

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

...

Palavras-chave: ..., ...;

Abstract

....

Keywords: Regression; GAMLSS; OLLST; Repeated measure in time

Listas de Figuras

3.1	Etapas do processamento digital de imagens que vai desde a aquisição de imagens até a identificação e descrição de objetos presente nelas.	8
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.	10
3.3	Disograma de um perceptron, ilustrando uma abordagem simplificada baseada na estrutura e função de um neurônio biológico.	12
4.1	Exemplo do dataset NYU Depth v2	16
4.2	Exemplo do dataset DIODE	17
4.3	Diagrama do método de transferência de domínio	18
4.4	Esquema de correção não-guiada.	19
4.5	Esquema de correção guiada com <i>Early Fusion</i>	19
4.6	Esquema de correção guiada com <i>Late Fusion</i>	20
4.7	Esquema do método LaMa (SUVOROV et al., 2022).	20

Lista de Tabelas

4.1	Características dos datasets utilizados no trabalho	15
6.1	Cronograma com as atividades realizadas para o desenvolvimento da pesquisa do ano de 2023	23
6.2	Cronograma com as atividades realizadas e pretendidas para o desenvolvimento da pesquisa do ano de 2024 e Janeiro de 2025.	24

Sumário

1	Introdução	1
1.1	Contextualização da pesquisa	1
1.2	Motivação e Justificativa	3
1.3	Objetivos	3
1.3.1	Objetivo Geral	3
1.3.2	Objetivos Específicos	3
2	Trabalhos Relacionados	4
3	Fundamentação Teórica	6
3.1	Processamento Digital de Imagens	6
3.1.1	Transformação de Intensidade	9
3.2	Deep Learning	10
3.3	Informação de profundidade	12
3.4	Modelos de estimativa de profundidade	13
4	Materiais e Métodos	14
4.1	Datasets	14
4.1.1	NYUv2	15
4.1.2	KITTI	15
4.1.3	SINTEL	15
4.1.4	ETH3D	15
4.1.5	DIODE	16

Sumário

4.2 Modelos Escolhidos	17
4.3 Protocolo de Avaliação	17
4.4 Método de Transformação de Intensidades (pós-processamento)	17
4.5 Correção de mapas de profundidade	18
4.5.1 Large Mask Inpainting	20
4.6 Análise com Aplicação	21
4.7 Considerações Metodológicas	21
5 Resultados e Discussões	22
5.1 Resultados Preliminares	22
5.2 Resultados Esperados	22
6 Cronograma	23
Referências	25
Apêndices A – Apêndice A	29

Capítulo 1

Introdução

1.1 Contextualização da pesquisa

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).

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

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. 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 (approx. 95% de esparsidade) e no caso de câmeras RGB-D ou câmeras 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 *stereo* requerem processos complexos de calibração e alinhamento (DONG et al., 2022).

O problema da estimativa monocular de profundidade (*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 estimativa monocular de profundidade utilizando toda a miríade de técnicas e metodologias dentro do universo do DL, 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 estimativa monocular de profundidade relativa baseados em DL através da abordagem quantitativa, utilizando métricas e *benchmarks* presentes na literatura, abordagem quali-

tativa e através de uma aplicação.

1.2 Motivação e Justificativa

1.3 Objetivos

1.3.1 Objetivo Geral

Este trabalho possui como objetivo geral a análise comparativa de estimadores monoculares de profundidade robustos capazes de produzir informação de profundidade de alta qualidade para imagens sob quaisquer circunstâncias.

1.3.2 Objetivos Específicos

- Estudo e escolha dos datasets que tenha as imagens apropriadas para teste.
- Estudo de modelos de estimação monocular de profundidade relativa do estado da arte.
- Análise e escolha entre os modelos estudados para implementação e testes.
- Implementação de método de pós-processamento para transferência do domínio relativo para métrico baseado em transformação de intensidade.
- Avaliação de desempenho perante métricas utilizadas na literatura para comparação entre os modelos no espaço relativo e métrico.
- Avaliação qualitativa dos resultados.
- Implementação de aplicação com os mapas de profundidade gerados.

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 consiste 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, Puhrsch 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.

A pesquisa realizada por Ranftl et al. (2020) 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 obtenção do *groundtruth*.

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 rectificado em multi resolução no processo de difusão.

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 Multiresolução:** 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.

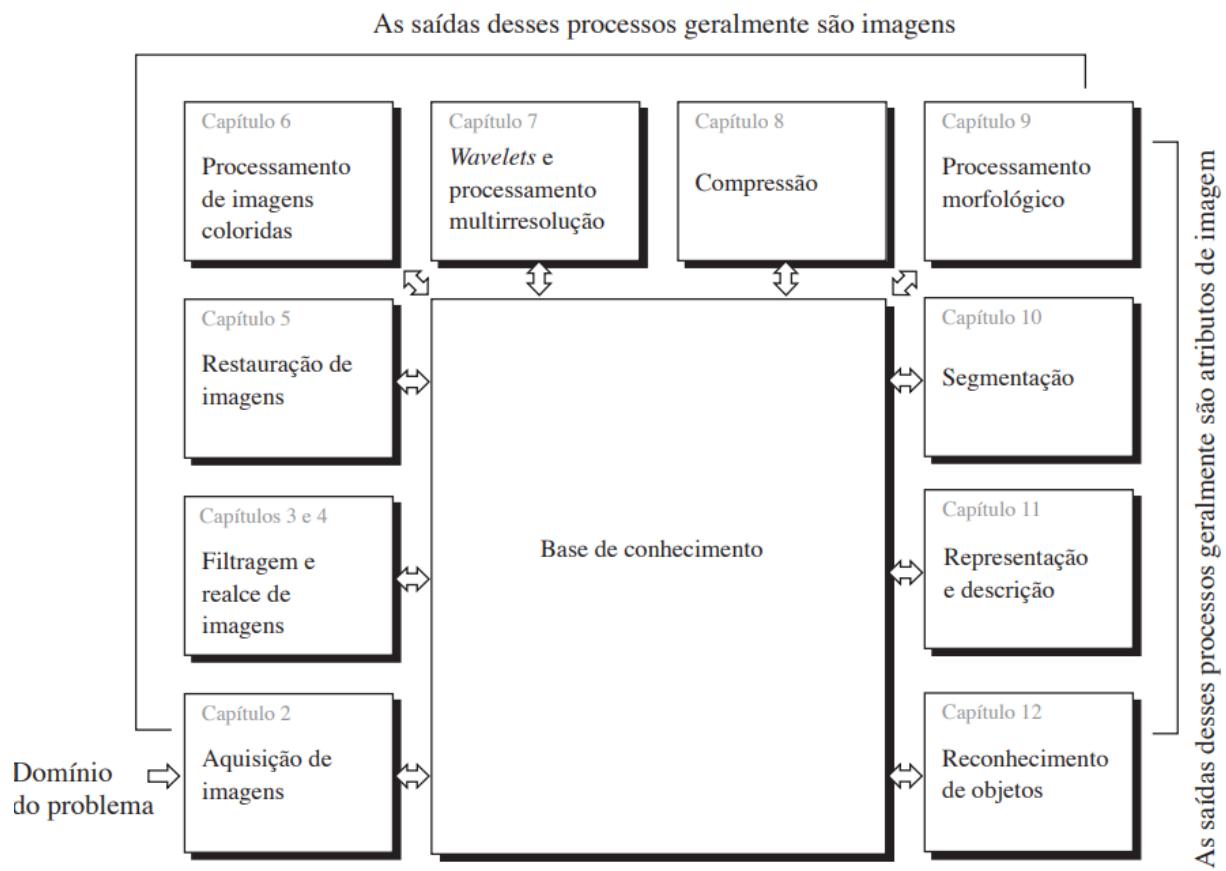


Figura 3.1: Etapas do processamento digital de imagens que vai desde a aquisição de imagens até a identificação e descrição de objetos presentes nelas.

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. Considere a transformação de intensidade ilustrada na Figura 3.2.a. Como mostra Gonzalez e Woods (2010), quando aplicamos essa transformação a cada pixel da imagem original f , geramos 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 3.2.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.

Essas transformações são exemplos de como ajustes nos valores de intensidade podem alterar significativamente a aparência e a utilidade da imagem, dependendo do objetivo

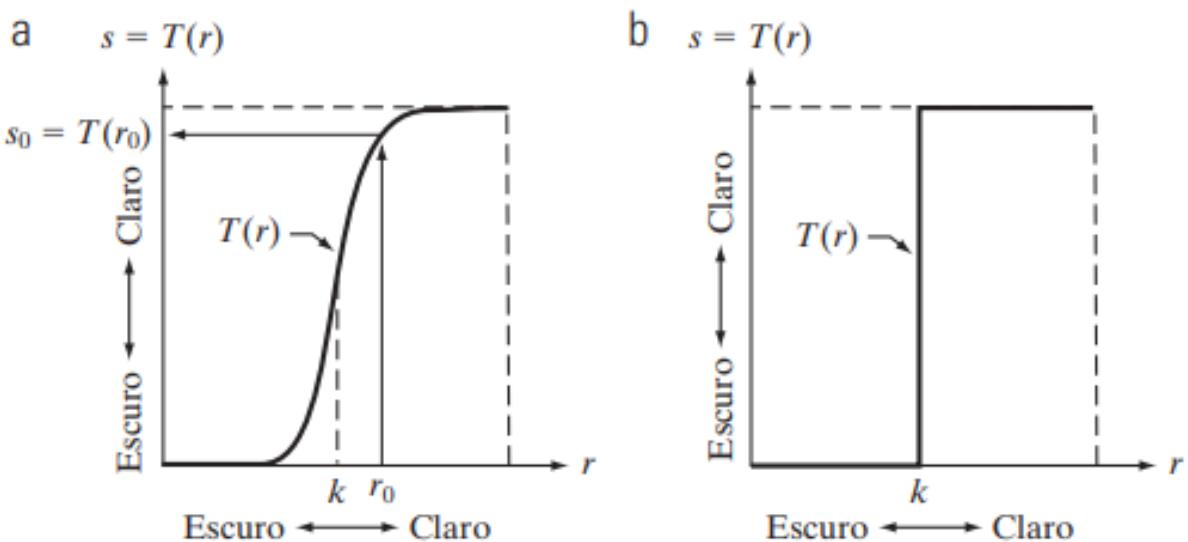


Figura 3.2: Funções para ajuste da intensidade em imagens. (a) Método de ampliação de contraste, que aumenta a diferença entre os níveis de intensidade, destacando áreas mais escuras e mais claras. (b) Método de binarização, que converte a imagem em dois níveis distintos de intensidade, geralmente preto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

do processamento.

3.2 Deep Learning

Deep Learning, ou Aprendizado Profundo, é uma subárea do aprendizado de máquina que se concentra em redes neurais artificiais com muitas camadas (ou "profundidade"). 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 e Papert (2021), 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 perceptrons 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 e Papert (2021), 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 se o neurônio "ativa" ou não. Tradicionalmente, a função de ativação usada em um perceptron simples é a função degrau, que retorna um valor binário - geralmente 0 ou 1 (HERTZ, 2018). Esse processo pode ser formalmente expresso pela fórmula:

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

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 3.3 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 a equação:

$$w_i = w_i + \Delta w_i$$

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

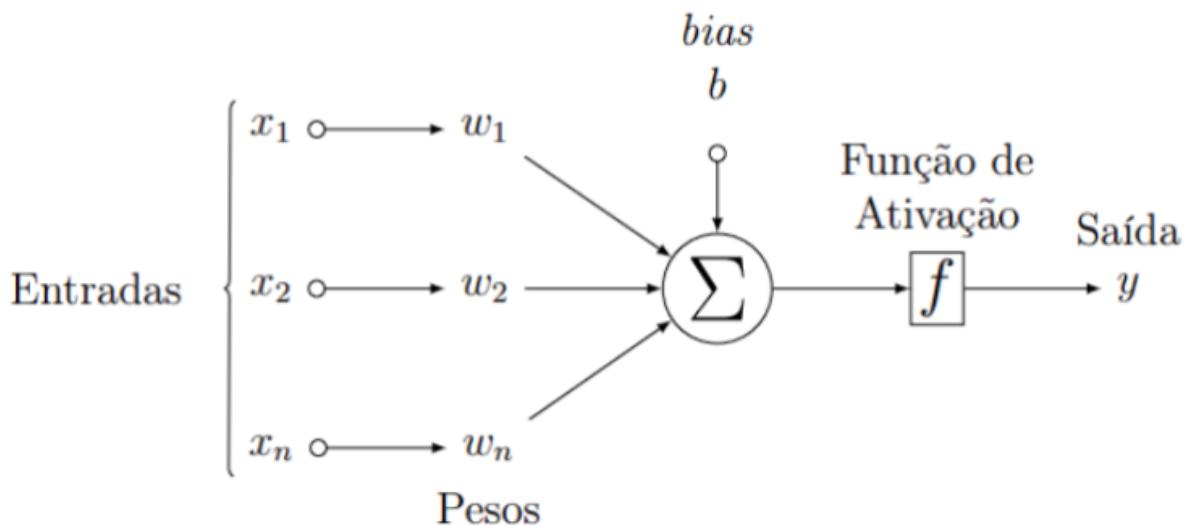


Figura 3.3: Disgrama de um perceptron, ilustrando uma abordagem simplificada baseada na estrutura e função de um neurônio biológico.

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, ele tem limitações, especialmente em relação a problemas de classificação não linear. Por exemplo, ele não consegue resolver problemas como o XOR, onde as classes não podem ser separadas por uma linha reta. Essa limitação levou ao desenvolvimento de redes neurais mais complexas, como os perceptrons multicamadas (MLPs), que usam várias camadas de neurônios para aprender representações mais complexas dos dados (BISHOP; NASRABADI, 2006).

O perceptron, no entanto, segundo Hertz (2018) 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 e classificação.

3.3 Informação de profundidade

Sensores de profundidade estão cada vez mais embarcados em equipamentos amplamente difundidos como dispositivos de realidade aumentada (Occulus, Kinect) 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).

Ainda segundo (CASTELLANO; TERRERAN; GHIDONI, 2023), cada uma das técnicas de aquisição de imagens de profundidade possui lados negativos que podem impactar os dados. Por exemplo, as câmeras ToF podem sofrer com invalidação de pixels próximos a cantos ou bordas de objetos devido à interferências entre os raios IR em superfícies des-contínuas ou reflexivas (HANSARD et al., 2012). Outros tipos de câmeras RGB-D mais comuns como o Microsoft Kinect ou Intel RealSense podem produzir valores inválidos em superfícies muito brilhantes ou reflexivas como espelhos, superfícies metálicas ou muito escuras (ZOLLHÖFER, 2019). Em ambientes internos, tais imagens podem conter até 50% de dados faltantes. (ZHANG et al., 2022) (ZHANG; FUNKHOUSER, 2018). Pontos cuja medição é desconhecida são representados por pixels totalmente pretos ou totalmente brancos (DOURADO; PEDRINO, 2020).

3.4 Modelos de estimativa de profundidade

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 viéses. 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).

Para avaliar os modelos de estimativa 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. 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 viéses e situações (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 equipamento Microsoft Kinect que utiliza a técnica de emissão de luz estruturada, que produz resultados precisos de informação de profundidade. Além dos pares RGB-D, também é disponibilizado os dados de leitura dos sensores puros 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, como observado na Figura 4.1. Entre as cenas observadas no *dataset*, podemos citar quartos, cozinhas, sala de aula, banheiro e etc. Além das informações de profundidade, a base de dados provém rótulos de segmentação de objetos e relações de suporte entre eles (LAHIRI; REN; LIN, 2024).

4.1.2 KITTI

4.1.3 SINTEL

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 e *ground truth* de profundidade capturado utilizando um

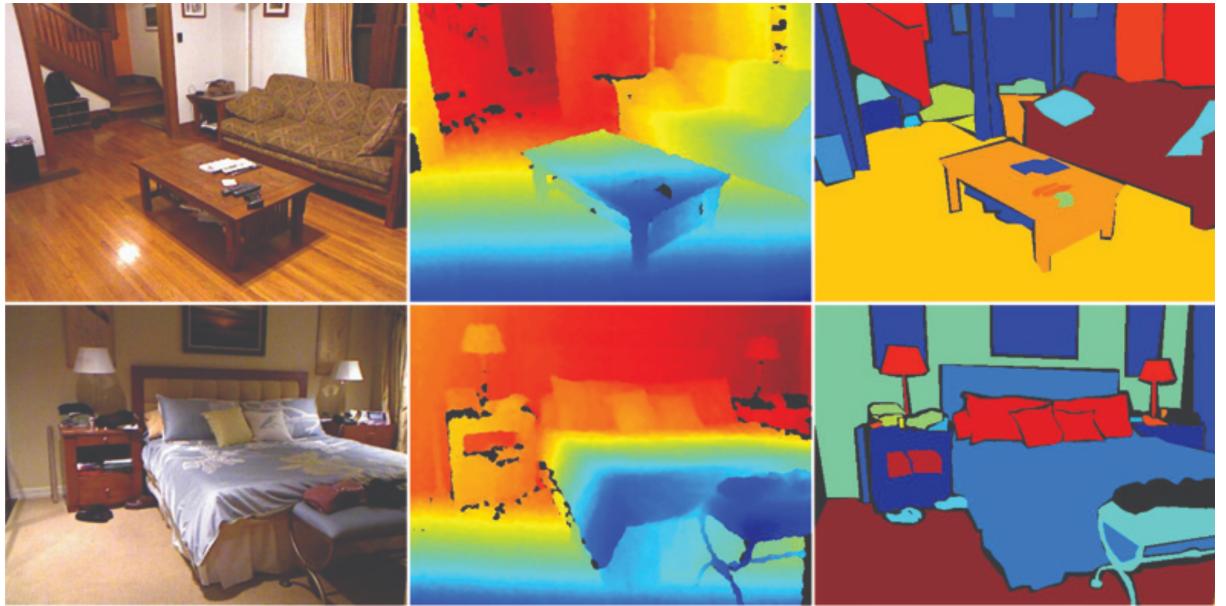


Figura 4.1: Exemplo do dataset NYU Depth v2

escâner a laser Faro Focus X 330. Oferece três versões de imagens de profundidade, uma correspondente à leitura bruta do sensor (*raw*), outra com *outliers* removidos por trabalho manual e uma ferramenta automática (*clean*) e uma com *outliers* e pontos observados por uma única imagem RGB removidos. A partição de teste não contém *groundtruth*. 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).

4.1.5 DIODE

O *dataset* DIODE (*Dense Indoor and Outdoor Depth Dataset*), é uma base de dados para estimativa monociliar de profundidade e consiste em 8574+25 imagens de ambientes internos e 16.884+446 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 escâner a laser Faro Focus S350. Alguns exemplos do *dataset* podem ser visualizados na Figura 4.2.

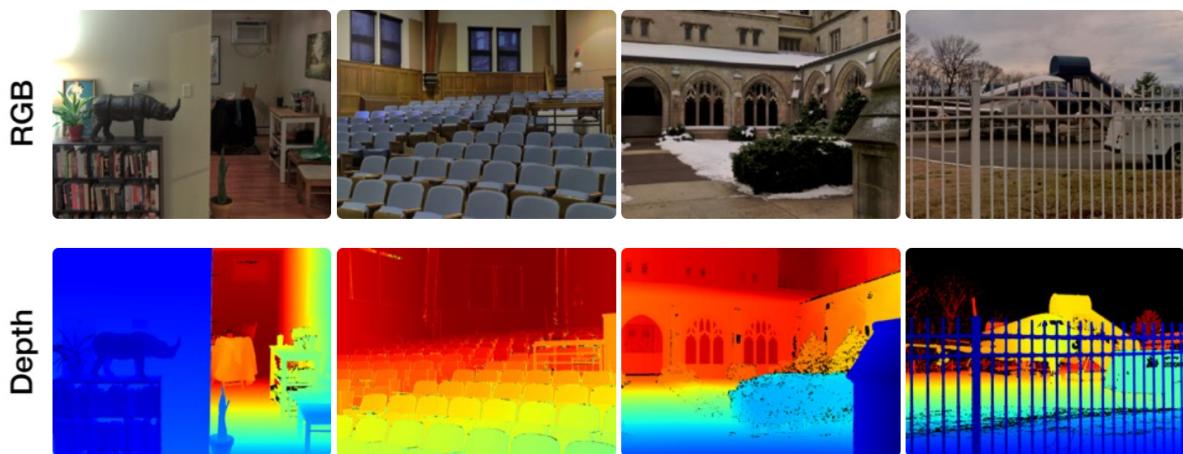


Figura 4.2: Exemplo do dataset DIODE

4.2 Modelos Escolhidos

4.3 Protocolo de Avaliação

4.4 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 milimetros), 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 transferência 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 4.3.

O método proposto diferencia-se do tradicional baseado em fator de escala e deslo-

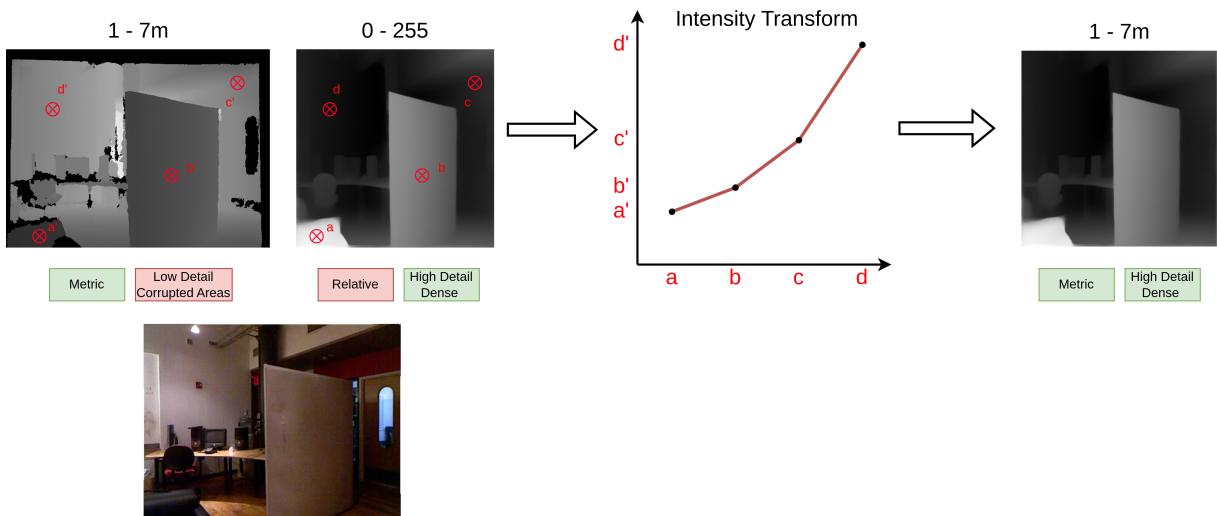


Figura 4.3: Diagrama do método de transferência de domínio

camento 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 protocolo de avaliação dos modelos de estimação de profundidade, mas sim o que é mais prevalente na literatura.

Aos conjuntos de dados que possuem leituras métricas de sensores, será comparado o resultado da técnica de pós-processamento e o resultado de estimadores métricos de profundidade.

4.5 Correção de mapas de profundidade

Para a tarefa de correção de mapas de profundidade utilizando redes neurais, o trabalho de (HU et al., 2022) propôs duas categorias principais que se diferenciam pelos dados utilizados:

- **Correção não-guiada** (Figura 4.4): Objetiva completar diretamente as partes faltantes utilizando como entrada somente o mapa de profundidade.
- **Correção guiada** (Figura 4.5 e 4.6): Objetiva completar as partes faltantes utilizando como entrada tanto o mapa de profundidade quanto a imagem RGB correspondente.

A escolha da categoria de correção depende da quantidade de erros nas imagens. Quando há uma pequena quantidade de pixels inválidos, a correção não-guiada pode ser

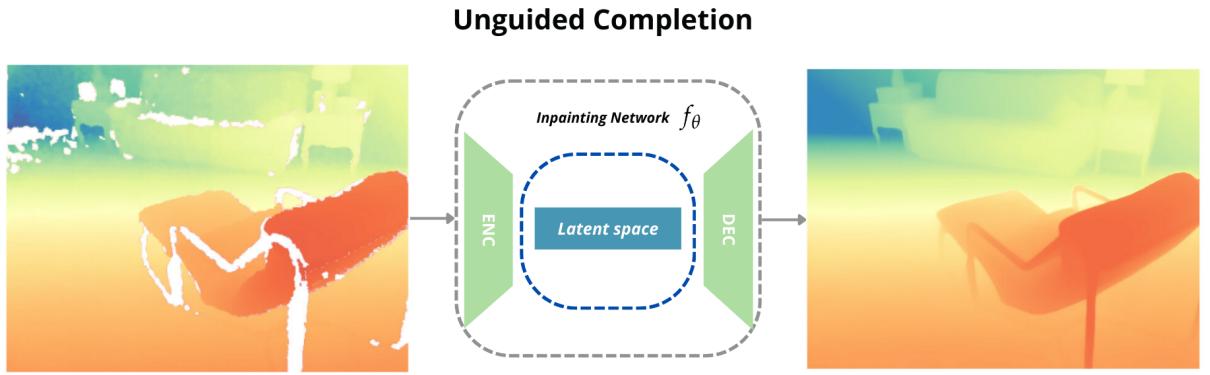


Figura 4.4: Esquema de correção não-guiada.

adequada, visto que não é necessário uma profunda extração de características dos dados. No entanto, no caso contrário, o uso de métodos guiados é indicado dado que existam grandes regiões com ausência de informação de profundidade ou que o mapa apresente uma grande esparsidão. Sendo necessário recorrer a extração dos atributos presentes na imagem RGB como bordas, contornos, estruturas de objetos não identificados pelo sensor e característica de descontinuidade de superfícies (HU et al., 2022).

Ainda no trabalho de (HU et al., 2022), nomeia-se outras subcategorias de técnicas de correção guiada. Uma delas é chamada de *Early Fusion* (Figura 4.5) e consiste em utilizar a imagem RGB concatenada ao mapa de profundidade com erros como entrada da rede neural. Essa técnica possui a vantagem de ser simples e de baixa complexidade. A outra, conhecida como *Late Fusion* (Figura 4.6) envolve transferir a fusão da imagem RGB com o mapa em ramos distintos da rede neural, chamados *RGB Encoder-Decoder* e *Depth Encoder-Decoder*.

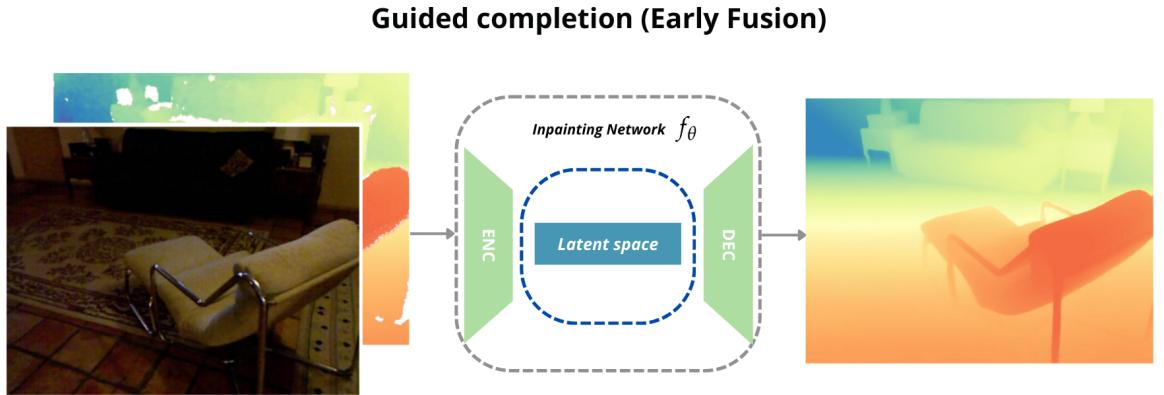


Figura 4.5: Esquema de correção guiada com *Early Fusion*.

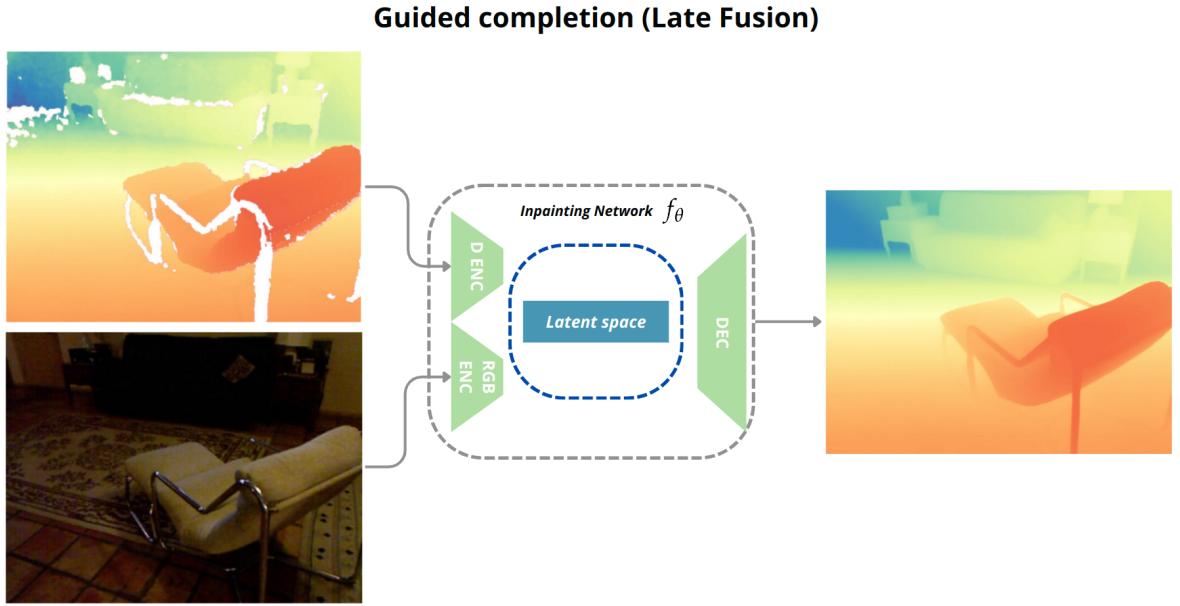


Figura 4.6: Esquema de correção guiada com *Late Fusion*.

4.5.1 Large Mask Inpainting

Image Inpainting refere-se ao processo de recuperar regiões faltantes de uma imagem a partir de informação já existente (ELHARROUSS et al., 2020). Para sintetizar as partes indicadas, é necessário que haja o aprendizado da estrutura global da imagem, sendo imprescindível um vasto campo receptivo na rede neural. Dessa forma, é proposto por (SUVOROV et al., 2022) o sistema LaMa, *Large Mask Inpainting* (Figura 4.7), que é composto por elementos capazes de explorar o campo receptivo apropriado para essa tarefa, sendo eles: i) convoluções rápidas de Fourier (do inglês, *Fast Fourier Convolutions*), ii) o uso de perda perceptual baseada em uma rede de segmentação e iii) uma estratégia de geração de máscaras para treinamento de alta cobertura.

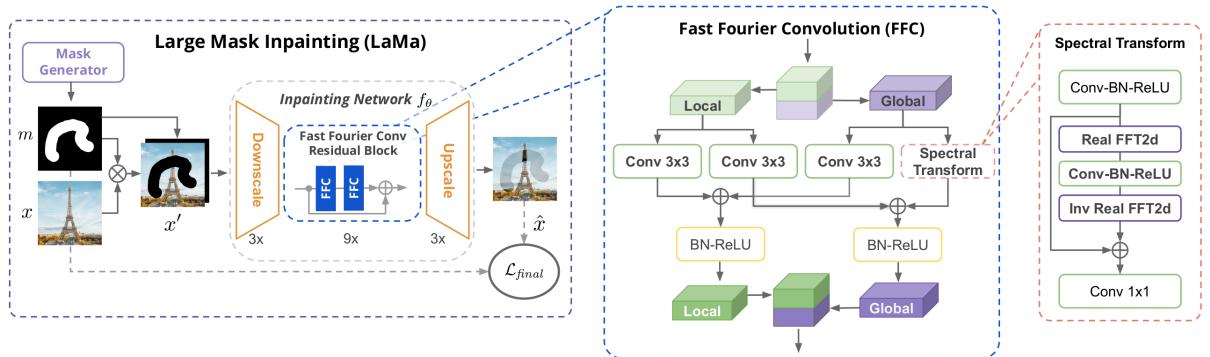


Figura 4.7: Esquema do método LaMa (SUVOROV et al., 2022).

4.6 Análise com Aplicação

4.7 Considerações Metodológicas

Capítulo 5

Resultados e Discussões

5.1 Resultados Preliminares

5.2 Resultados Esperados

Capítulo 6

Cronograma

A presente visa expor as atividades já 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

- 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/>>.
- 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.
- 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.
- ELHARROUSS, O. et al. Image inpainting: A review. *Neural Processing Letters*, Springer, v. 51, p. 2007–2028, 2020.
- 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.
- 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. 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.
- 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, 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.
- 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.
- 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.
- 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.
- 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.; PAPERT, S. *Perceptron*. [S.l.]: April, 2021. v. 9.
- 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.
- 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.
- SUVOROV, R. et al. Resolution-robust large mask inpainting with fourier convolutions. In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. [S.l.: s.n.], 2022. p. 2149–2159.

- 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.; FUNKHOUSER, T. Deep depth completion of a single rgb-d image. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2018. p. 175–185.
- 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.
- 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

...