

UNIVERSIDADE FEDERAL DO ACRE

**Gustavo Moreira Oliveira de Castro**

**Estimação de mapas de profundidade em imagens de câmeras monoculares  
utilizando modelos de difusão**

**RIO BRANCO  
2024**

UNIVERSIDADE FEDERAL DO ACRE

Gustavo Moreira Oliveira de Castro

Estimação de mapas de profundidade em imagens de câmeras monoculares  
utilizando modelos de difusão

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

Orientadora:  
Prof. Dr. Roger Fredy Larico Chavez

RIO BRANCO  
2024

Gustavo Moreira Oliveira de Castro

Estimação de mapas de profundidade em imagens de câmeras monoculares  
utilizando modelos de difusão

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

**Estimação de mapas de profundidade em imagens de câmeras monoculares  
utilizando modelos de difusão**

...

**Palavras-chave:** ...; ...; ...; ....

# Abstract

....

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

# Listas de Figuras

1.1	Exemplo de mapa de profundidade do <i>dataset Nyu Depth V2</i> . No primeiro quadro, a imagem RGB, no segundo, o mapa de profundidade em escala de cinza, no terceiro uma colorização artificial para o mapa de profundidade.	2
1.2	Exemplo de imagem RGB com mapa de profundidade apresentando leituras inválidas.	3
4.1	Esquema de correção não-guiada.	8
4.2	Esquema de correção guiada com <i>Early Fusion</i> .	9
4.3	Esquema de correção guiada com <i>Late Fusion</i> .	10
4.4	Esquema do método LaMa (SUVOROV et al., 2022).	10

# **Lista de Tabelas**

# Sumário

<b>1</b>	<b>Introdução</b>	<b>1</b>
1.1	Motivação . . . . .	3
1.2	Objetivos . . . . .	3
1.2.1	Objetivo Geral . . . . .	3
1.2.2	Objetivos Específicos . . . . .	3
<b>2</b>	<b>Trabalhos Relacionados</b>	<b>5</b>
<b>3</b>	<b>Marco Teórico</b>	<b>7</b>
3.1	. . . . .	7
<b>4</b>	<b>Materiais e Métodos</b>	<b>8</b>
4.0.1	Correção de mapas de profundidade . . . . .	8
4.0.2	Large Mask Inpainting . . . . .	9
4.1	Datasets . . . . .	10
4.2	Metodologia da Pesquisa . . . . .	11
<b>5</b>	<b>Resultados Preliminares</b>	<b>12</b>
	Referências	13
	Apêndices A – MEUS ARTIGOS?	16

# Capítulo 1

## Introdução

Informação de profundidade é uma das representações mais úteis para o entendimento de ambientes físicos (LASINGER et al., 2019) (ZHOU; KRAHÉNBUHL; 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), como exemplificado na Figura 1.1, que mostra uma cena em imagem RGB, seu mapa de profundidade e uma versão colorizada. Pontos cuja medição é desconhecida são representados por pixels totalmente pretos ou totalmente brancos (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).



Figura 1.1: Exemplo de mapa de profundidade do *dataset Nyu Depth V2*. No primeiro quadro, a imagem RGB, no segundo, o mapa de profundidade em escala de cinza, no terceiro uma colorização artificial para o mapa de profundidade.

Sensores de profundidade estão cada vez mais embarcados em equipamentos amplamente difundidos como dispositivos de realidade aumentada (Oculus, 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 (BRANS-COMBE, 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 próximo a cantos ou bordas de objetos devido à interferências entre os raios IR em superfícies descontínuas ou reflexivas (HANSARD et al., 2012), exemplificado na Figura 1.2. 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 (ZOLLMÖFER, 2019). Em ambientes internos, tais imagens podem conter até 50% de dados faltantes. (ZHANG et al., 2022) (ZHANG; FUNKHOUSER, 2018).

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 (HU et al., 2012).

Ainda de acordo com (HU et al., 2012), o problema de preenchimento de valores in-

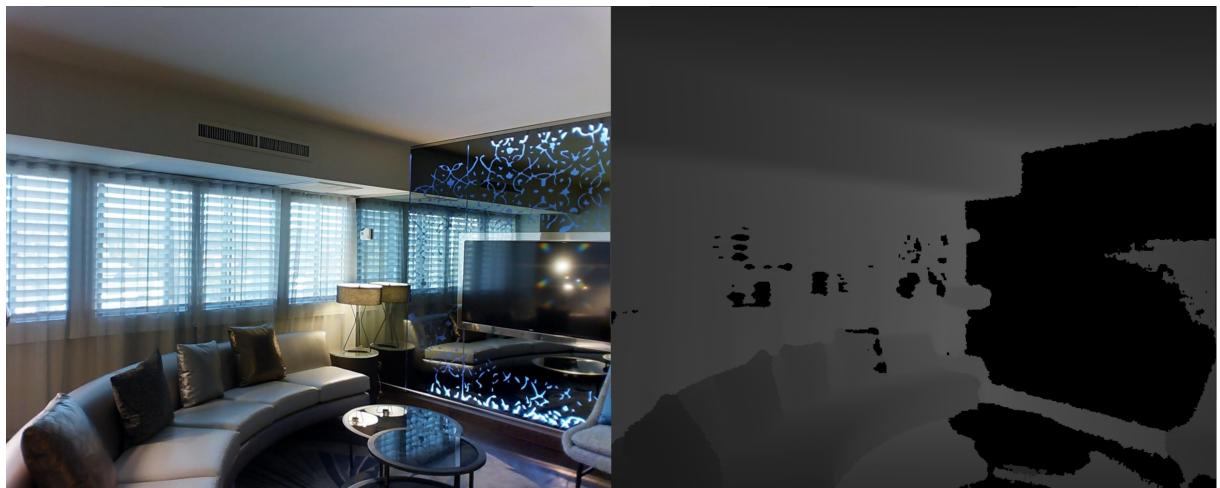


Figura 1.2: Exemplo de imagem RGB com mapa de profundidade apresentando leituras inválidas.

determinados em mapas de profundidade depende inteiramente do sensor utilizado para sua captura. No caso do LiDAR, são produzidos mapas esparsos (approx. 95% de esparcidade) e no caso de câmeras RGB-D ou câmeras ToF são produzidos mapas com partes faltantes em determinadas superfícies ou bordas. Os pixels inválidos normalmente são representados pelo valor 0 (DOURADO; PEDRINO, 2020). De acordo com (ZHANG; FUNKHOUSER, 2018), o problema de correção de imagens de profundidade implica em fazer novas previsões para áreas em que o sensor não retornou dados válidos.

Neste cenário, o presente trabalho apresenta uma abordagem baseada em técnicas de *inpainting* utilizando redes neurais artificiais para realizar a correção de erros de mapas de profundidade capturados por câmeras RGB-D e sensores ToF.

## 1.1 Motivação

## 1.2 Objetivos

### 1.2.1 Objetivo Geral

Este trabalho possui como objetivo geral realizar a correção de mapas de profundidade com erros através de redes neurais artificiais e técnicas de *inpainting*.

### 1.2.2 Objetivos Específicos

- Realizar revisão de literatura de métodos de correção de mapas de profundidade.

- Escolher um método de *inpainting* baseado em redes neurais artificiais.
- Estabelecer um método de geração de máscaras de erro para treinamento.
- Corrigir mapas de profundidade com algoritmo de *inpainting* baseado em correção não-guiada.
- Corrigir mapas de profundidade com algoritmo de *inpainting* baseado em correção guiada.
- Corrigir mapas de profundidade utilizando abordagem morfológica.
- Comparar as diferentes abordagens através de métricas de avaliação perante a diferentes níveis de pixels inválidos.

# Capítulo 2

## Trabalhos Relacionados

acho que vale a pena organizar os trabalhos por método empregado, i) métodos determinísticos (não inteligentes), ii) redes neurais convolucionais, iii) transformers, iv) modelos geratitivos (gans e diffusion e lama e oq mais tiver)

Em Liu, Gong e Liu (2012) é 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, um filtro de preservação de bordas é aplicado 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 otimizar 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.

Buscando 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 enquanto mantém 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 *tokens* 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. **talvez não tenha ficado bem explicado**

Ainda utilizando GANs, Wang et al. (2022) introduziram

# Capítulo 3

## Marco Teórico

### 3.1

# Capítulo 4

## Materiais e Métodos

### 4.0.1 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 4.1): Objetiva completar diretamente as partes faltantes utilizando como entrada somente o mapa de profundidade.
- **Correção guiada** (Figura 4.2 e 4.3): Objetiva completar as partes faltantes utilizando como entrada tanto o mapa de profundidade quanto a imagem RGB correspondente.

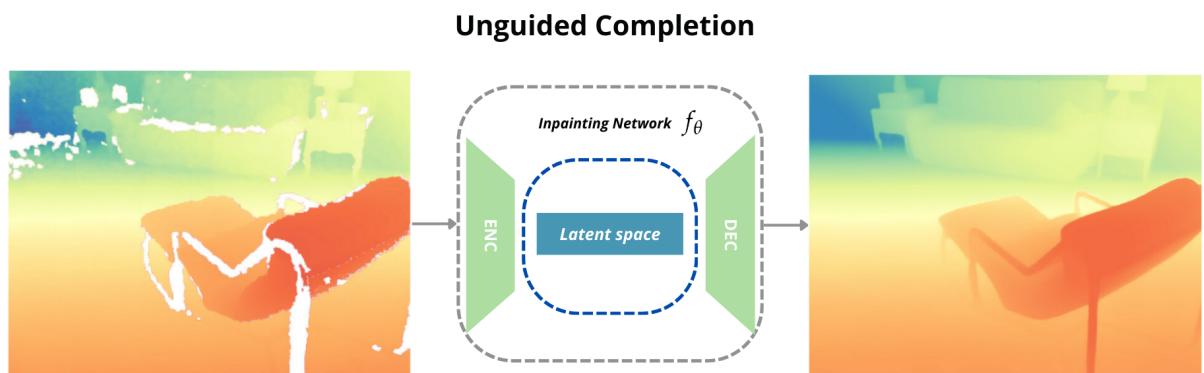


Figura 4.1: Esquema de correção não-guiada.

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 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 4.2) 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 4.3) envolve transferir a fusão da imagem RGB com o mapa em ramos distintos da rede neural, chamados *RGB Encoder-Decoder* e *Depth Encoder-Decoder*.

**Guided completion (Early Fusion)**

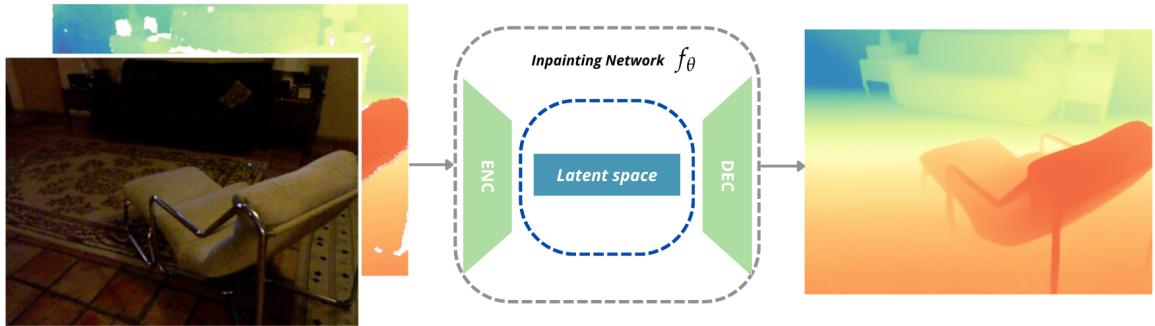


Figura 4.2: Esquema de correção guiada com *Early Fusion*.

#### 4.0.2 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 4.4), 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.

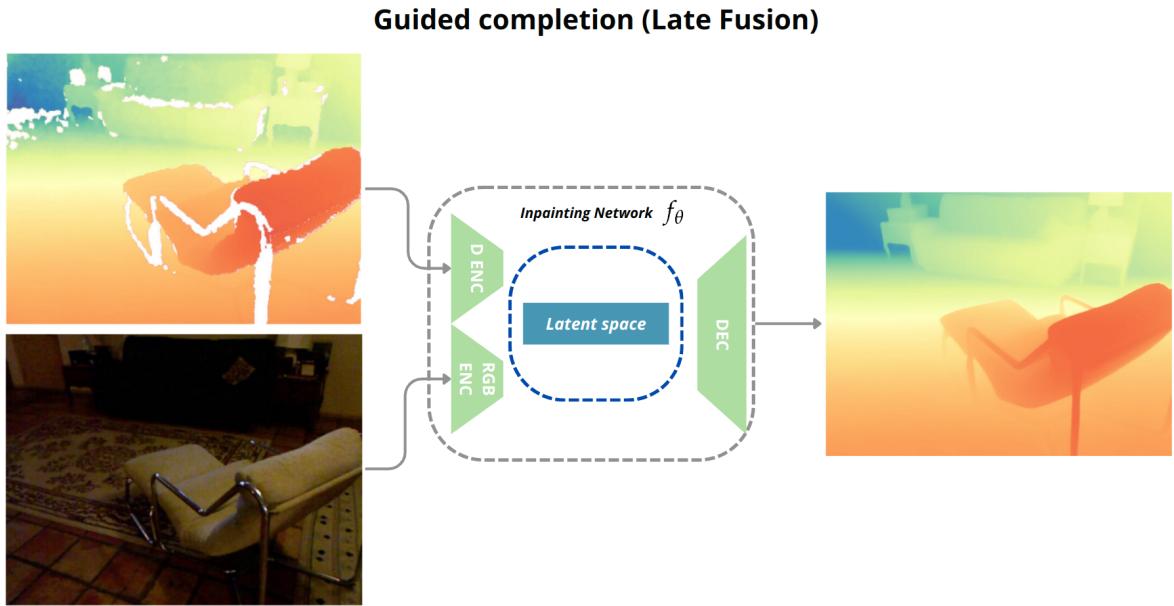


Figura 4.3: Esquema de correção guiada com *Late Fusion*.

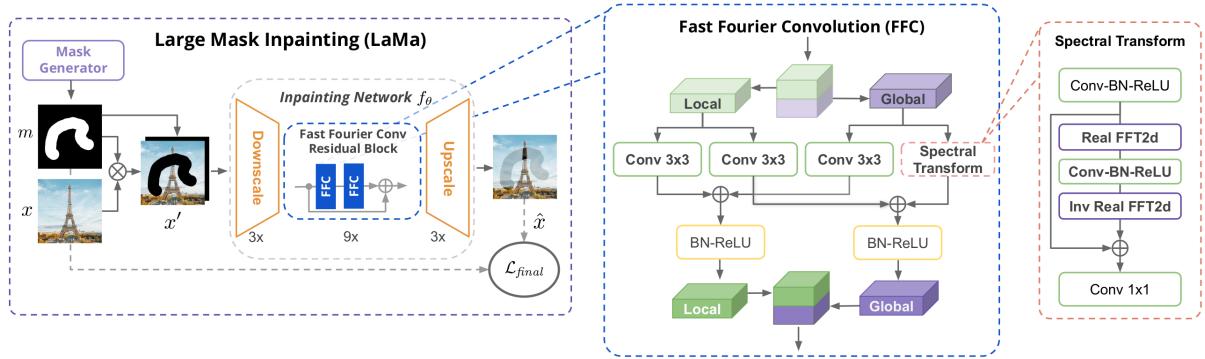


Figura 4.4: Esquema do método LaMa (SUVOROV et al., 2022).

## 4.1 Datasets

O presente trabalho exige um tipo de base de dados pouco encontrado na literatura, trios de imagem RGB, mapa de profundidade com erros e um outro mapa de profundidade denso e completo. De acordo com (ZHANG; FUNKHOUSER, 2018), uma das maneiras de se obter esses dados seria capturar imagens com uma câmera RGB-D de baixo custo e alinha-las com outra captura simultânea de um sensor mais preciso, porém essa abordagem é muito custosa, além de que não há disponibilidade de grandes conjuntos para treinamento.

O presente projeto pretende utilizar como base de dados principal o **Hypersim**. Um dataset para compreensão de cenas baseado em cenas sintéticas fotorrealistas. Contendo 77.400 imagens de 461 cenas *indoor* com pares de RGB e mapas de profundidade calculado

deterministicamente, além de outras informações como normais de superfície, rótulos de segmentação e detecção de objetos e entre outros (ROBERTS et al., 2021).

## 4.2 Metodologia da Pesquisa

# Capítulo 5

## Resultados Preliminares

# Referências

- 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.
- 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.
- 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.
- 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.
- 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.
- ROBERTS, M. et al. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In: *Proceedings of the IEEE/CVF international conference on computer vision*. [S.l.: s.n.], 2021. p. 10912–10922.
- 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.
- 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.
- 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.

- 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.
- 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 - MEUS ARTIGOS?

...