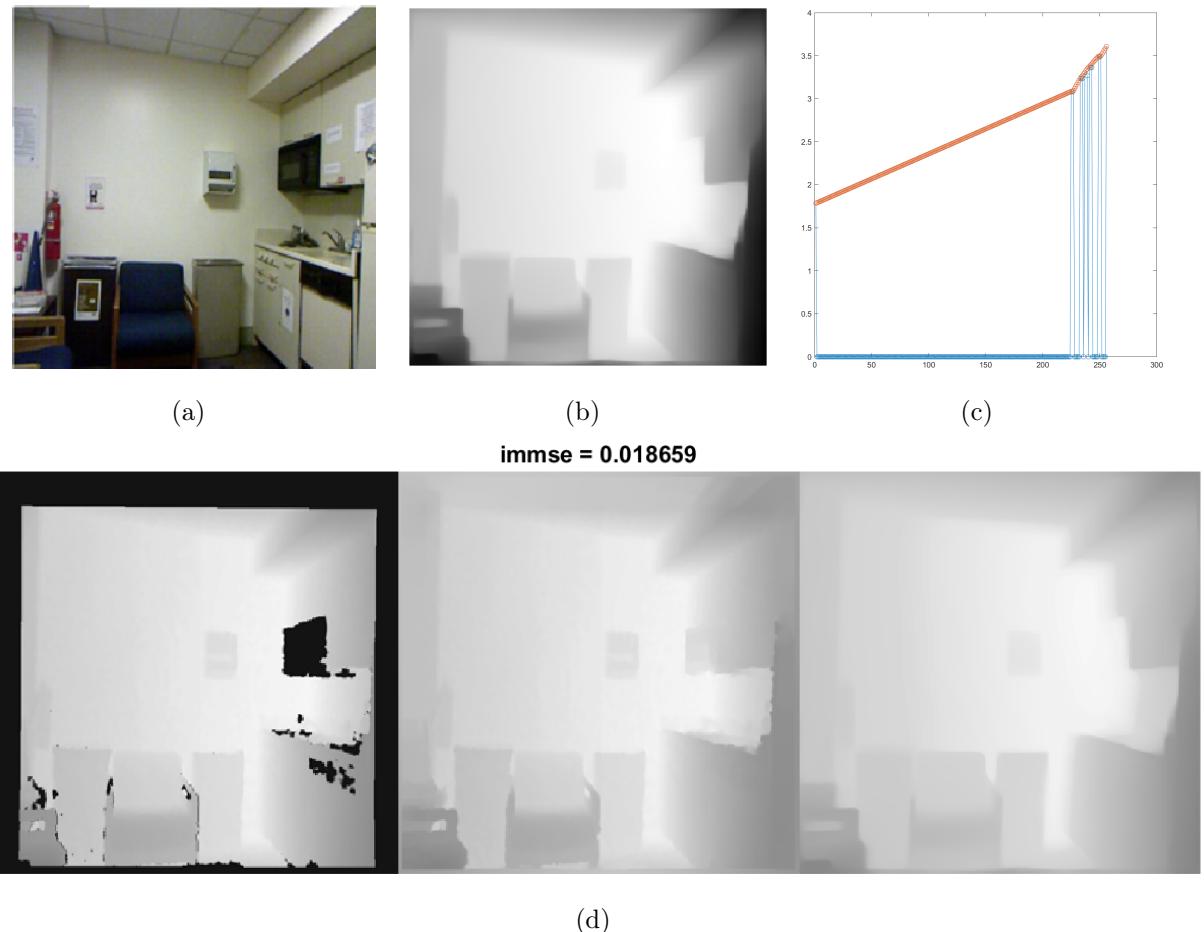


Figura 5.2: **Exemplo 2.** (a) Imagem RGB, (b) Resultado relativo do modelo, (c) Transformação de intensidades de transferência de domínio relativo para métrico, (d) Em ordem: Mapa de profundidade do *dataset*, mapa de profundidade do *dataset* corrigido e mapa de profundidade estimado e corrigido com transferência de domínio.

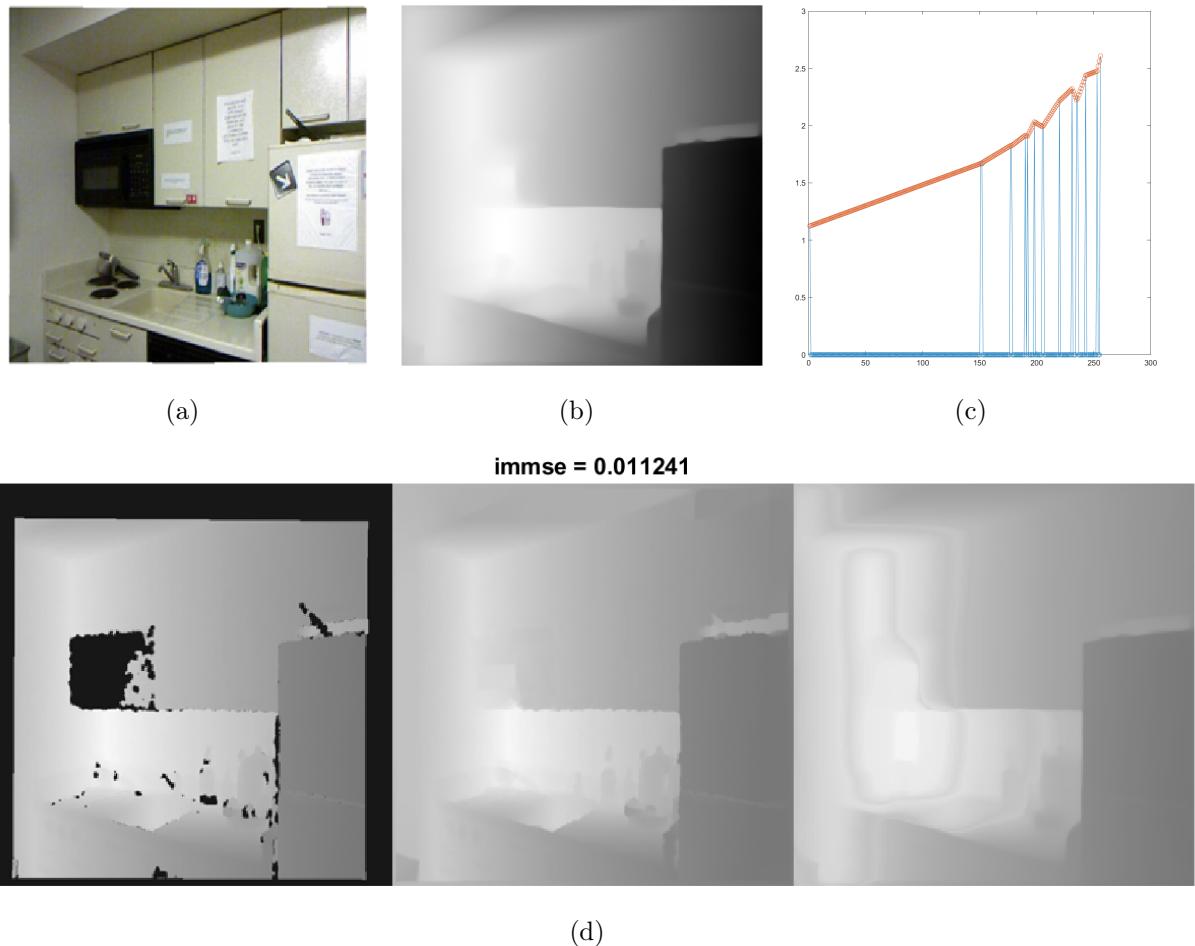


Figura 5.3: **Exemplo 3.** (a) Imagem RGB, (b) Resultado relativo do modelo, (c) Transformação de intensidades de transferência de domínio relativo para métrico, (d) Em ordem: Mapa de profundidade do *dataset*, mapa de profundidade do *dataset* corrigido e mapa de profundidade estimado e corrigido com transferência de domínio.

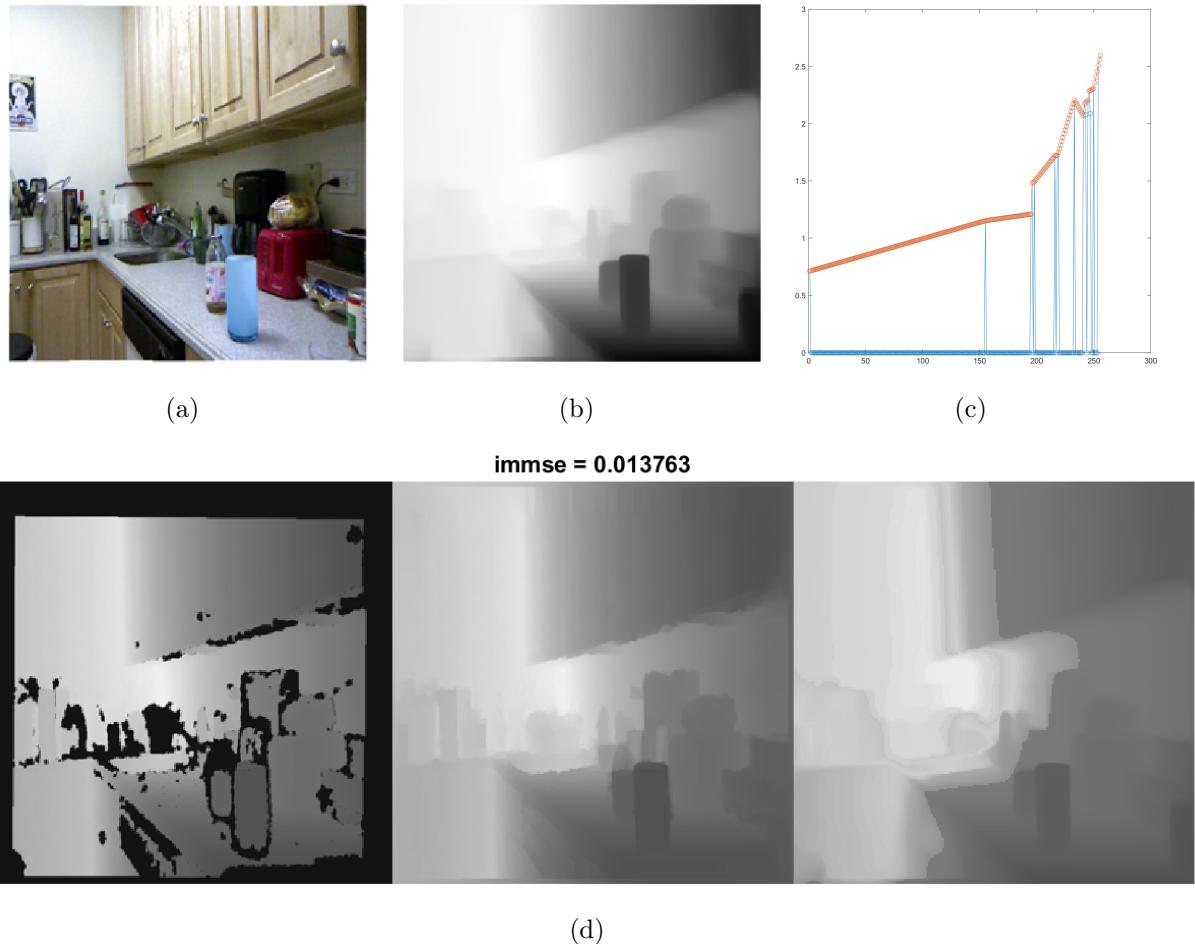
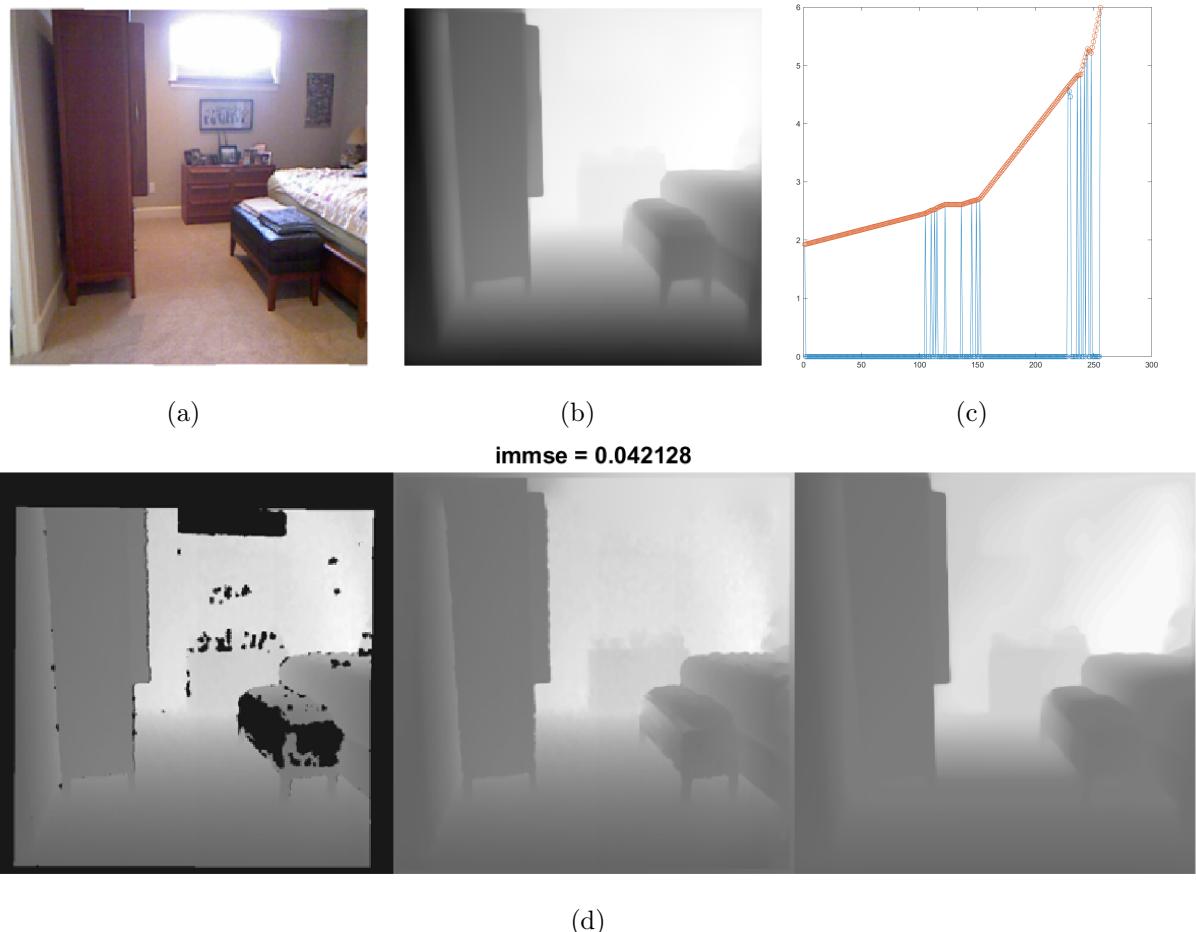


Figura 5.4: **Exemplo 4.** (a) Imagem RGB, (b) Resultado relativo do modelo, (c) Transformação de intensidades de transferência de domínio relativo para métrico, (d) Em ordem: Mapa de profundidade do *dataset*, mapa de profundidade do *dataset* corrigido e mapa de profundidade estimado e corrigido com transferência de domínio.



Capítulo 6

Cronograma

Este capítulo visa expor as atividades realizadas e futuras considerando os prazos estipulados para finalização da pesquisa científica e defesa de dissertação. A tabela 6.1 mostra as atividades realizadas no ano de 2023 e a Tabela 6.2 para o ano 2024 e Janeiro de 2025.

Tabela 6.1: Cronograma com as atividades realizadas para o desenvolvimento da pesquisa do ano de 2023

Tabela 6.2: Cronograma com as atividades realizadas e pretendidas para o desenvolvimento da pesquisa do ano de 2024 e Janeiro de 2025.

Referências

- AGGARWAL, C. C. et al. *Neural networks and deep learning*. [S.l.]: Springer, 2018. v. 10.
- AMOS, E. *Kinect*. 2014. Acessado em 13 de agosto, 2024. Disponível em: <<https://commons.wikimedia.org/w/index.php?curid=33217678>>.
- BIRKL, R.; WOKF, D.; MÜLLER, M. Midas v3. 1—a model zoo for robust monocular relative depth estimation. *arXiv preprint arXiv:2307.14460*, 2023.
- BISHOP, C. M.; NASRABADI, N. M. *Pattern recognition and machine learning*. [S.l.]: Springer, 2006. v. 4.
- BRANSCOMBE, M. *How Microsoft is making its most sensitive HoloLens depth sensor yet*. 2018. <<https://www.zdnet.com/article/how-microsoft-is-making-its-most-sensitive-hololens-depth-sensor-yet/>>.
- BUTLER, D. J. et al. A naturalistic open source movie for optical flow evaluation. In: A. Fitzgibbon et al. (Eds.) (Ed.). *European Conf. on Computer Vision (ECCV)*. [S.l.]: Springer-Verlag, 2012. (Part IV, LNCS 7577), p. 611–625.
- CASTELLANO, R.; TERRERAN, M.; GHIDONI, S. Performance evaluation of depth completion neural networks for various rgb-d camera technologies in indoor scenarios. In: SPRINGER. *International Conference of the Italian Association for Artificial Intelligence*. [S.l.], 2023. p. 351–364.
- DING, M. et al. Learning depth-guided convolutions for monocular 3d object detection. In: *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition workshops*. [S.l.: s.n.], 2020. p. 1000–1001.
- DING, X. et al. Stereo depth estimation under different camera calibration and alignment errors. *Applied Optics*, Optica Publishing Group, v. 50, n. 10, p. 1289–1301, 2011.
- DONG, X. et al. Towards real-time monocular depth estimation for robotics: A survey. *IEEE Transactions on Intelligent Transportation Systems*, IEEE, v. 23, n. 10, p. 16940–16961, 2022.
- DOURADO, A. M. B.; PEDRINO, E. C. Multi-objective cartesian genetic programming optimization of morphological filters in navigation systems for visually impaired people. *Applied Soft Computing*, Elsevier, v. 89, p. 106130, 2020.
- DU, R. et al. Depthlab: Real-time 3d interaction with depth maps for mobile augmented reality. In: *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. [S.l.: s.n.], 2020. p. 829–843.

- EIGEN, D.; PUHRSCH, C.; FERGUS, R. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, v. 27, 2014.
- FARKHANI, S. et al. Sparse-to-dense depth completion in precision farming. In: *Proceedings of the 3rd International Conference on Vision, Image and Signal Processing*. [S.l.: s.n.], 2019. p. 1–5.
- FOLEY, H. J. *Sensation and perception*. [S.l.]: Routledge, 2019.
- GEIGER, A.; LENZ, P.; URTASUN, R. Are we ready for autonomous driving? the kitti vision benchmark suite. In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2012.
- GODARD, C. et al. Digging into self-supervised monocular depth estimation. In: *Proceedings of the IEEE/CVF international conference on computer vision*. [S.l.: s.n.], 2019. p. 3828–3838.
- GONZALEZ, R.; WOODS, R. *Digital Image Processing, Global Edition*. Pearson Education, 2018. ISBN 9781292223070. Disponível em: <<https://books.google.com.br/books?id=P8AoEAAAQBAJ>>.
- GONZALEZ, R. C.; WOODS, R. E. Processamento digital de imagem. *Pearson, ISBN-10: 8576054019*, v. 10, p. 11–27, 2010.
- GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep learning*. [S.l.]: MIT press, 2016.
- HANSARD, M. et al. *Time-of-flight cameras: principles, methods and applications*. [S.l.]: Springer Science & Business Media, 2012.
- HAYKIN, S. *Redes neurais: princípios e prática*. [S.l.]: Bookman Editora, 2001.
- HAYKIN, S. *Neural networks and learning machines, 3/E*. [S.l.]: Pearson Education India, 2009.
- HERTZ, J. A. *Introduction to the theory of neural computation*. [S.l.]: Crc Press, 2018.
- HOIEM, D.; EFROS, A. A.; HEBERT, M. Automatic photo pop-up. In: *ACM SIGGRAPH 2005 Papers*. [S.l.: s.n.], 2005. p. 577–584.
- HU, G. et al. A robust rgb-d slam algorithm. In: *IEEE. 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. [S.l.], 2012. p. 1714–1719.
- HU, H. et al. Deep learning-based monocular 3d object detection with refinement of depth information. *Sensors*, MDPI, v. 22, n. 7, p. 2576, 2022.
- HU, J. et al. Deep depth completion from extremely sparse data: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, v. 45, n. 7, p. 8244–8264, 2022.
- JAIN, A. K. *Fundamentals of digital image processing*. [S.l.]: Prentice-Hall, Inc., 1989.

- JARITZ, M. et al. Sparse and dense data with cnns: Depth completion and semantic segmentation. In: IEEE. *2018 International Conference on 3D Vision (3DV)*. [S.l.], 2018. p. 52–60.
- KE, B. et al. Repurposing diffusion-based image generators for monocular depth estimation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2024. p. 9492–9502.
- KHAN, F.; SALAHUDDIN, S.; JAVIDNIA, H. Deep learning-based monocular depth estimation methods—a state-of-the-art review. *Sensors*, MDPI, v. 20, n. 8, p. 2272, 2020.
- KOPF, J.; RONG, X.; HUANG, J.-B. Robust consistent video depth estimation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2021. p. 1611–1621.
- LAHIRI, S.; REN, J.; LIN, X. Deep learning-based stereopsis and monocular depth estimation techniques: a review. *Vehicles*, MDPI, v. 6, n. 1, p. 305–351, 2024.
- LASINGER, K. et al. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *arXiv preprint arXiv:1907.01341*, 2019.
- LECUN, Y. et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, Ieee, v. 86, n. 11, p. 2278–2324, 1998.
- MA, F. et al. Sparse depth sensing for resource-constrained robots. *The International Journal of Robotics Research*, SAGE Publications Sage UK: London, England, v. 38, n. 8, p. 935–980, 2019.
- MERTAN, A.; DUFF, D. J.; UNAL, G. Single image depth estimation: An overview. *Digital Signal Processing*, Elsevier, v. 123, p. 103441, 2022.
- NIELSEN, M. A. *Neural networks and deep learning*. [S.l.]: Determination press San Francisco, CA, USA, 2015. v. 25.
- O'SHEA, K.; NASH, R. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015.
- PADHY, R. P. et al. Monocular vision-aided depth measurement from rgb images for autonomous uav navigation. *ACM Transactions on Multimedia Computing, Communications and Applications*, ACM New York, NY, v. 20, n. 2, p. 1–22, 2023.
- PARK, H.; LEE, Y.; KO, J. Enabling real-time sign language translation on mobile platforms with on-board depth cameras. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, ACM New York, NY, USA, v. 5, n. 2, p. 1–30, 2021.
- PLACED, J. A. et al. A survey on active simultaneous localization and mapping: State of the art and new frontiers. *IEEE Transactions on Robotics*, IEEE, v. 39, n. 3, p. 1686–1705, 2023.
- RAJAPAKSHA, U. et al. Deep learning-based depth estimation methods from monocular image and videos: A comprehensive survey. *ACM Computing Surveys*, ACM New York, NY, 2024.

- RANFTL, R. et al. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 44, n. 3, p. 1623–1637, 2020.
- ROSENBLATT, F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, American Psychological Association, v. 65, n. 6, p. 386, 1958.
- RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. (1986) de rumelhart, ge hinton, and rj williams, learning internal representations by error propagation, parallel distributed processing: Explorations in the microstructures of cognition, vol. i, de rumelhart and jl mcclelland (eds.) cambridge, ma: Mit press, pp. 318-362. 1988.
- RUSS, J. C. *The image processing handbook*. [S.l.]: CRC press, 2006.
- SAXENA, A.; CHUNG, S.; NG, A. Learning depth from single monocular images. *Advances in neural information processing systems*, v. 18, 2005.
- SAXENA, A.; SUN, M.; NG, A. Y. Make3d: Learning 3d scene structure from a single still image. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 31, n. 5, p. 824–840, 2008.
- SCHOPS, T.; SATTLER, T.; POLLEFEYS, M. Bad slam: Bundle adjusted direct rgbd slam. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2019. p. 134–144.
- SCHÖPS, T. et al. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2017.
- SEE, A. R.; SASING, B. G.; ADVINCULA, W. D. A smartphone-based mobility assistant using depth imaging for visually impaired and blind. *Applied Sciences*, MDPI, v. 12, n. 6, p. 2802, 2022.
- SILBERMAN, N. et al. Indoor segmentation and support inference from rgbd images. In: SPRINGER. *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7–13, 2012, Proceedings, Part V 12*. [S.l.], 2012. p. 746–760.
- SONG, Z. et al. Self-supervised depth completion from direct visual-lidar odometry in autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, IEEE, v. 23, n. 8, p. 11654–11665, 2021.
- SPENCER, J. et al. The third monocular depth estimation challenge. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2024. p. 1–14.
- STACHNISS, C.; LEONARD, J. J.; THRUN, S. Simultaneous localization and mapping. *Springer Handbook of Robotics*, Springer, p. 1153–1176, 2016.
- SZELISKI, R. *Computer vision: algorithms and applications*. [S.l.]: Springer Nature, 2022.

- TATENO, K. et al. Cnn-slam: Real-time dense monocular slam with learned depth prediction. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2017. p. 6243–6252.
- VASILJEVIC, I. et al. DIODE: A Dense Indoor and Outdoor DEpth Dataset. *CoRR*, abs/1908.00463, 2019. Disponível em: <<http://arxiv.org/abs/1908.00463>>.
- WU, X. et al. Sparse fuse dense: Towards high quality 3d detection with depth completion. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. [S.l.: s.n.], 2022. p. 5418–5427.
- WULFF, J. et al. Lessons and insights from creating a synthetic optical flow benchmark. In: A. Fusiello et al. (Eds.) (Ed.). *ECCV Workshop on Unsolved Problems in Optical Flow and Stereo Estimation*. [S.l.]: Springer-Verlag, 2012. (Part II, LNCS 7584), p. 168–177.
- XIE, Z. et al. Ultradepth: Exposing high-resolution texture from depth cameras. In: *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*. [S.l.: s.n.], 2021. p. 302–315.
- YANG, L. et al. Depth anything: Unleashing the power of large-scale unlabeled data. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2024. p. 10371–10381.
- YANG, L. et al. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024.
- ZHANG, Y. et al. Indepth: Real-time depth inpainting for mobile augmented reality. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, ACM New York, NY, USA, v. 6, n. 1, p. 1–25, 2022.
- ZHAO, C. et al. Monocular depth estimation based on deep learning: An overview. *Science China Technological Sciences*, Springer, v. 63, n. 9, p. 1612–1627, 2020.
- ZHOU, B.; KRÄHENBÜHL, P.; KOLTUN, V. Does computer vision matter for action? *Science Robotics*, American Association for the Advancement of Science, v. 4, n. 30, p. eaaw6661, 2019.
- ZOLLHÖFER, M. Commodity rgb-d sensors: Data acquisition. *RGB-D image analysis and processing*, Springer, p. 3–13, 2019.

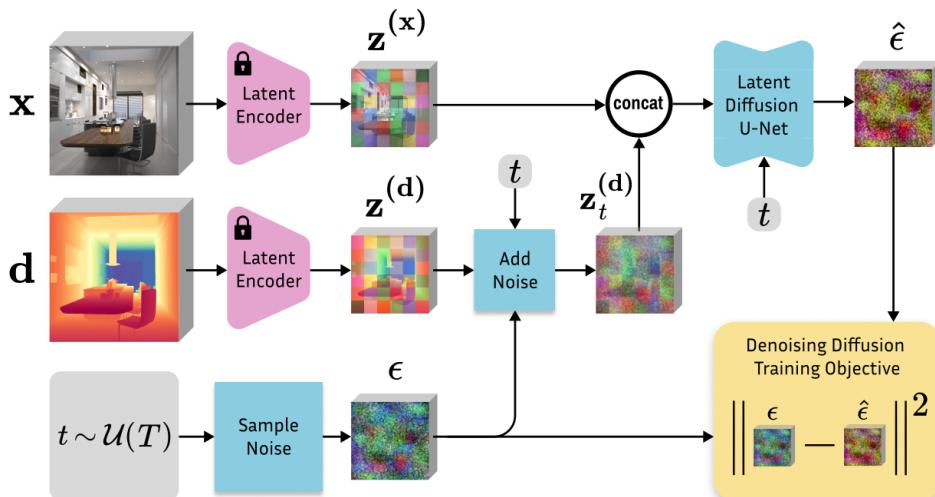
APÊNDICES A – Apêndice A

A.1 Modelos Estudados

A.1.1 Marigold

Os autores (KE et al., 2024) apresentaram o Marigold, um modelo de difusão latente baseado em *Stable Diffusion* (SD). Modelos fundacionais de visão computacional, como o SD, são treinados com dados em larga escala e em uma vasta gama de domínios. O seu funcionamento é baseado na premissa de que a tarefa de MDE possui como pilar a plena compreensão das representações visuais do mundo, portanto, é possível aproveitar o conhecimento prévio de um modelo fundacional de difusão para transformá-lo em um estimador de profundidade. Desse modo é proposto um protocolo de ajuste fino (ou do inglês, *fine-tuning*) que objetiva a adaptação de um modelo de difusão latente para um modelo de MDE, mostrado na Figura A.1.

Figura A.1: Protocolo de ajuste fino do Marigold.



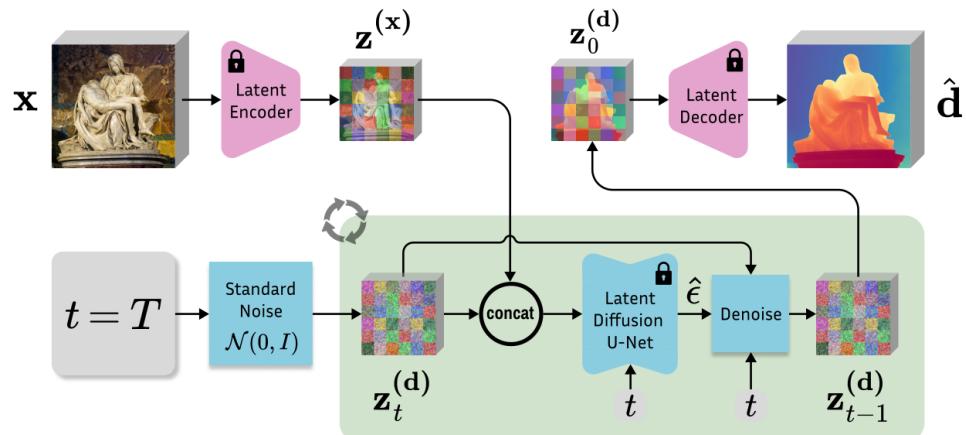
Fonte: Ke et al. (2024)

Um VAE (*Variational Auto Encoder*) pré-treinado, proveniente do SD, é empregado

para projetar a imagem RGB e o mapa de profundidade em um espaço latente de menor dimensionalidade. O processo de ajuste fino se concentra na componente U-Net, otimizando uma função objetivo. Essa função compara o código latente obtido a partir do ruído de difusão com a representação latente gerada pela U-Net após o processo de remoção de ruído (*denoising*). É importante destacar que apenas os parâmetros da U-Net são ajustados durante o treinamento.

O processo de inferência funciona de acordo com a Figura A.2. Dado uma imagem RGB de entrada, ela é processada com por meio do mesmo VAE utilizado no *fine-tuning*) e concatenada com o ruído já no espaço latente. Em seguida, passa pela U-Net para remover o ruído a cada iteração. Depois da execução de T passos de difusão, a imagem é descodificada pelo VAE em uma imagem de 3 canais, que representa o mapa de profundidade final.

Figura A.2: Processo de inferência do Marigold.



Fonte: Ke et al. (2024)