

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

| | | |
|-----|--|----|
| 4.1 | Exemplo do dataset NYU Depth v2 | 11 |
| 4.2 | Exemplo do dataset DIODE | 12 |
| 4.3 | Diagrama do método de transferência de domínio | 13 |
| 4.4 | Esquema de correção não-guiada. | 14 |
| 4.5 | Esquema de correção guiada com <i>Early Fusion</i> | 14 |
| 4.6 | Esquema de correção guiada com <i>Late Fusion</i> | 15 |
| 4.7 | Esquema do método LaMa (SUVOROV et al., 2022). | 15 |

Lista de Tabelas

| | | |
|-----|--|----|
| 4.1 | Características dos datasets utilizados no trabalho | 10 |
| 6.1 | Cronograma com as atividades realizadas para o desenvolvimento da pesquisa do ano de 2023 | 18 |
| 6.2 | Cronograma com as atividades realizadas e pretendidas para o desenvolvimento da pesquisa do ano de 2024 e Janeiro de 2025. | 19 |

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 | 7 |
| 3.1 | Processamento Digital de Imagens | 7 |
| 3.2 | Deep Learning | 7 |
| 3.3 | Informação de profundidade | 7 |
| 3.4 | Modelos de estimativa de profundidade | 8 |
| 4 | Materiais e Métodos | 9 |
| 4.1 | Datasets | 9 |
| 4.1.1 | NYUv2 | 10 |
| 4.1.2 | KITTI | 10 |
| 4.1.3 | SINTEL | 10 |
| 4.1.4 | ETH3D | 10 |
| 4.1.5 | DIODE | 11 |
| 4.2 | Modelos Escolhidos | 12 |

Sumário

| | | |
|----------|---|-----------|
| 4.3 | Protocolo de Avaliação | 12 |
| 4.4 | Método de Transformação de Intensidades (pós-processamento) | 12 |
| 4.5 | Correção de mapas de profundidade | 13 |
| 4.5.1 | Large Mask Inpainting | 15 |
| 4.6 | Análise com Aplicação | 16 |
| 4.7 | Considerações Metodológicas | 16 |
| 5 | Resultados e Discussões | 17 |
| 5.1 | Resultados Preliminares | 17 |
| 5.2 | Resultados Esperados | 17 |
| 6 | Cronograma | 18 |
| | Referências | 20 |
| | Apêndices A – Apêndice A | 24 |

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

Em Liu, Gong e Liu (2012) foi proposto um algoritmo de *inpainting* para aprimorar os mapas de profundidade capturados com Kinect, estendendo o método original de marcha rápida (*Fast Marching Method* - FMM) para reconstruir regiões desconhecidas incorporando uma imagem RGB alinhada como guia. Em seguida, aplica-se um filtro de preservação de bordas para reduzir o ruído nas regiões de separação de objetos.

No trabalho de Zhang e Funkhouser (2018), foi desenvolvido um esquema que utiliza uma rede neural para inferir as normais de superfície e os limites de oclusão a partir de uma imagem RGB. Essas predições são combinadas com mapas de profundidade de câmeras RGB-D através de um método de otimização para computar a profundidade resultante de todos os pixels da imagem. A rede neural possui arquitetura totalmente convolucional utilizando a VGG-16 como *backbone* e é treinada com as normais de superfície e limites de oclusão computados a partir da renderização de uma malha tridimensional reconstruída a partir de múltiplos ângulos de visão.

Dourado e Pedrino (2020) introduz o método NSGA2CGP, uma abordagem de otimização multi-objetivo que integra programação cartesiana genética para a otimização de filtros morfológicos em escala de cinza para completar mapas de profundidade utilizados para algoritmo detector de caminho livre. O objetivo é minimizar tanto os erros quanto a complexidade dos elementos estruturantes dado as limitações energéticas dos sistemas de navegação embarcados para pessoas com deficiência visual. Além do erro, também foram mensurados na aplicação, o consumo de energia e tempo de execução.

É proposto por Fujii, Hachiuma e Saito (2020) um método para *inpainting* de imagens RGB-D utilizando uma rede generativa adversarial (*Generative Adversarial Network* - GAN) objetivando restaurar simultâneamente a textura e geometria de regiões faltantes

levando em consideração as informações complementares de cor e profundidade com uma abordagem de fusão tardia, resultando na restauração tanto dos canais RGB quanto de mapas de profundidade.

Com o objetivo de complementar o trabalho de Zhang e Funkhouser (2018), os autores Huang et al. (2019) desenvolveram um *framework* para completar mapas de profundidade buscando preservar a clareza das bordas dos objetos mantendo a estrutura da imagem, evitando o cenário onde as redes neurais aprendem meramente a interpolar os valores de profundidade. É empregado o mecanismo de atenção própria para reunir informação das características de normais de superfície e limites de oclusão. Os autores afirmam que são alcançados tempos de execução menores em relação ao estado da arte.

Em Rho, Ha e Kim (2022), é apresentada uma arquitetura para correção de mapas de profundidade esparsos em três estágios. Uma estrutura dupla de *encoder-decoder* baseada em *transformers* para extrair características dos *tokes* das imagens RGB e mapas de profundidade esparsos. Um módulo de atenção guiada (GAM, do inglês *Guided-Attention Module*) para fusionar os dados das duas modalidades distintas. Um método para fusionar os resultados dos ramos e capturar as dependências intermodais.

Ainda utilizando GANs, Wang et al. (2022) propôs uma rede de dois ramos projetada para estimar mapas de profundidade completos a partir de pares de mapas de profundidade incompletos e imagens RGB. No primeiro ramo, uma estrutura de *encoder-decoder* é empregada para predizer valores de profundidade densos. No segundo ramo, é utilizada uma estrutura de GAN que possui como entrada as características do mapa incompleto com a imagem RGB atuando como condicionamento para gerar uma predição de mapa de profundidade denso e um mapa de confiança, que é avaliado por uma rede discriminadora do mapa real. Um módulo de fusão adaptativa chamado W-AdaIN é empregado para propagar características entre os ramos.

Um módulo de codificação esparsa de convolução espacial aliado a filtragem bilateral é introduzido por Wu et al. (2022). Primeiramente um dicionário convolucional e uma codificação esparsa são aprendidos pela rede via *self-supervised learning* para preencher pequenas áreas do mapa incompleto, resultando em uma imagem de profundidade inicial. Em seguida, um filtro conjunto bilateral hierárquico é construído utilizando a imagem RGB correspondente para preencher partes maiores de dados faltantes, dado que valores de profundidade em *pixels* adjacentes são similares em partes com cores parecidas.

Com o foco em corrigir dados de profundidade de câmeras ToF equipadas em *smartphones*, o trabalho de Zhang et al. (2022) desenvolveu uma arquitetura de redes neurais

convolucionais baseadas em dois ramos *encoder-decoder* com *skip-connections* entre os estágios do *decoder* para cada estágio correspondente dos dois *encoders*. Sendo um dos ramos para extração de característica da imagem RGB inicializado com os pesos do *dataset* ImageNet e congelado durante o processo de treinamento, e outro para o mapa de profundidade incompleto. É empregado também um módulo de decodificação com dilatação para reconstruir a imagem a partir de dados de uma área maior. É proposta uma função de perda híbrida baseada na imposição da geometria da cena e um método de aumento de dados para remoção de artefatos.

Zhang et al. (2023) desenvolve uma abordagem para correção de mapas de profundidade esparsos utilizando uma arquitetura que combina *transformers*, levando em conta características globais e extração de características locais via CNN. A estrutura da rede possui um único ramo em formato *encoder-decoder* piramidal com *skip-connections* entre seus estágios correspondentes. Como unidade fundamental do modelo de correção de profundidade, é proposto um bloco composto por unidades convolucionais com módulos de atenção e *transformers*(*Convolutional Attention and Transformer Block*, JCAT).

No trabalho de Ran, Yuan e Shibasaki (2023), os autores introduziram um paradigma de aprendizado por poucas amostras para correção de mapas de profundidade esparsos. Primeiramente um DDPM é inicialmente pré-treinado em imagens RGB sem dados de profundidade para servir como *backbone* da rede. Em um segundo estágio, é empregado um modelo de fusão baseado na operação de convolução guiada que tem como entrada um mapa de profundidade esparso, sendo capaz de inferir mapas de profundidade densos, incorporando informação de múltiplas escalas extraído do estágio DDPM.

Capítulo 3

Fundamentação Teórica

3.1 Processamento Digital de Imagens

3.2 Deep Learning

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 estimação 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 | Esparça | 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

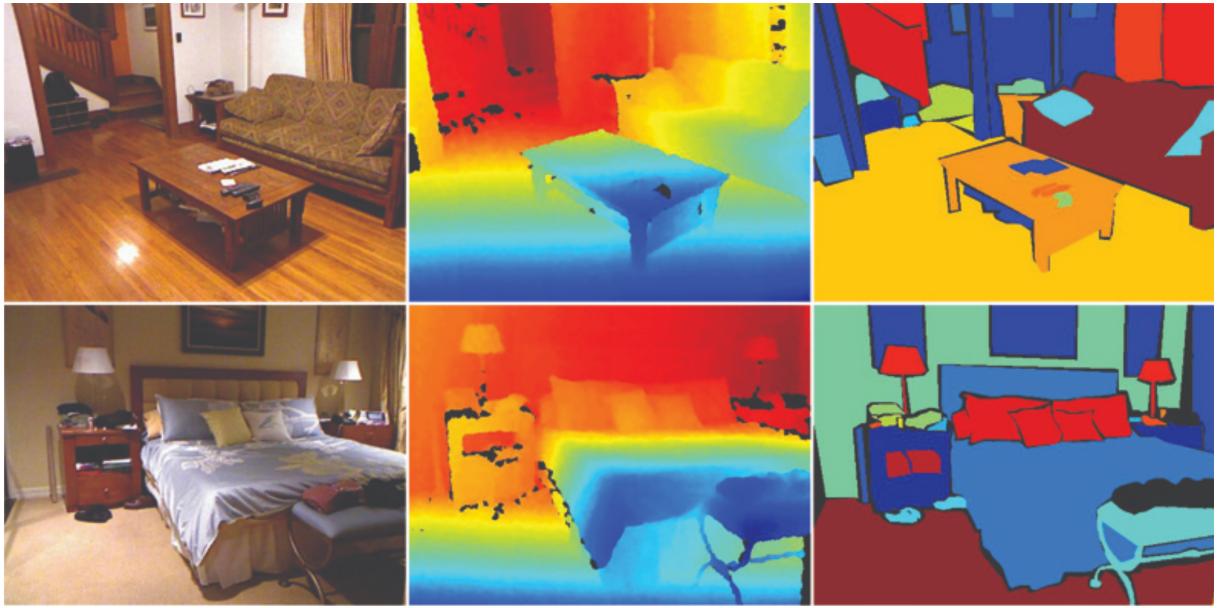


Figura 4.1: Exemplo do dataset NYU Depth v2

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 monocular 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.

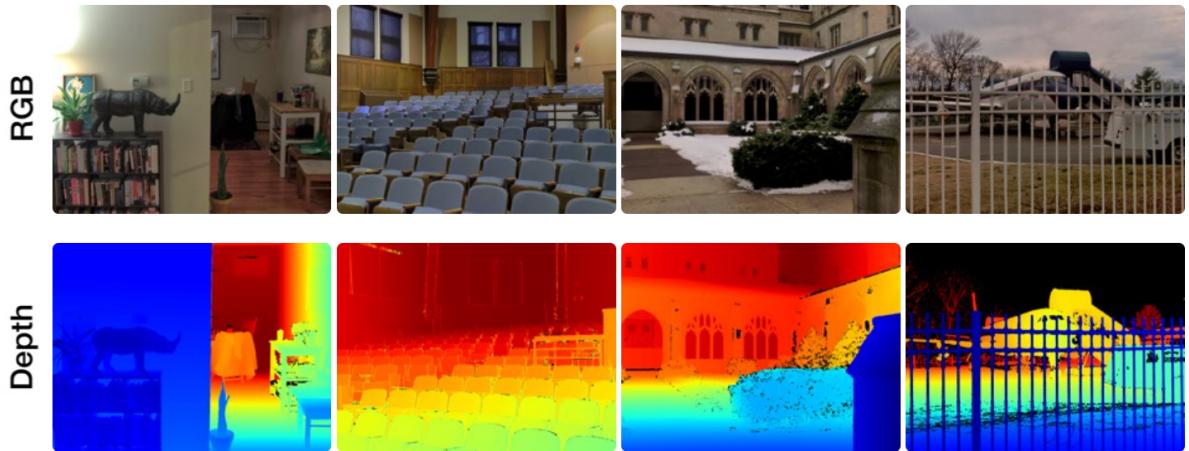


Figura 4.2: Exemplo do dataset DIODE

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 esparsidade. 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-

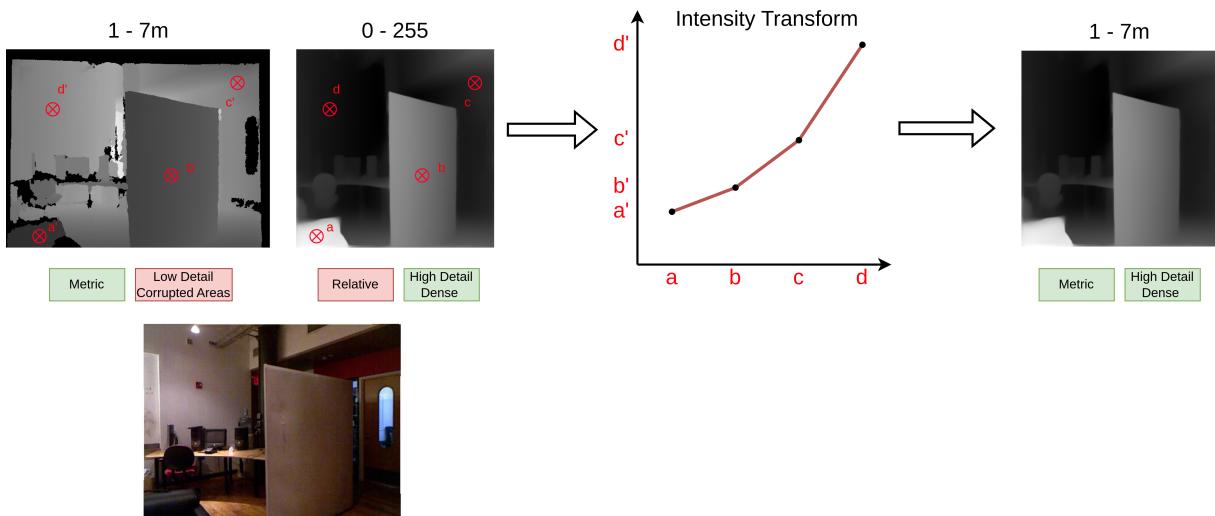


Figura 4.3: Diagrama do método de transferência de domínio

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 estimativa 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

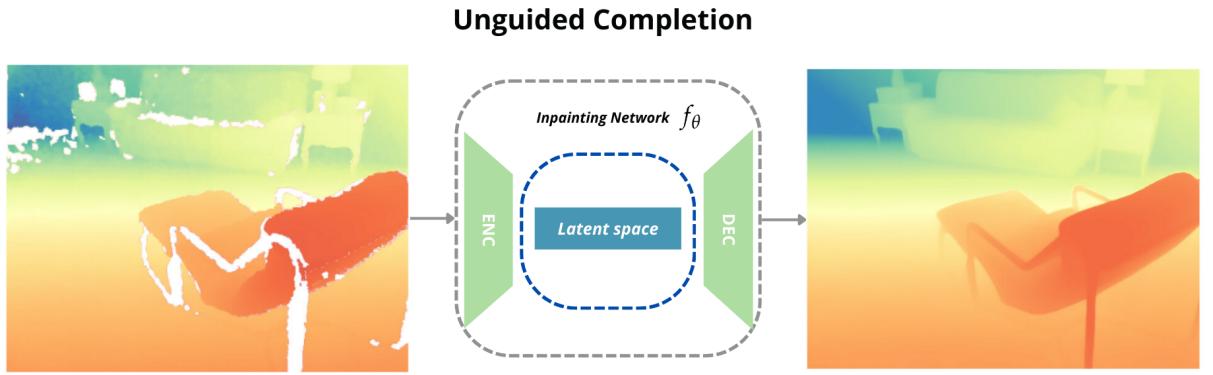


Figura 4.4: Esquema de correção não-guiada.

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*.

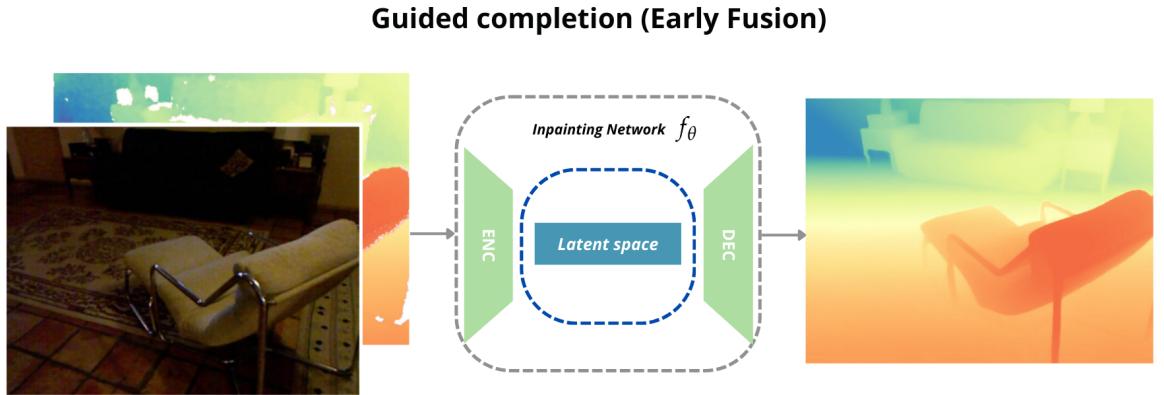


Figura 4.5: Esquema de correção guiada com *Early Fusion*.

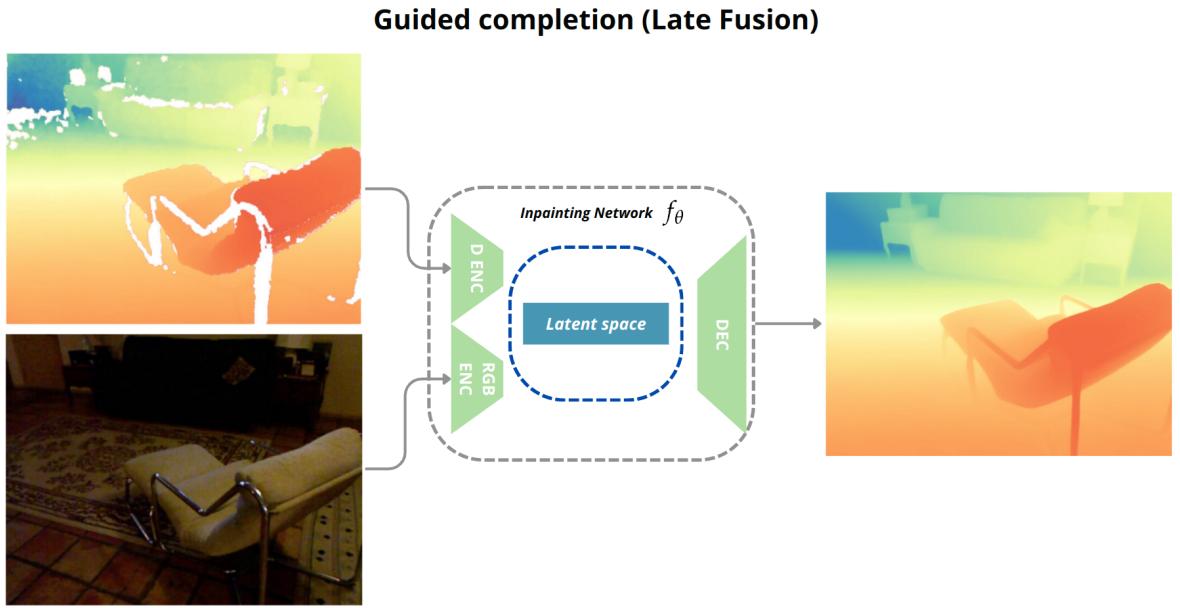


Figura 4.6: Esquema de correção guiada com *Late Fusion*.

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.

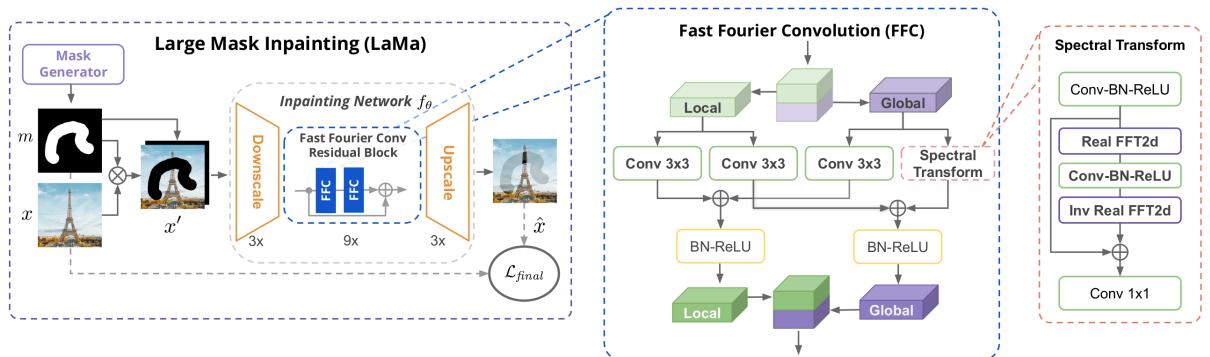


Figura 4.7: Esquema do método LaMa (SUVOROV et al., 2022).

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.
- 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.
- FUJII, R.; HACHIUMA, R.; SAITO, H. Rgb-d image inpainting using generative adversarial network with a late fusion approach. In: SPRINGER. *International Conference on Augmented Reality, Virtual Reality and Computer Graphics*. [S.l.], 2020. p. 440–451.
- 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.

- HANSARD, M. et al. *Time-of-flight cameras: principles, methods and applications*. [S.l.]: Springer Science & Business Media, 2012.
- 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.
- HUANG, Y.-K. et al. Indoor depth completion with boundary consistency and self-attention. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. [S.l.: s.n.], 2019. p. 0–0.
- 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.
- LIU, J.; GONG, X.; LIU, J. Guided inpainting and filtering for kinect depth maps. In: IEEE. *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*. [S.l.], 2012. p. 2055–2058.
- 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.
- 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.

- RAN, W.; YUAN, W.; SHIBASAKI, R. Few-shot depth completion using denoising diffusion probabilistic model. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2023. p. 6558–6566.
- 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.
- RHO, K.; HA, J.; KIM, Y. Guideformer: Transformers for image guided depth completion. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2022. p. 6250–6259.
- 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.
- WANG, H. et al. Rgb-depth fusion gan for indoor depth completion. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. [S.l.: s.n.], 2022. p. 6209–6218.
- WU, H. et al. Joint self-supervised and reference-guided learning for depth inpainting. *Computational Visual Media*, Springer, v. 8, n. 4, p. 597–612, 2022.
- 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. Completionformer: Depth completion with convolutions and vision transformers. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2023. p. 18527–18536.
- 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

...