UNIVERSIDADE FEDERAL FLUMINENSE

ROGER DANTE RIPAS MAMANI

## MELHORIA DA ESTIMATIVA DE PROFUNDIDADE MONOCULAR USANDO O MODELO DE REFLEXÃO DE PHONG

NITERÓI 2022

#### ROGER DANTE RIPAS MAMANI

### MELHORIA DA ESTIMATIVA DE PROFUNDIDADE MONOCULAR USANDO O MODELO DE REFLEXÃO DE PHONG

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do Grau de Mestre em Computação. Área de concentração: CIÊNCIA DA COMPUTAÇÃO.

Orientador: LEANDRO AUGUSTO FRATA FERNANDES

> NITERÓI 2022

#### Ficha catalográfica automática - SDC/BEE Gerada com informações fornecidas pelo autor

R588m Ripas Mamani, Roger Dante Melhoria da Estimativa de Profundidade Monocular usando o Modelo de Reflexão de Phong / Roger Dante Ripas Mamani. -2022. 76 f. Orientador: Leandro Augusto Frata Fernandes. Dissertação (mestrado)-Universidade Federal Fluminense, Instituto de Computação, Niterói, 2022. 1. Visão computação, Niterói, 2022. 1. Visão computaçãonal. 2. Estimativa de Profundidade Monocular. 3. Modelo de Reflexão de Phong. 4. Produção intelectual. I. Fernandes, Leandro Augusto Frata, orientador. II. Universidade Federal Fluminense. Instituto de Computação. III. Título. CDD - XXX

Bibliotecário responsável: Debora do Nascimento - CRB7/6368

#### ROGER DANTE RIPAS MAMANI

#### MELHORIA DA ESTIMATIVA DE PROFUNDIDADE MONOCULAR USANDO O MODELO DE REFLEXÃO DE PHONG

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do Grau de Mestre em Computação. Área de concentração: CIÊNCIA DA COMPUTAÇÃO.

Aprovada em outubro de 2022.

BANCA EXAMINADORA Prof. Leandro Augusto Frata Fernandes - Orientador, UFF Prof. Esteban Walter-Gonzalez Clua, UFF

Prof. Raquel Esperanza Patiño Escarcina, Universidad Católica San Pablo

Niterói 2022

Dedicatória(s): Dedico este trabalho a Deus, a meus pais Amalia e Mario, a minha irmã Karen e a minha namorada Brigitte por sempre me dar forças e seu apoio ao longo desta longa jornada. Sem todos eles não teria conseguido concluir esta etapa.

# Agradecimentos

Agradeço aos professores dos cursos do mestrado por me permitirem descobrir novos horizontes e principalmente ao meu orientador Leandro pelo apoio e dedicação incondicionais para que este trabalho possa dar resultados muito interessantes e que me incentivou ainda mais a continuar pesquisando.

Gostaria de agradecer também ao CAPES pelo apoio financeiro para que este trabalho pudesse ser realizado.

### Resumo

A estimativa de profundidade com base em uma única imagem RGB é uma tarefa importante e desafiadora, com aplicações em robótica, veículos autônomos e outras áreas. Com o avanço do aprendizado profundo, várias abordagens de estimativa de profundidade de imagem única surgiram com resultados notáveis. No entanto, ainda é possível observar deficiências nos mapas de profundidade gerados pelas técnicas atuais. Dois problemas comuns são a irregularidade das superfícies planas e a definição borrada das bordas da superfície em diferentes profundidades, o que pode levar à fusão de objetos distintos que se tornam indistinguíveis no mapa de profundidade. Esta dissertação apresenta uma rede neural profunda que usa o resultado produzido por alguma abordagem de estimativa de profundidade de imagem única existente e o aprimora adicionando os detalhes que o mapa de profundidade requer para ser nítido. Treinamos a rede neural de realce de profundidade (Depth Enhancer Neural Network - DENN) proposta usando uma nova função de perda que compara a imagem colorida de entrada da cena com a imagem colorida produzida pela renderização da cena usando o modelo de reflexão de Phong. Para renderização, calculamos os vetores normais das superfícies do mapa de profundidade aprimorado, calculamos as direções da luz por pixel e estimamos o parâmetro de reflexão difusa da superfície em cada pixel usando um modelo de estimativa de albedo. As direções de luz foram calculadas por pixel como uma etapa de pré-processamento para ser usada no treinamento para que não sejam recalculadas a cada vez em cada época do treinamento. Nossos experimentos mostram uma clara melhoria na nitidez da imagem de profundidade produzida pelo DENN, levando ao aprimoramento de bordas e regularidade de superfícies planas sem comprometer objetos não planares.

**Palavras-chave**: Estimativa de Profundidade Monocular — Albedo — Modelo de Reflexão de Phong — Visão Computacional

### Abstract

Depth estimation based on a single RGB image is an important and challenging task, with applications in robotics, autonomous vehicles, and other areas. With the advance of deep learning, several single-image depth estimation approaches have emerged with remarkable results. However, it is still possible to observe deficiencies in the depth maps generated by current techniques. Two common issues are the irregularity of planar surfaces and the blurred definition of surface edges at different depths, which can lead to merging distinct objects that become indistinguishable in the depth map. This dissertation presents a deep neural network that takes the result produced by some existing single-image depth estimation approach and enhances it by adding the details that the depth map requires to be sharp. We train the proposed Depth Enhancer Neural Network (DENN) using a new loss function that compares the input color image of the scene to the color image produced by rendering the scene using the Phong reflection model. For rendering, we compute the normal vectors of the surfaces from the enhanced depth map, calculate light directions per pixel and estimate the surface diffuse reflection parameter on each pixel using an albedo estimation model. Light directions were calculated per pixel as a preprocessing stage to be used in training so they are not recalculated each time in each training epoch. Our experiments show a clear improvement in the sharpness of the depth image produced by the DENN, leading to edge enhancement and regularity of planar surfaces without compromising non-planar objects.

**Keywords**: Monocular Depth Estimation — Albedo — Phong Reflection Model — Computer Vision

# Lista de Figuras

1	Dispositivos especiais para obter mapas de profundidade	12
2	Detalhes imprecisos nas bordas e superfícies das imagens em profundidade. Profundidades verdadeiras (linha superior) e profundidades geradas (linha inferior) a partir de algumas abordagens (Bhat, Alhashim e Wonka (2021), Yin et al. (2019))	13
3	(a) Imagem original. Representações 2D da profundidade: (b) em tons de cinza; (c) em mapa de cor Jet; (d) em mapa de cor Magma	16
4	(a) Imagem RGB; (b) Albedo	17
5	Mapas de importância gerados a partir de uma imagem RGB. Fonte: Choi et al. (2021)	18
6	Superfícies mostrando uma paleta de cores usada para representar mapas de normais. Fonte: Autoria própria	19
7	(a) Imagem RGB; (b) Mapa de normais	19
8	Visualização dos componentes do modelo de Phong e do modelo resultante. Fonte: Brad Smith, Wikimedia Commons.	20
9	Exemplo de uma arquitetura de rede neural convolucional. Fonte: Tabian, Fu e Sharif Khodaei (2019).	21
10	Exemplo de operação de convolução sobre os pixels uma imagem. Fonte: Autoria própria	21
11	O diagrama de fluxo de treinamento do DENN. Os retângulos pontilhados e tracejados denotam o estágio de pré-processamento das direções da luz e a inferência do modelo, respectivamente.	34
12	Exemplos de mapas de profundidade produzidos pelas técnicas de Yin et al. (2019) (linhas 2 e 3), Bhat, Alhashim e Wonka (2021) (linhas 4 e 5) e Song, Lim e Kim (2021) (linhas 6 e 7) em escalas de cinza e mapas de cores.	36

	٠	٠
V	1	1
V	I	L

13	<ul> <li>(a) Imagem RGB; (b-d) Exemplos de albedos gerados por algumas técnicas:</li> <li>(b) Nestmeyer e Gehler (2017); (c) Lettry, Vanhoey e Van Gool (2018);</li> <li>(d) Yunfei Liu et al. (2020)</li></ul>	37
14	Exemplos de imagens necessárias: (a) Imagem RGB; (b) Albedo estimado; e exemplos de imagens geradas (apresentadas em mapas de cores) para calcular o mapa de importância: (c) Mapa de saliência; (d) Magnitude do gradiente e (e) Mapa de importância	38
15	Mapas de características produzidos pela DENN. Nossa arquitetura con- siste em quatro camadas convolucionais seguidas por uma camada total- mente conectada por canal	38
16	(a) Imagem RGB; (b) Albedo prior; (c) Mapa de normais; (d) Mapa de direções de luz.	41
17	(a) Imagem RGB; (b) Mapa de profundidade; (c) Mapa de normais	41
18	(a) Imagem RGB; (b) Imagem sintética renderizada pelo modelo de reflexão de Phong.	43
19	(a) Imagem RGB; (b) Imagens sintéticas renderizadas por diferentes fontes de luz usando o modelo reflexão de Phong	44
20	(a) Imagem RGB; (b) Profundidade da imagem em visualização 3D; (c) Conjunto de direções de luz candidatas ao redor da profundidade da cena 3D	45
21	Algoritmo para calcular o mapa de direções de luz	45
22	(a) Imagem RGB; (b) Mapa de direções de luz.	46
23	Exemplos de pares de imagens do conjunto de dados NYU Depth-V2. Linha 1: imagens RGB; linha 2: profundidades verdadeiras	50
24	Perda média nas fases de treinamento (acima) e validação (abaixo) depois de 80 épocas.	53
25	Exemplos de resultados produzidos pela abordagem proposta por Yin et al. (2019) (profundidade preliminar) e pela nossa abordagem (profundidade produzida por DENN)	54
26	(a) Profundidade preliminar; (b) Profundidade produzida por DENN; (c) Di- ferença entre as profundidades.	55

27	Exemplos de mapas de profundidade melhorando durante o treinamento	68
28	Mais exemplos de mudanças na profundidades durante o treinamento que mostram a melhoria dos mapas de profundidade	69
29	Exemplos de mapas de normais melhorando durante o treinamento	70
30	Mais exemplos de mudanças nos mapas de normais durante o treinamento	
	que mostram a melhoria dos mapas de normais.	71

# Lista de Tabelas

1	Funções de ativação com suas equações e gráficos	23
2	Comparação entre os mapas de profundidade preliminares e os mapas de profundidade produzidos por DENN em métricas padrão. Fonte em negrito indica melhores resultados ou empate	56
3	Comparação entre os mapas de profundidade preliminares e os mapas de profundidade produzidos por DENN em métricas DBE. Fonte em negrito indica melhores resultados ou empate	56
4	Tempos médios de execução de DENN, avaliados com três técnicas diferen- tes de estimativa de profundidade monocular e uma técnica de estimativa de albedo	57
5	Comparação entre os mapas de profundidade preliminares e os mapas de profundidade produzidos por DENN em métricas padrão	58
6	Comparação entre os mapas de profundidade preliminares e os mapas de profundidade produzidos por DENN em métricas DBE	58

## Lista de Abreviaturas e Siglas

- **CNN** Convolutional Neural Network (Rede Neural Convolucional)
- **CRF** Conditional Random Field (Campo Aleatório Condicional)
- GAN Generative Adversarial Network (Rede Adversária Geradora)
- LSTM Long Short-Term Memory (Memória Prolongada de Curto Prazo)
- MRF Markov Random Field (Campo Aleatório de Markov)

# Sumário

1	Intro	odução	12
	1.1	Motivação e Ideia Central	13
	1.2	Objetivos	14
	1.3	Contribuições	14
	1.4	Estrutura da Dissertação	14
2	Fund	lamentação Teórica	16
	2.1	Imagens de Profundidade	16
	2.2	Estimativa de Profundidade baseada em uma Única Imagem	17
	2.3	Albedo	17
	2.4	Mapa de Importância	18
	2.5	Mapa de Saliência	18
	2.6	Magnitude do Gradiente	19
	2.7	Mapa de Normais	19
	2.8	Modelo de Reflexão de Phong	20
	2.9	Redes Neurais Convolucionais	20
		2.9.1 Operação de Convolução	20
	2.10	Camada de Dropout	21
	2.11	Funções de Ativação	22
	2.12	Função de Perda	22
	2.13	Métricas de Avaliação	22
		2.13.1 Métricas Padrão	23

		2.13.2	Métricas DBE	24
3	Trat	oalhos l	Relacionados	25
	3.1	Prime	iras pesquisas	25
	3.2	Métod	os baseados em redes neurais profundas	26
	3.3	Estade	o da arte	28
	3.4	Funçõ	es de perda	31
	3.5	Métod	os que tentam melhorar outras abordagens	32
4	Abo	rdagen	Proposta	33
	4.1	Abord	agens Prévias Necessárias	33
		4.1.1	Abordagens para Estimativa de Profundidade Monocular	34
		4.1.2	Abordagens para Estimativa de Albedo	35
	4.2	Etapa	de Treinamento	35
		4.2.1	Cálculo do Mapa de Importância	35
		4.2.2	Desenho da Arquitetura e Descrição do DENN	38
		4.2.3	Renderização da cena baseada no Modelo de Reflexão de Phong $\ .$ .	39
			4.2.3.1 Cálculo do Mapa de Normais	40
			4.2.3.2 Cálculo do Modelo de Reflexão do Phong	42
		4.2.4	Pré-processamento do Mapa de Direções da Luz	42
			4.2.4.1 Renderizações Geradas para cada Direção de Luz	42
			4.2.4.2 Seleção da Melhor Direção de Luz por Pixel	44
		4.2.5	Função de Perda	45
	4.3	Etapa	de Inferência	47
5	Exp	erimen	tos e Resultados	48
	5.1	Seleçã	o e Construção do Conjunto de Dados	48
		5.1.1	Levantamento de Conjuntos de Dados Disponíveis	48

			5.1.1.1	Conjuntos de Dados de Ambientes Internos	48
			5.1.1.2	Conjuntos de Dados de Ambientes Externos	49
		5.1.2	Conjunt	o de Dados Selecionado: NYU Depth-V2 Dataset	50
	5.2	Ferran	nentas Es	colhidas	50
	5.3	Metod	ologia Ex	perimental	51
		5.3.1	Seleção	de Abordagens de Estimativa de Profundidade Monocular	
			e Estima	ativa de Albedo para Treinamento	51
		5.3.2	Treinam	ento e Otimização dos Hiperparâmetros	51
		5.3.3	Teste		52
	5.4	Avalia	ção Quali	tativa	52
	5.5	Avalia	ção Quan	titativa	55
	5.6	Tempo	o de Exec	ução	57
	5.7	Discus	são		58
6	Con	clusões			59
	6.1	Trabal	lhos Futu	ros	59
Rł	REFERÊNCIAS				61
Ap	Apêndice A - MUDANÇAS NA PROF. DURANTE O TREINAMENTO			68	
Ap	Apêndice B - MUDANÇAS NAS NORMAIS DURANTE O TREINAMENTO       70				70

### 1 Introdução

Estimar o mapa de profundidade de uma cena a partir de uma única imagem RGB é um problema mal colocado e inerentemente ambíguo devido à distorção natural da projeção em perspectiva. É um dos problemas de visão computacional mais desafiadores, cuja solução poderia substituir o uso de dispositivos especiais como LiDAR e câmeras de profundidade baseadas em visão binocular, como a ZED Camera da Stereolabs Inc., ou triangulação passiva de padrões infravermelhos, como o Kinect Sensor da Microsoft e o Structure Sensor da Occipital Inc, mostrados na Figura 1. Além de serem mais caros que as câmeras convencionais, esses dispositivos também tendem a consumir mais energia.



Figura 1: Dispositivos especiais para obter mapas de profundidade

Apesar de ser um dos problemas mais desafiadores, a estimativa de profundidade de imagem única já foi resolvida (até certo ponto) por várias pesquisas e aplicada em robótica, veículos autônomos e outras áreas (Mo et al. (2020); Xue et al. (2020); Ye et al. (2020)). Infelizmente, em praticamente todas as pesquisas até hoje, ainda existem falhas na distinção de superfícies em uma cena. Essas deficiências podem ser vistas principalmente nas bordas dos objetos devido a interpolações erradas, valores incorretos ou ruídos. Essas bordas mal definidas dão a impressão de falta de superfícies discretas no mapa de profundidade. Idealmente, o mapa de profundidade produzido deve apresentar uma definição nítida das descontinuidades da superfície dos objetos. Outro problema comum encontrado nos resultados destas pesquisas é a irregularidade das superfícies planas. Alguns exemplos desses problemas são mostrados na Figura 2.



Figura 2: Detalhes imprecisos nas bordas e superfícies das imagens em profundidade. Profundidades verdadeiras (linha superior) e profundidades geradas (linha inferior) a partir de algumas abordagens (Bhat, Alhashim e Wonka (2021), Yin et al. (2019)).

#### 1.1 Motivação e Ideia Central

As deficiências acima mencionadas impedem o uso da maioria das abordagens de estimativa de profundidade de imagem única existentes para problemas práticos, onde a estimativa correta do limite dos objetos e a estimativa precisa da distância da câmera às superfícies são cruciais para realizar a manipulação de objetos ou evitar uma colisão. Neste trabalho, para superar essas deficiências, propomos uma nova rede neuronal profunda, Depth Enhanced Neural Network (DENN), que aprende a adicionar detalhes ausentes às superfícies representadas na imagem aproximada de profundidade obtida da saída de um modelo de estimativa de profundidade monocular existente usado como estimador preliminar de profundidade. Usamos explicitamente uma informação visual rica em detalhes que os modelos encontrados na literatura utilizam de forma indireta por não considerar comparações com a imagem colorida durante o treinamento. Isso porque as funções de perda utilizadas nesses modelos só olham a profundidade anotada.

Nossa rede neuronal profunda adiciona esses detalhes à imagem inicial da profundidade, corrigindo as normais das superfícies da cena obtidas dos mapas de profundidade. Essa correção das normais é baseada no uso do modelo de reflexão de Phong, onde em cada época do treinamento da rede, é gerada uma imagem renderizada que é comparada à imagem original, ambos no espaço de cor L\*a\*b\*, dando o valor da função de perda.

### 1.2 Objetivos

O principal objetivo deste trabalho é apresentar uma metodologia para a melhoria das imagens de profundidade estimadas por abordagens baseadas em uma única imagem. Um segundo objetivo é avaliar os resultados obtidos com a técnica proposta de forma qualitativa através da exibição de detalhes melhorados entre imagens comparadas e de forma quantitativa através de métricas de avaliação próprias para este tipo de problemas.

### 1.3 Contribuições

As contribuições deste trabalho são:

- Apresentar terminologia relacionada à questão da estimativa de profundidade de forma clássica e monocular.
- Revisão das primeiras pesquisas e do estado da arte dos trabalhos relacionados com a estimativa de profundidade monocular.
- Apresentar o DENN, um novo modelo para melhorar as imagens de profundidade produzidas por abordagens de estimativa de profundidade de imagem única.
- Apresentar um novo procedimento desenhado para calcular o valor de perda usado para treinar o modelo DENN.

#### 1.4 Estrutura da Dissertação

O conteúdo deste trabalho está organizado da seguinte forma:

- O Capítulo 2 apresenta a definição dos conceitos teóricos necessários para a compreensão da metodologia desenvolvida para a resolução do problema. Aqui descrevemos conceitos como estimativa de profundidade clássica e monocular, o albedo, o modelo de reflexão de Phong e conceitos relacionados a redes neurais convolucionais.
- O Capítulo 3 apresenta os diferentes trabalhos relacionados na área de estimativa de profundidade monocular, desde os trabalhos que iniciaram esta linha de pesquisa até aqueles do estado da arte.

- O Capítulo 4 apresenta a metodologia proposta para melhorar a estimativa de profundidade monocular. Descrevemos o fluxo de trabalho e cada um de suas etapas em detalhe, desde a aquisição do conjunto de dados até a etapa de inferência.
- O Capítulo 5 mostra os detalhes dos experimentos realizados com a proposta, as ferramentas e configurações utilizadas durante o treinamento e os testes da etapa de inferência. Também são apresentados os resultados obtidos, que são avaliados qualitativamente e quantitativamente.
- O Capítulo 6 apresenta as conclusões da proposta desenvolvida, algumas limitações encontradas e trabalhos futuros.
- Finalmente, são apresentadas as Referências e os Apêndices A e B que mostram o progresso da melhora dos mapas de profundidade e dos mapas de normais durante as épocas no treinamento.

### 2 Fundamentação Teórica

Neste capítulo, serão apresentados os conceitos fundamentais relacionados à profundidade, o albedo extraído de uma imagem, a função de um mapa de importância, o modelo de iluminação utilizado, os conceitos de redes neurais profundas e as métricas de avaliação utilizadas neste tipo de problemas. Todos esses conteúdos serão necessários para a compreensão da proposta deste trabalho.

#### 2.1 Imagens de Profundidade

Uma imagem de profundidade é uma representação bidimensional 2D de uma cena tridimensional 3D, onde cada pixel que compõe a imagem 2D fornece informações sobre a distância de um ponto em uma superfície visível da cena 3D a partir de um ponto de vista. Os pixels da imagem 2D são expressos em uma faixa de valores, onde os valores extremos representam os pontos mais próximos ou mais distantes da cena dependendo da interpretação do autor.

Normalmente, uma forma de exibir essa representação 2D é colocar o intervalo de valores da imagem em um intervalo normalizado que pode resultar em uma imagem em tons de cinza (com valores de 0 a 255). Outra maneira de exibir a representação 2D da profundidade é por meio de um mapa de cores. Na Figura 3 são mostradas as representações em escala de cinza e em mapas de cores de uma imagem de profundidade.



Figura 3: (a) Imagem original. Representações 2D da profundidade: (b) em tons de cinza; (c) em mapa de cor Jet; (d) em mapa de cor Magma.

### 2.2 Estimativa de Profundidade baseada em uma Única Imagem

Nos últimos anos e com o uso generalizado de redes neurais e abordagens baseadas em aprendizado profundo, a estimativa de um mapa de profundidade a partir de uma única imagem (problema conhecido como "estimativa de profundidade monocular") tornou-se um problema desafiador que recebeu muita atenção em visão computacional. Este problema, aparentemente mal imposto devido à falta de informação fornecida por uma única imagem, está ganhando popularidade atualmente, pois pode levar a aplicações que utilizam menos recursos computacionais e substituir dispositivos ou sensores que fornecem a profundidade de forma tradicional. As Seções 3.1, 3.2 e 3.3 descrevem este tópico com mais detalhes.

#### 2.3 Albedo

Também conhecido como fator de refletância de superfície, em termos físicos, é a razão entre a luz incidente em uma superfície e a luz refletida. As superfícies claras e brilhantes têm um albedo maior do que as superfícies escuras e foscas. Em computação gráfica, o albedo pode ser interpretado como a cor característica de um objeto e faz parte das chamadas imagens intrínsecas junto com o sombreamento que compõem uma imagem.

Na literatura é possível encontrar diferentes técnicas que permitem estimar o albedo de uma imagem (Zhou, Krahenbuhl e Efros (2015), Nestmeyer e Gehler (2017), Lettry, Vanhoey e Van Gool (2018), Yunfei Liu et al. (2020), dentre outros.). Na Figura 4, podemos ver um exemplo de uma imagem colorida e seu respectivo albedo gerado pela técnica de Lettry, Vanhoey e Van Gool (2018).



(a)



(b)

Figura 4: (a) Imagem RGB; (b) Albedo.

#### 2.4 Mapa de Importância

Devido a que uma imagem contém informações relevantes que permitem sua melhor interpretação e melhor reconhecimento e dependendo da informação que mais precisamos destacar, um mapa de importância permite destacar aquelas características que tornam a imagem o mais interpretável e reconhecível possível de acordo com uma necessidade e minimiza ou elimina informações sem importância. Assim, as regiões que precisamos destacar mais na imagem podem ser claramente identificadas.

Dependendo do tipo de regiões que deseja destacar em uma imagem, existem diferentes tipos de técnicas que permitem a identificação dessas regiões. A Figura 5 mostra um exemplo de mapas de importância obtidos de uma imagem com a técnica de Choi et al. (2021).



Figura 5: Mapas de importância gerados a partir de uma imagem RGB. Fonte: Choi et al. (2021).

#### 2.5 Mapa de Saliência

O mapa de saliência é um mapa que destaca as regiões onde o sistema de visão se concentra primeiro ou com mais destaque. Pode ser visto como um mapa de calor onde cada pixel na imagem reflete um grau de importância. Esse tipo de mapa é amplamente utilizado em técnicas de aprendizado profundo, como nas CNNs, onde ajudam à rede a se concentrar em determinadas áreas de pixels. Ao contrário do mapa de importância, o mapa de saliência pode destacar muitas regiões onde existem características que talvez não queiramos preservar.

#### 2.6 Magnitude do Gradiente

A magnitude do gradiente é uma quantidade escalar que indica a rapidez com que a intensidade da imagem muda de acordo com a vizinhança de cada pixel (x, y) em uma determinada direção do gradiente.

#### 2.7 Mapa de Normais

O mapa de normais é uma textura de imagem que está associada às superfícies presentes em uma cena. Cada ponto dessa textura corresponde diretamente aos eixos X, Y e Z no espaço 3D e fornece informações sobre a direção exata na qual a normal desse ponto está orientada. É comum expressar a orientação de cada normal no mapa de normais usando informações RGB.

A Figura 6 mostra exemplos geométricos com a paleta de cores usada para representar as orientações das normais nas superfícies em um mapa de normais. A Figura 7 mostra um exemplo do mapa de normais de uma cena real.



Figura 6: Superfícies mostrando uma paleta de cores usada para representar mapas de normais. Fonte: Autoria própria.



(a)



Figura 7: (a) Imagem RGB; (b) Mapa de normais.

#### 2.8 Modelo de Reflexão de Phong

O modelo de reflexão de Phong é um modelo de iluminação apresentado por Phong (1975) e que atribui um brilho aos pontos de uma superfície modelada. Este modelo descreve a forma como a superfície de um objeto reflete a luz, como uma combinação de três componentes: um componente de reflexão ambiente correspondente à luz espalhada pela cena, um componente de reflexão difusa característica de superfícies rugosas e um componente de reflexão especular característico de superfícies lustrosas. A Figura 8 mostra uma ilustração visual dos três componentes do modelo de reflexão Phong e do modelo resultante.



Figura 8: Visualização dos componentes do modelo de Phong e do modelo resultante. Fonte: Brad Smith, Wikimedia Commons.

Para calcular a iluminação em cada ponto da superfície usando o modelo de Phong, existe uma equação descrita com detalhe na Seção 4.2.3.

#### 2.9 Redes Neurais Convolucionais

Uma rede neural convolucional é um tipo de rede neural artificial composta por blocos de múltiplas camadas (e.g., camadas convolucionais, camadas de pooling e camadas totalmente conectadas, dentre outros) que, através da operação de convolução aplicada às camadas convolucionais usando kernels ou filtros de convolução, permitem extrair mapas de características dos dados de entrada que podem ser imagens, áudio, séries temporais, sinais, etc. A Figura 9 mostra uma representação de uma rede neural convolucional.

#### 2.9.1 Operação de Convolução

Esta operação consiste em utilizar um kernel ou filtro que percorre toda a imagem para obter um mapa de características. Primeiro, o tamanho da janela do filtro no canto



Figura 9: Exemplo de uma arquitetura de rede neural convolucional. Fonte: Tabian, Fu e Sharif Khodaei (2019).

superior esquerdo é definido. Depois, a janela do filtro se move progressivamente da esquerda para a direita um número pré-definido de caixas (hiperparâmetro Stride) até chegar ao final da imagem. Em cada porção da imagem que encontra, é realizado o cálculo de convolução, permitindo obter na saída um mapa de características que indica onde estão localizadas as características desejadas na imagem. A Figura 10 mostra um exemplo de uma operação de convolução aplicada aos pixels de uma imagem.



Figura 10: Exemplo de operação de convolução sobre os pixels uma imagem. Fonte: Autoria própria.

### 2.10 Camada de Dropout

Permite realizar a técnica de dropout, que é um método de regularização que a cada iteração do processo de treinamento da rede, desativa diferentes neurônios aleatoriamente a fim de reduzir o sobreajuste do modelo. Esta técnica possui um parâmetro que indica a probabilidade de os neurônios serem ativados, sendo que este parâmetro pode ser ajustado para um valor de 0 a 1.

#### 2.11 Funções de Ativação

São funções usadas em redes neurais que calculam uma soma ponderada das entradas e viéses para determinar se um neurônio será ativado ou não. A escolha de uma função de ativação correta será refletida nos resultados obtidos pela rede neural. Algumas das funções de ativação mais conhecidas são:

- Linear: Dá um valor de saída igual ao valor de entrada.
- **ReLU**: Dá um valor de saída de 0 se o valor de entrada for negativo, caso contrário, deixa o valor de saída igual ao valor de entrada.
- Sigmoid: Dá um valor de saída no intervalo de [0, 1], onde os valores de entrada mais altos tendem a 1 e os valores mais baixos tendem a 0.
- TanH: Dá um valor de saída no intervalo [-1, 1], onde os valores de entrada mais altos tendem a 1 e valores mais baixos tendem a -1.
- SoftSign: Dá um valor de saída no intervalo de [-1, 1]. É uma alternativa à função TanH que converge exponencialmente, enquanto a função SoftSign converge polinomialmente.

A Tabela 1 mostra as funções de ativação descritas acima, com suas respectivas equações e gráficos.

#### 2.12 Função de Perda

É uma função que compara a saída obtida da rede neural com o alvo para medir quão bem a rede está sendo treinada. O objetivo do treinamento é minimizar o valor da perda entre os dados alvos e os dados resultantes da rede. Existe uma grande variedade de funções de perda, como o Erro Quadrado Médio ou a Entropia Categórica Cruzada, e dependendo do tipo de problema que deseja resolver, se pode criar uma função de perda mais adequada.

#### 2.13 Métricas de Avaliação

Para medir o desempenho dos mapas de profundidade produzidos por nossa abordagem em comparação com os mapas de profundidade gerados por técnicas preliminares de estimativa de profundidade monocular, usamos dois tipos de métricas de avaliação: um

Função de ativação	Equação	Gráfico
Lineal	f(x) = x	122 33 33 34 -23 -23 -23 -23 -23 -23 -23 -23
Del II	f(x) = max(0, x)	
ReLU	J(x) = max(0, x)	-10.0 -7.5 -50 -2.5 00 2.3 50 7.3 10.0
Sigmoid	$f(x) = \frac{1}{1 + e^{-x}}$	
TanH	$f(x) = \frac{2}{3}$ 1	100 1075 109 103 103 100 103 100 100 100 100
	$J(x) = \frac{1}{1+e^{-2x}} - 1$	-10.0 -7.5 -5.0 -2.3 0.0 2.3 5.0 7.3 10.0
SoftSign	$f(x) = \frac{x}{(1+ x )}$	237 237 232 009 -228 -339 -319 -106 -75 -5.0 -2.5 0.0 2.5 50 7.5 16.0

Tabela 1: Funções de ativação com suas equações e gráficos.

conjunto de métricas padrão que comparam mapas de profundidade com a profundidade verdadeira e um conjunto de métricas relacionadas aos erros das bordas das profundidades.

#### 2.13.1 Métricas Padrão

Para medir a eficiência das abordagens de estimativa de profundidade monocular, Eigen, Puhrsch e Fergus (2014) propuseram um conjunto de métricas padrão: diferença relativa absoluta (Abs-Rel, Equação 2.1), erro quadrado relativo (Sq-Rel, Equação 2.2), erro quadrático médio (RMSE, Equação 2.3), RMSE-Log (Equação 2.4) e precisão com um limiar  $(\delta_t, Equação 2.5)$ ; onde para cada equação,  $\hat{d}_i \in d_i$  denotam a profundidade estimada e a profundidade verdadeira do pixel i, respectivamente, e N é o número do pixels no mapa de profundidade:

Abs-Rel = 
$$\frac{1}{N} \sum_{i} \left| \hat{d}_{i} - d_{i} \right| / d_{i}$$
 (2.1)

$$\operatorname{Sq-Rel} = \frac{1}{N} \sum_{i} \left\| \hat{d}_{i} - d_{i} \right\|^{2} / d_{i}$$

$$(2.2)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i} \left\| \hat{d}_{i} - d_{i} \right\|^{2}}$$
(2.3)

RMSE-Log = 
$$\sqrt{\frac{1}{N}\sum_{i} \left\| \log(\hat{d}_{i}) - \log(d_{i}) \right\|^{2}}$$
 (2.4)

$$\delta_t = \text{Percentagem de } d_i \text{ tal que } max\left(\frac{gt}{pred}, \frac{pred}{gt}\right) < 1.25^t$$
 (2.5)

#### 2.13.2 Métricas DBE

Outro conjunto de métricas, foi proposta por Koch et al. (2018) para medir a eficiência das abordagens baseadas na qualidade das bordas dos objetos na cena. Elas incluem as duas métricas de erro de borda de profundidade (do inglês Depth Boundary Errors (DBE)): o erro de precisão ( $\varepsilon_{\text{DBE}}^{acc}$ , Equação 2.6) e o erro de completude ( $\varepsilon_{\text{DBE}}^{comp}$ , Equação 2.7); onde  $y_{bin}$  e  $y_{bin}^*$  representam os contornos dos mapas de profundidade extraídos com a técnica de bordas estruturadas para a profundidade predita e a profundidade verdadeira, respectivamente,  $e \in e^*$  representam as transformadas de distância euclidiana para a profundidade predita e a profundidade verdadeira:

$$\varepsilon_{\text{DBE}}^{acc}(Y) = \frac{1}{\sum_{i} \sum_{j} y_{bin;i,j}} \sum_{i} \sum_{j} e_{i,j}^* \cdot y_{bin;i,j}$$
(2.6)

$$\varepsilon_{\text{DBE}}^{comp}(Y) = \frac{1}{\sum_{i} \sum_{j} y_{bin;i,j}^*} \sum_{i} \sum_{j} e_{i,j} \cdot y_{bin;i,j}^*$$
(2.7)

### **3** Trabalhos Relacionados

Neste capítulo, serão apresentadas pesquisas relacionadas à estimativa de profundidade monocular e alguns estudos que tentam melhorar tais profundidades estimadas. Descreveremos a evolução ao longo do tempo das diferentes técnicas que foram desenvolvidas para resolver este problema e, além disso, apresentaremos uma classificação das técnicas de acordo com suas funções de perda utilizadas.

#### 3.1 Primeiras pesquisas

As pesquisas iniciais sobre o tema apontam para meados dos anos 2000 com abordagens baseados em Markov Random Field (Campo Aleatório de Markov) (MRF), como no trabalho de Saxena, Sung Chung e Andrew Ng (2005), que usa características locais e globais em múltiplas escalas da imagem para obter as profundidades em pontos individuais e a relação entre as profundidades de vários pontos. Já no trabalho de Saxena, Schulte, Ng et al. (2007), os autores usam o MRF para capturar sinais monoculares tais como variações de textura e gradientes, cor, opacidade, etc. e incorporá-los em um sistema estereoscópico, enquanto que Saxena, Sung H Chung e Andrew Y Ng (2008) combinam as pesquisas acima mencionadas para obter melhores resultados. Outro trabalho deste tipo é o de Liu, Gould e Koller (2010), onde a classe semântica das peças que compõem a cena é inferida através de MRF. Outro trabalho relevante é apresentado por Wedel et al. (2006), que estima a profundidade da cena através do dimensionamento de regiões em quadros de vídeo e o aplica em ambientes externos para detectar obstáculos estacionários.

Nesses primeiros estudos já era possível ver imagens com estilos que representavam profundidade, onde os objetos mais próximos tinham uma coloração diferente em relação aos objetos mais distantes.

#### 3.2 Métodos baseados em redes neurais profundas

Em anos posteriores e graças ao grande salto que ocorreu em várias áreas da visão computacional devido ao hardware disponível que possibilitou novas pesquisas com o uso das redes neurais profundas, inúmeras pesquisas resultaram em imagens de profundidade estimadas que revelam mais detalhes dos objetos nas cenas. Assim, trabalhos baseados em redes neurais, Convolutional Neural Network (Rede Neural Convolucional) (CNN), sistemas híbridos e outros foram introduzidos.

Por exemplo, no trabalho de Eigen, Puhrsch e Fergus (2014) foram usadas duas redes profundas: uma que faz a predição global na imagem e outra que a refina localmente. Sua abordagem foi testada em dois conjuntos de dados que são bem conhecidos na comunidade e que têm sido muito requisitados desde aqueles anos: NYU (Silberman et al. (2012)) e KITTI (Geiger et al. (2013)). O trabalho de Liu, Shen e Lin (2015) propõe o uso de um sistema híbrido composto de uma CNN e um Conditional Random Field (Campo Aleatório Condicional) (CRF) Contínuo, pois considera os valores de profundidade como uma característica contínua e, por conseguinte, o problema de estimativa de profundidade passa a ser como um aprendizado CRF contínuo. Fayao Liu et al. (2015), em um trabalho posterior, melhoram a pesquisa citada antes adicionando um método de agrupamento para gerar superpixels, que é usado para a CNN, melhorando a velocidade de sua abordagem. O trabalho de Zhang et al. (2015) aborda o problema da segmentação de objetos e a ordenação na profundidade destes em imagens monoculares. Para isso, eles utilizam uma CNN que prevê as segmentações das instâncias e depois um MRF que infere a ordem de profundidade dos objetos segmentados. No trabalho de Chakrabarti, Shao e Shakhnarovich (2016), uma CNN é utilizada para obter características da estrutura geométrica local, através da predição das derivadas da profundidade em diferentes ordens, orientações e escalas em cada localização de imagem.

No trabalho de Kuznietsov, Stuckler e Leibe (2017), os autores propõem uma abordagem semi-supervisionada usando uma CNN que é ajudada por alguns sinais obtidos de um LiDAR (imagens estéreo e a profundidade verdadeira dispersa) para gerar mapas de profundidade densos fotoconsistentes. A pesquisa de Mancini et al. (2017) propõe uma rede neural profunda, treinada em um conjunto de dados sintéticos. Eles também mostram como o uso de uma rede Long Short-Term Memory (Memória Prolongada de Curto Prazo) (LSTM) ajuda a aliviar algumas das limitações neste tipo de problemas, por exemplo, reduzindo a sobrecarga computacional. O trabalho de Godard, Mac Aodha e Brostow (2017) apresenta uma abordagem que trata a estimativa de profundidade mo-

27

nocular como um problema de reconstrução da imagem. Eles usam uma CNN que utiliza imagens estéreo retificadas e, a fim de melhorar sua qualidade, propõe uma nova função de perda entre as disparidades produzidas pelas imagens estéreo.

O trabalho de Harsányi et al. (2018) apresenta uma rede neural profunda híbrida que combina uma CNN existente de outro trabalho e as etapas de codificação e decodificação de uma U-net. Fu et al. (2018), em seu trabalho, aplica um método de discretização crescente do espaçamento nas imagens que compõem a cena e com os valores de profundidade discreta treina uma CNN usando uma perda de regressão ordinal. O trabalho de Duan et al. (2018) apresenta uma rede baseada na arquitetura ResNet que estima a profundidade monocular. Depois, para melhorar este resultado, eles aplicam uma sub-rede de melhoria baseada em uma CNN. Alhashim e Wonka (2018) apresentam uma CNN baseada no aprendizado de transferência onde a rede é dividida em um codificador e um decodificador. O codificador é uma Dense-Net 169 pré-treinada como um classificador e para a parte do decodificador são utilizados blocos de camadas convolucionais para gerar a imagem de profundidade. No trabalho de He, Wang e Hu (2018), eles incorporam a distância focal em sua rede, mas primeiro geram um conjunto de dados de distância focal variável a partir de conjuntos de dados conhecidos que têm distância focal fixa. Depois, com as imagens de entrada correspondentes, eles treinam a rede proposta com base em um modelo VGG pré-treinado e adicionam as informações obtidas a partir das distâncias focais.

No trabalho de Atapour-Abarghouei e Breckon (2018), é apresentada uma abordagem que utiliza uma Generative Adversarial Network (Rede Adversária Geradora) (GAN) para aprender sobre um conjunto de dados sintéticos RGB-D que, através da transferência de estilo e adaptação ao domínio de dados reais, resulta em imagens de profundidade de domínio do mundo real. No trabalho de Xian et al. (2018), os autores utilizam um conjunto de dados de imagens estéreo chamado Depth in the Wild (DIW) para gerar um novo conjunto de dados de mapas de disparidade (Relative Depth from Web (ReDWeb) 2) utilizando um método de fluxo óptico. Depois, utilizando uma CNN, eles calculam a profundidade com base nos mapas do conjunto de dados encontrados anteriormente. O trabalho de Jiao et al. (2018) apresenta uma abordagem onde eles resolveram o problema da estimativa de profundidade monocular e segmentação semântica dado que uma vez que são conhecidas algumas informações semânticas dos objetos, a compreensão da profundidade da cena melhora. Eles empregam uma CNN multitarefa composta de uma sub-rede de predição de profundidade e uma sub-rede de rotulagem semântica que fazem o compartilhamento do conhecimento através de unidades de troca lateral. No trabalho de Yaoxin Li et al. (2018), eles utilizam uma GAN condicional que aprende a ajustar a profundidade da imagem gerada por 3 entradas dadas no gerador: imagens RGB, nuvens de pontos e ruído aleatório.

#### 3.3 Estado da arte

Com o avanço nas pesquisas que utilizaram aprendizado profundo, o problema de estimar uma imagem de profundidade monocular tornou-se popular nos últimos anos, onde muitos trabalhos usavam mais sistemas híbridos ou predição de profundidade combinada com outros problemas.

Por exemplo, no trabalho de Man et al. (2019), sua abordagem estimou a orientação do plano do solo a partir de uma única imagem. Para isso eles desenharam uma rede que tinha uma sub-rede que estimava aproximadamente a profundidade e outra sub-rede que estimava a superfície normal. Mern et al. (2019) apresentam uma abordagem que utiliza uma R-CNN e um auto-encoder baseado em unidades recorrentes fechadas convolucionais (C-GRUs) treinadas em sequências de vídeo obtidas do simulador AirSim da Microsoft. O trabalho de Casser et al. (2019) apresenta uma abordagem que introduz a estrutura geométrica de cenas em movimento no aprendizado para estimativa de profundidade. Isso é feito usando uma rede de estimadores de ego-movimento baseada em uma CNN que permite determinar a profundidade inicial. Depois, para refinar isso, eles usam um estimador de movimento de objetos 3D na cena. Hu et al. (2019) apresentam uma abordagem que extrai características da imagem em várias escalas e as mescla. Desta forma eles desenharam uma rede que é dividida em quatro partes: um codificador que extrai as características em múltiplas escalas, um decodificador usado para decodificar as características até obter uma de escala 1/2, um módulo de fusão multiescala e um módulo de refinamento.

Lee, Han et al. (2019) apresentam seu trabalho baseada em CNN contendo um decodificador onde uma camada de orientação planar local é colocada em cada camada de decodificação, depois as saídas de todas as camadas são combinadas para obter a imagem de profundidade com resolução total. No trabalho de Lee e Kim (2019), eles usam uma CNN que estima a profundidade relativa e global entre pares de regiões mas em diferentes escalas. Depois, esses mapas são recombinados para obter a profundidade final. Tosi et al. (2019) apresentam uma abordagem onde eles estimam a profundidade monocular e que usa a disparidade gerada por imagens estéreo para refinar a profundidade final. O trabalho de Yin et al. (2019) mostra a importância das restrições geométricas 3D para a estimativa de profundidade, assim, eles obtém um mapa de normais virtuais preditos a partir da profundidade verdadeira que gera uma nuvem de pontos 3D que auxilia na melhoria do mapa de profundidade anteriormente encontrado por meio de uma rede codificador-decodificador.

Tan et al. (2019) usa uma GAN cujo gerador é uma rede codificador-decodificador com blocos de convolução transpostos que gera uma imagem de profundidade que é depois comparada com a profundidade verdedira no discriminador. O trabalho de Lin et al. (2019) apresenta uma abordagem que tenta resolver dois problemas ao mesmo tempo: estimativa de profundidade e segmentação semântica. Para fazer isso, eles avaliam quais características podem ser compartilhadas para resolver ambos os problemas e quais características devem ser separadas. Logo, para a parte de estimativa de profundidade, eles desenharam uma rede baseada na arquitetura AlexNet dividida em 3 partes: uma que estima a profundidade em nível global, uma que prediz os mapas de gradiente da profundidade estimada e outra que refina a profundidade. No trabalho de Wofk et al. (2019), eles desenvolvem uma abordagem de estimativa de profundidade para sistemas embarcados que funcionam em tempo real. Assim, eles usam uma rede codificador-decodificador onde a poda de rede é aplicada no lado do decodificador para reduzir a complexidade e a latência computacional.

O trabalho de Klingner et al. (2020) apresenta uma rede híbrida que resolve o problema de segmentação semântica para orientar a estimativa de profundidade. Eles desenvolvem isso em amostras de vídeo onde, por meio da segmentação prévia, distinguem objetos estáticos e dinâmicos na cena para refinar a profundidade. Wang et al. (2020) propõe uma abordagem que resolve o problema de estimativa de profundidade com base na abordagem de dividir e conquistar visto sob segmentação semântica. Seu modelo decompõe a cena em blocos semânticos e então prediz a profundidade para cada segmento semântico, de modo que aqueles na mesma categoria semântica compartilhem o mesmo decodificador de profundidade. No final, os segmentos de profundidade local são unidos fazendo a predição de sua escala e deslocamento no nível global da imagem. No trabalho de Guizilini et al. (2020), eles propõem um método que estima a profundidade de vídeos monoculares usando uma arquitetura codificadora-decodificadora baseada em uma rede que introduz blocos de empacotamento e desempacotamento 3D que aprendem a preservar informações importantes para obter profundidade em tempo real. Cao et al. (2020) apresentam uma abordagem que utiliza um conjunto de dados próprio (Relative Depth in Stereo (RDIS)) gerado a partir de filmes estéreos para obter profundidades relativas. Em seguida, eles

pré-treinam um modelo ResNet com este conjunto de dados. Posteriormente, para o processo de ajuste fino, eles formulam a estimativa de profundidade como um problema de classificação em um conjunto de dados RGB-D.

O trabalho de Bhat, Alhashim e Wonka (2021) destaca como o processamento global das imagens pode ajudar a estimar melhor a profundidade. Por esta razão, os autores desenharam uma arquitetura que consiste em duas partes: a primeira, que é um bloco codificador-decodificador clássico, e a segunda, que refina a saída do decodificador com um bloco de pós-processamento composto por um codificador de transformadores. Gurram et al. (2021) apresenta uma abordagem híbrida com uma arquitetura de duas partes: uma parte aprende de forma supervisionada enquanto a outra aprende de forma autosupervisionada. A parte supervisionada usa um conjunto de dados do mundo virtual com imagens RGB e profundidade de verdade que é treinada primeiro e depois dá lugar à parte autosupervisionada que usa os princípios da estrutura do movimento (SfM) graças a uma câmera montada em um carro no mundo real. No trabalho de Song, Lim e Kim (2021), eles apresentam uma rede baseada na arquitetura codificador-decodificador onde as pirâmides Laplacianas são incorporadas na etapa de decodificação, pois elas preservam as informações locais dos dados gerenciados. Ranftl, Bochkovskiy e Koltun (2021) apresenta uma abordagem baseada no uso de transformadores de visão ao invés de redes convolucionais como na maioria dos trabalhos. Para isso, eles primeiro dividem a imagem em tokens através do ResNet-50, depois esses tokens são passados por transformadores que realizam a função de um codificador e, finalmente, as predições obtidas são combinadas usando um decodificador. No trabalho de Queiroz Mendes et al. (2021), eles propõem uma FCN leve e rápida projetada para navegação autônoma no mundo real que se concentra em encontrar sinais geométricas da imagem de entrada para a predição da imagem de profundidade. Eles também adicionam outros métodos à sua arquitetura, como: terminação de profundidade, segmentação semântica, estimativa das normais da superfície, entre outros para melhorar a profundidade final.

No trabalho de Masoumian et al. (2021), eles apresentam uma rede composta por duas partes: a primeira é uma ResNet que extrai características da imagem e a segunda uma série de Graph Convolutional Networks (GCN) para a estimativa de profundidade ao invés de CNNs clássicas, já que as CNNs não consideram a geometria da profundidade. Chang, Zhang e Xiong (2021) apresenta uma rede híbrida baseada em transformadores de visão para a parte de codificação e um decodificador baseado em uma CNN. Para aproveitar melhor os transformadores, eles desenharam uma supervisão de atenção que serve de guia para os transformadores. A supervisão de atenção é obtida calculando o mapa de atenção de cada pixel na imagem com base na profundidade verdadeira. Shuai Li et al. (2021) apresenta uma abordagem que considera a relação hierárquica entre os objetos que compõem a cena como chave principal. É por isso que desenharam uma rede que consiste em três partes. A primeira parte é um módulo para a representação de relacionamentos hierárquicos, que por meio da semântica extrai as relações entre objetos entre si e entre seus vizinhos. A segunda parte é um módulo para extrair características e relações. A terceira parte é o módulo de restrição semântica da profundidade predita e que refina a profundidade global. No trabalho de Chawla et al. (2021), eles propõem uma abordagem que utiliza dados de sistemas de posicionamento global (GPS) como entrada para a rede, pois esses dados fornecem a distância entre vários quadros capturados do vídeo. Os dados fornecidos pelo GPS são utilizados apenas no treinamento da rede. Li, Luo e Xiao (2022) propõe uma abordagem que estima a profundidade primeiro grosseiramente usando um módulo de atenção de bloco convolucional normalizado (NCBAM) e depois o refina usando uma rede que usa a imagem colorida original para resolver o problema de completude da profundidade.

#### 3.4 Funções de perda

Várias funções de perda têm sido aplicadas no treinamento de modelos de estimativa de profundidade monocular. Por exemplo, no trabalho de Eigen, Puhrsch e Fergus (2014) aplicaram o Root Mean Square Error (RMSE). Eles foram seguidos por Bhat, Alhashim e Wonka (2021), que também aplicaram uma perda de densidade bin-center que usa a distância de Chamfer. A perda de entropia cruzada foi aplicada por Zhang et al. (2015), Wofk et al. (2019), Yin et al. (2019) e pela GAN proposta por Tan et al. (2019). Lin et al. (2019) usou a soma da entropia cruzada e dois termos relacionados à distância euclidiana entre a profundidade verdadeira e as profundidades estimadas.

A aplicação das perdas L1 e L2 é bastante popular. Por exemplo, He, Wang e Hu (2018) utilizaram a função de perda BerHu, que consiste nas normas das perdas L1 e L2. Casser et al. (2019) considerou a soma da perda L1 e o Índice de Similaridade Estrutural (SSIM). Tosi et al. (2019) usou uma perda que é a soma dos termos baseados na perda de BerHu, SSIM e disparidade de gradiente. Klingner et al. (2020) aplicou uma perda mínima de reprojeção (baseada em SSIM) e uma perda de suavidade, enquanto Atapour-Abarghouei e Breckon (2018) aplicaram a perda L1 para treinar o gerador de sua GAN. Duan et al. (2018) utilizou a soma das perdas L2 adaptadas para avaliar imagens de baixa e alta resolução produzidas por seu modelo. Hu et al. (2019) resumiu a perda de
profundidade L1 à perda de gradientes de profundidade e normais de superfície. Man et al. (2019) também somou três termos para compor sua função de perda: a profundidade e as perdas normais baseadas na perda L2 e um termo de consistência geométrica para perdas anteriores. Song, Lim e Kim (2021) usaram uma função composta por um termo de perda de dados baseado em uma perda de L2 e um termo de perda de gradiente baseado em somatórios de perdas L1. O uso mais envolvido da perda de L1 foi apresentado por Long Chen et al. (2020). Eles usaram uma perda composta por uma Perda de Correspondência de Patch (inspirada na perda de correlação cruzada normalizada de média zero), Perda de Reconstrução de Visualização (com base em uma perda de L1), Perda de Suavidade de Disparidade (com base na disparidade de gradiente de norma L1) e Perda de Consistência de Disparidade (com base em uma perda L1).

Outras estratégias para definir a função de perda incluem o uso de CRFs (Liu, Shen e Lin (2015)), classificação (Xian et al. (2018), Cao et al. (2020)), divergência KL (Chakrabarti, Shao e Shakhnarovich (2016)), regressão ordinal em valores discretos de profundidade (Fu et al. (2018), Lee e Kim (2019) e atenção (Jiao et al. (2018)).

Ao contrário de todas as funções de perda discutidas aqui, nossa função de perda proposta usa a comparação perceptual das cores na imagem fornecida como entrada para o estimador de mapa de profundidade preliminar com uma imagem sintética renderizada para a cena usando o mapa produzido por nossa abordagem. O detalhe de nossa função de perda é descrita na Seção 4.2.5.

### 3.5 Métodos que tentam melhorar outras abordagens

Existem também alguns trabalhos que tentam melhorar os resultados obtidos por as abordagens de estimativa de profundida monocular. Por exemplo, Tian Chen et al. (2020) usa uma rede neural com blocos de atenção espacial que focam em áreas onde as imagens podem ser melhoradas (em bordas ou descontinuidades) ou como no trabalho de Parida, Srivastava e Sharma (2021), que usa uma rede neural que leva como entradas uma série de ecos binaurais, imagens RGB e propriedades de material estimadas de vários objetos na cena para corrigir as imagens em profundidade.

A principal diferença entre nossa abordagem e essas técnicas que também melhoram os mapas de profundidade é que nossa técnica usa uma metodologia de treinamento não supervisionado e precisa apenas de imagens RGB e não de outros sinais ou dados mais difíceis de obter ou processar.

## 4 Abordagem Proposta

Nossa abordagem visa melhorar o mapa de profundidade estimado a partir de uma única imagem RGB por uma técnica de estimativa de profundidade existente. O modelo proposto (ou seja, o DENN) realiza essa tarefa adicionando detalhes ao mapa de profundidade fornecido, corrigindo a nitidez das arestas da superfície e aplanando regiões planas. Fazer essas melhorias é um processo aprendido por DENN durante o treinamento, onde a função de perda proposta avalia se a imagem colorida original é semelhante a uma renderização sintética da cena. Assumindo uma versão simplificada do modelo de reflexão de Phong ao renderizar a cena e fixando o albedo e as direções das fontes de luz, a correção de renderização resulta da correção das normais de superfície e, portanto, do mapa de profundidade.

Dividimos nossa abordagem em duas etapas. A primeira etapa é o treinamento, a totalidade da Figura 11 mostra o fluxo de processamento realizado por nossa abordagem durante esta etapa. Dentro do treinamento temos uma etapa anterior de pré-processamento de direções de luz que é mostrado no retângulo cinza pontilhado na mesma figura. A segunda etapa é a inferência, que é destacada pelo retângulo cinza tracejado na parte superior da Figura 11.

Para mais detalhes sobre cada etapa e cada parte da proposta, dividimos o capítulo da seguinte forma: a Seção 4.1 apresenta as técnicas que usamos como modelos preliminares de estimativa de profundidade e estimativa de albedo; a Seção 4.2 apresenta as partes que compõem o treinamento; a Seção 4.3 descreve os procedimentos utilizados para a etapa de inferência da abordagem.

## 4.1 Abordagens Prévias Necessárias

Nossa abordagem precisa de duas entradas iniciais baseadas na imagem original: seu mapa de profundidade predito a partir de uma técnica preliminar que estima a profundidade monocular e sua imagem de albedo obtida de uma técnica que extrai o albedo.



Figura 11: O diagrama de fluxo de treinamento do DENN. Os retângulos pontilhados e tracejados denotam o estágio de pré-processamento das direções da luz e a inferência do modelo, respectivamente.

#### 4.1.1 Abordagens para Estimativa de Profundidade Monocular

A entrada primária do DENN é o mapa de profundidade aproximado previsto por alguma técnica de estimativa de imagem de profundidade. Conforme descrito no Capítulo 3, existem várias técnicas de estimativa de profundidade na literatura que podem ser usadas

para gerar o mapa de profundidade inicial exigido por nossa abordagem.

No diagrama de fluxo da Figura 11, o processo que estima a profundidade monocular é representado pelo retângulo denominado "Modelo Preliminar de Estimativa da Profundidade". A Figura 12 mostra alguns resultados (linha 2 a linha 7) das técnicas utilizadas para gerar o mapa de profundidade da imagem original (linha 1). A linha 2 é o resultado da técnica de Yin et al. (2019), a linha 4 é o resultado da técnica de Bhat, Alhashim e Wonka (2021) e a linha 6 é o resultado da técnica de Song, Lim e Kim (2021), todos esses mapas estão em representação em escala de cinza. As linhas 3, 5 e 7 são os mapas de profundidade das mesmas técnicas mencionadas acima, mas em mapas de cores.

#### 4.1.2 Abordagens para Estimativa de Albedo

Outro dado requerido por nossa abordagem é o albedo estimado a partir da imagem RGB de entrada. Na literatura, existem diferentes métodos para decompor imagens em seus componentes intrínsecos (albedo e sombreamento). Para nossa abordagem, precisamos apenas do albedo da imagem original.

No diagrama de fluxo da Figura 11, o processo que estima o albedo é representado pelo retângulo denominado "Modelo de Estimativa de Albedo". A Figura 13 mostra alguns resultados (Figuras 13b, 13c e 13d) das técnicas utilizadas para gerar o albedo da imagem original (Figura 13a).

## 4.2 Etapa de Treinamento

Com base nos dois dados (mapa de profundidade e albedo) da imagem original obtida pelas técnicas anteriores, procedemos com a etapa de treinamento de nossa abordagem, que inclui os seguintes processos.

#### 4.2.1 Cálculo do Mapa de Importância

Como nossa abordagem se concentra em destacar as bordas internas e externas dos objetos no mapa de profundidade, definimos um mapa de importância J para focar DENN em áreas com detalhes que podem produzir variações de sombreamento, como cantos, saliências e ranhuras de objetos. Calculamos o mapa de importância como:  $J = S \odot (1 - G)$ , onde  $\odot$  denota o produto Hadamard, S é o mapa de saliência da imagem colorida de entrada e G é a magnitude do gradiente do albedo. Usamos o canal L\* das imagens no



Figura 12: Exemplos de mapas de profundidade produzidos pelas técnicas de Yin et al. (2019) (linhas 2 e 3), Bhat, Alhashim e Wonka (2021) (linhas 4 e 5) e Song, Lim e Kim (2021) (linhas 6 e 7) em escalas de cinza e mapas de cores.



Figura 13: (a) Imagem RGB; (b-d) Exemplos de albedos gerados por algumas técnicas: (b) Nestmeyer e Gehler (2017); (c) Lettry, Vanhoey e Van Gool (2018); (d) Yunfei Liu et al. (2020).

espaço de cores L\*a\*b\* para calcular  $S \in G$  e normalizá-los para o intervalo [0, 1].

O mapa de saliência S indica regiões ricas em arestas, sejam essas arestas relacionadas a descontinuidades em materiais, superfícies ou iluminação. Calculamos S como a soma dos níveis da pirâmide Laplaciana proposta por Burt e Adelson (1987). O mapa de magnitude do gradiente G, por outro lado, destaca apenas descontinuidades em materiais e superfícies com materiais diferentes, pois o albedo é invariante às condições de iluminação. Usamos o operador de Sobel para encontrar a magnitude do gradiente absoluto aproximado em cada ponto do mapa de albedo. Finalmente, multiplicando S e o complemento de G, obtemos nosso J.

No diagrama de fluxo da Figura 11, este processo é representado pelo retângulo denominado "Cálculo do Mapa de Importância". A Figura 14 mostra exemplos dos mapas de saliência (Figura 14c) das imagens coloridas de entrada (Figura 14a) usando o canal L\* do espaço de cores L\*a\*b, das magnitudes do gradiente (Figura 14d) dos albedos (Figura 14b) usando o canal L\* do espaço de cores L\*a\*b\* e dos mapas de importância (Figura 14e) calculados usando nossa abordagem.



Figura 14: Exemplos de imagens necessárias: (a) Imagem RGB; (b) Albedo estimado; e exemplos de imagens geradas (apresentadas em mapas de cores) para calcular o mapa de importância: (c) Mapa de saliência; (d) Magnitude do gradiente e (e) Mapa de importância.



Figura 15: Mapas de características produzidos pela DENN. Nossa arquitetura consiste em quatro camadas convolucionais seguidas por uma camada totalmente conectada por canal.

#### 4.2.2 Desenho da Arquitetura e Descrição do DENN

Nossa abordagem introduz um modelo de rede neural (DENN) que usa como entrada a imagem de profundidade gerada por uma técnica anterior e o mapa de importância gerado na Seção 4.2.1. Durante o treinamento do DENN, nossa rede aprende a adicionar detalhes à imagem de profundidade.

No diagrama de fluxo da Figura 11, este processo é representado pelo retângulo de-

nominado "Modelo de Melhoramento da Profundidade". A Figura 15 mostra os mapas de características produzidos por nossa arquitetura de melhoramento de profundidade. Ele recebe como entrada um mapa de profundidade e um mapa de importância, ambos com valores num intervalo [0, 1]. Quatro camadas convolucionais processam o mapa de profundidade, produzindo mapas de características com a mesma resolução da entrada e 4, 16, 64 e 256 canais, respectivamente. Cada kernel de convolução tem um campo receptivo de  $3 \times 3$ , bias e a ativação do Softsign. Durante o treinamento, aplicamos dropout entre as camadas para evitar o sobreajuste do modelo. A arquitetura é flexível, só que para os experimentos foram configurados os hiperparâmetros de acima após uma varredura de hiperparâmetros; isso será explicado detalhadamente na Seção 5.3.2.

Usamos o mapa de importância para suprimir os vetores de características calculados pela última camada convolucional para regiões onde os detalhes não são aparentes na imagem colorida de entrada, voltando a atenção da rede para regiões com porções interessantes das superfícies. Fazemos isso multiplicando todos os 256 canais de um determinado vetor pelo valor de importância associado ao seu respectivo pixel no mapa de importância. Como o mapa de importância é composto por valores constantes, a esperança durante o treinamento é que valores próximos de zero no mapa de importância reduzam a influência dessas regiões no ajuste dos pesos da rede.

No último estágio da rede, usamos uma camada totalmente conectada por canal para resumir cada vetor de características em um valor na faixa [-1, +1]. Esta camada inclui um bias e a ativação de Tangente Hiperbólica (Tanh). Corrigimos o mapa de profundidade dado como entrada somando o mapa de características resultante a ele.

#### 4.2.3 Renderização da cena baseada no Modelo de Reflexão de Phong

Para que nosso modelo de rede neural DENN adicione detalhes à imagem do mapa de profundidade, ele usa uma função de perda que considera um termo baseado no modelo de reflexão de Phong (Phong (1975)):

$$\hat{x} = k_a i_a + \sum_{m \in \text{ ligths}} \left( k_d \left( \vec{L}_m \cdot \vec{N} \right) i_{m,d} + k_s \left( \vec{R}_m \cdot \vec{V} \right)^{\alpha} i_{m,s} \right).$$
(4.1)

Na Equação 4.1,  $\hat{x}$  é a iluminação de um determinado ponto da superfície,  $k_a$ ,  $k_d$ , e  $k_s$  são, respectivamente, as constantes de reflexão ambiente, difusa e especular para o material naquele ponto, enquanto  $\alpha$  é uma constante de brilho. Em relação às fontes de luz,  $i_a$  controla a iluminação ambiente, e  $i_{m,d}$  e  $i_{m,s}$  são as intensidades dos componentes especular e difuso da fonte de luz m. Além disso, · denota o produto escalar,  $\vec{L}_m$  é a direção unitária deste ponto na superfície em direção à fonte de luz m e  $\vec{R}_m$  é a direção unitária que um raio de luz perfeitamente refletido tomaria a partir deste ponto na superfície. Finalmente,  $\vec{N}$  é o vetor normal unitário neste ponto da superfície e  $\vec{V}$  é a direção unitária que aponta para o observador.

Devido a que não conseguimos reproduzir uma técnica que nos permita encontrar o componente especular da cena, então assumimos que materiais lambertianos estão por toda a cena, o que significa que não trabalhamos com o componente especular, ou seja,  $k_s = (0, 0, 0)$ . Também consideramos a ausência de luz ambiente, ou seja,  $i_a = (0, 0, 0)$ , e uma única fonte de luz branca (preferencial) definida para cada pixel na imagem, ou seja,  $i_{\text{pref},d} = (1, 1, 1)$ . Após simplificações, a Equação 4.1 é escrita como:

$$\hat{x} = k_d \left( \vec{L}_{\text{pref}} \cdot \vec{N} \right). \tag{4.2}$$

Assim, com base nesta equação simplificada, vamos gerar uma imagem renderizada onde cada pixel desta imagem está associado a um ponto visível na cena. A informação necessária para gerar uma imagem sintética  $\hat{X}$  da cena avaliando a Equação 4.2 por pixel é: (i) um mapa de parâmetros de reflexão difusa de superfície (também conhecido como albedo prior); (ii) um mapa normal; e (iii) um mapa de direções de luz. Exemplos desses mapas são mostrados nas Figuras 16b, 16c e 16d respectivamente.

Calculamos o mapa normais conforme apresentado na Seção 4.2.3.1, que é derivado do mapa de profundidade produzido pelo modelo preliminar de estimativa de profundidade monocular (Seção 4.1.1). Também calculamos os mapas de direções de luz usando o procedimento descrito na Seção 4.2.4.

#### 4.2.3.1 Cálculo do Mapa de Normais

Como mencionado acima, um dos termos necessários para gerar a imagem sintética  $\hat{X}$  é o mapa normal da imagem. Isso é calculado a partir de uma imagem de profundidade. No nosso caso, a imagem de profundidade que usaremos é aquela produzida por DENN (Seção 4.2.2) e o vetor normal  $\vec{N}$  atribuído a cada pixel é calculado usando o produto vetorial dos pixels vizinhos. Na literatura também existem outras técnicas para o cálculo de mapas de normais a partir de mapas de profundidade, como é o caso do Barron e Malik (2014) que é baseado em convoluções.

No diagrama de fluxo da Figura 11, este processo é representado pelo retângulo deno-



Figura 16: (a) Imagem RGB; (b) Albedo prior; (c) Mapa de normais; (d) Mapa de direções de luz.



Figura 17: (a) Imagem RGB; (b) Mapa de profundidade; (c) Mapa de normais.

minado "Calculo de Mapa de Normais". A Figura 17c mostra alguns exemplos de mapas de normais gerados com a técnica do produto vetorial dos pixels vizinhos a partir do mapas de profundidade da Figura 17b.

#### 4.2.3.2 Cálculo do Modelo de Reflexão do Phong

Com os componentes necessários (i.e. albedo, mapa de normais e mapa de direções da luz), podemos obter as imagens sintéticas renderizadas com base no modelo de reflexão de Phong que posteriormente será usado na função de perda.

No diagrama de fluxo da Figura 11, este processo é representado pelo retângulo denominado "Renderização usando o Modelo de Reflexão de Phong". A Figura 18b mostra exemplos dessas imagens sintéticas renderizadas.

#### 4.2.4 Pré-processamento do Mapa de Direções da Luz

O modelo de reflexão que usamos durante o treinamento para renderizar a cena (Equação 4.2) requer a direção que parte da superfície para uma fonte de luz estimada para cada pixel como a direção da fonte de luz preferida, ou seja, a direção  $\vec{L}_{\rm pref}$  da qual a maior quantidade de luz incidente provavelmente virá.

Para este trabalho, desenvolvemos um algoritmo para estimar um mapa de direções de luz considerando uma coleção de candidatos de direção de luz e os mapas de profundidade e albedo gerados por técnicas preliminares para a imagem colorida de entrada. É importante notar que a geração de mapas de direções de luz é realizada uma vez por imagem do conjunto de treinamento, durante uma etapa de pré-processamento da etapa de treinamento do DENN. No diagrama de fluxo da Figura 11, este processo é representado pelo retângulo pontilhado inferior denominado "Pre-processamento de Direções de Luz".

Na Seção 4.2.4.1, descrevemos o procedimento usado para gerar o conjunto de imagens renderizadas a partir de uma coleção de direções de luz. Na Seção 4.2.4.2, descrevemos o procedimento utilizado para encontrar a melhor direção de luz em cada pixel de acordo com o conjunto de imagens renderizadas.

#### 4.2.4.1 Renderizações Geradas para cada Direção de Luz

No pré-processamento das direções de luz, precisamos de um conjunto de renderizações geradas por uma coleção de direções de luz. Para isso, usamos a Equação 4.2 para obter





Figura 18: (a) Imagem RGB; (b) Imagem sintética renderizada pelo modelo de reflexão de Phong.

as imagens sintéticas renderizadas usando cada direção de luz incluída em um conjunto disponível. Dependendo do número de luzes que temos, geraremos um número igual de imagens sintéticas com base no modelo de reflexão de Phong.

Lembremos que para obter as imagens sintéticas também é necessário ter o albedo e o mapa de normais de cada imagem. Esse processo é representado no diagrama de fluxo da Figura 11 pelo retângulo denominado "Renderização usando o Modelo de Reflexão de Phong" na parte inferior.

A Figura 19b mostra alguns exemplos de diferentes imagens sintéticas renderizadas por várias fontes de luz.



Figura 19: (a) Imagem RGB; (b) Imagens sintéticas renderizadas por diferentes fontes de luz usando o modelo reflexão de Phong.

#### 4.2.4.2 Seleção da Melhor Direção de Luz por Pixel

Para cada pixel, entre as renderizações geradas para cada direção de luz, escolhemos a direção que leva ao cor perceptualmente mais próxima daquela observada na imagem colorida de entrada. A Figura 21 resume o algoritmo proposto.

Na Figura 21, X, N, A e  $\lambda$  denotam, respectivamente, a imagem colorida de entrada, o mapa de normais calculado para o mapa de profundidade preliminar previsto para X, o mapa de albedo previsto para X e o conjunto de direções de luz candidatas que foram criadas como vértices de uma icosfera (Figura 20).

Para calcular as distâncias entre um par de cores do mesmo pixel na imagem colorida original e sua versão renderizada, primeiro aplicamos um filtro de média móvel simples de tamanho  $3 \times 3$  para reduzir o ruído no par de pixels (Figura 21, linhas 1 e 2), depois calculamos a distância euclidiana das cores médias no par de pixels (Figura 21, linha 6).

No diagrama de fluxo da Figura 11, este processo é representado pelo retângulo denominado "Seleção de Melhor Direção de Luz" na parte inferior. A Figura 22b mostra alguns exemplos de mapas de direções de luz produzidos pelo nosso algoritmo.



Figura 20: (a) Imagem RGB; (b) Profundidade da imagem em visualização 3D; (c) Conjunto de direções de luz candidatas ao redor da profundidade da cena 3D.

**Input:** X: imagem colorida; N: mapa de normais; A: mapa de albedo;  $\lambda$ : conjunto de direções de luz candidatas

**Output:** *L*: mapa de direções de luz

- 1:  $X' \leftarrow X$  convertido para L\*a\*b\* e depois convoluído com um kernel de média móvel simples
- 2:  $\chi \leftarrow$  conjunto de renderizações computadas usando a Equação 4.2 com N, A e cada direção de luz  $\vec{L} \in \lambda$ , convertido para L\*a\*b\* e depois convoluído com um kernel de média móvel simples
- 3: for cada coordenada de pixel *ij* do

```
4:
          d_{min} \leftarrow \infty
          for cada imagem \hat{X}' \in \chi do
 5:
               d \leftarrow \text{distância entre as cores em } X'_{ii} \in \hat{X}'_{ii}
 6:
 7:
               if d < d_{min} then
                    L_{ij} \leftarrow \vec{L} usado para renderizar a imagem atual \hat{X}'
 8:
 9:
                    d_{min} \leftarrow d
10:
               end if
          end for
11:
12: end for
```

Figura 21: Algoritmo para calcular o mapa de direções de luz.

#### 4.2.5 Função de Perda

Nossa função de perda compara a imagem colorida de entrada X com a imagem sintética  $\hat{X}$  renderizada usando o modelo de reflexão do Phong:

$$\mathcal{L}\left(\hat{X}, X\right) = \sum_{i=1}^{H} \sum_{j=1}^{W} \|\hat{X}_{ij} - X_{ij}\|_{2}^{2} J_{ij}.$$
(4.3)

As imagens comparadas estão no espaço de cores L\*a\*b\* para capturar a diferença perceptual das cores. Aplicamos a redução de soma ao lote de pares de imagens. Na Equação 4.3,  $||v||_2^2$  é a norma Euclidiana ao quadrado do vetor v, J é o mapa de impor-





Figura 22: (a) Imagem RGB; (b) Mapa de direções de luz.

tância calculado para X (Seção 4.2.1), e W e H são, respectivamente, a largura e a altura das imagens. Em nossa experiência, regiões de imagem mais amplas e menos críticas dominam o valor de perda se não multiplicarmos a distância de cor quadrada ao mapa de importância, ofuscando o erro associado a detalhes e bordas do objeto.

No diagrama de fluxo da Figura 11, este processo é representado pelo retângulo denominado "Cálculo de Função de Perda".

## 4.3 Etapa de Inferência

Quando nossa rede termina de ser treinada, o que significa que nossa função de perda é mínima, podemos prosseguir para a etapa de inferência para testar o resultado do treinamento do DENN. Neste ponto, precisaremos apenas das técnicas preliminares de estimativa de profundidade monocular e a extração do albedo, primeiro para que possamos gerar o mapa de importância da imagem e depois para inserir esta imagem e a imagem de profundidade como as duas entradas em nosso modelo DENN, obtendo assim a imagem de profundidade produzida por DENN. Este processo é representado pelo retângulo tracejado denominado "Inferência" na parte superior no diagrama de fluxo da Figura 11.

## **5** Experimentos e Resultados

Para avaliar o desempenho da metodologia proposta, foram realizados experimentos usando diferentes técnicas de estimação de profundidade monocular na fase de treinamento e na fase de teste. Os detalhes da preparação dos dados necessários, as ferramentas utilizadas, a implementação e dos experimentos são descritos neste capítulo.

## 5.1 Seleção e Construção do Conjunto de Dados

Como peça chave para nosso modelo proposto, por estar baseado em uma rede neural, precisamos de um conjunto de dados relacionado ao problema de estimativa de profundidade. Para o nosso caso, o conjunto de dados escolhido deve conter apenas um conjunto de cenas em imagens de cores RGB para as etapas de treinamento e inferência. As profundidades verdadeiras das cenas do conjunto de dados serão necessárias apenas para a parte da avaliação quantitativa.

#### 5.1.1 Levantamento de Conjuntos de Dados Disponíveis

Na literatura sobre o problema de estimativa de profundidade monocular, diferentes conjuntos de dados têm sido utilizados para resolver este problema. Alguns deles são conjuntos de dados extraídos do mundo real, enquanto outros foram projetados sinteticamente. De maneira mais geral, todos esses conjuntos de dados podem ser divididos em conjuntos de dados de ambientes internos e conjuntos de dados de ambientes externos.

#### 5.1.1.1 Conjuntos de Dados de Ambientes Internos

 Make3D: Foi introduzido em 2008 por Saxena, Sun e Ng (2008). Este conjunto de dados inclui imagens de cores RGB e suas profundidades verdadeiras. Contém 400 imagens para treinamento e 134 para teste. As resolução das imagens é de  $2272\times1704$  pixels para imagens RGB e $305\times55$  pixels para imagens de profundidade.

- NYU Depth-V2: Foi introduzido em 2012 por Silberman et al. (2012). É um dos conjuntos de dados interiores mais utilizados na resolução de problemas de estimativa de profundidade monocular. Ele contém pares de imagens RGB e profundidades verdadeiras. É composto por 120 mil imagens de treinamento e 654 imagens de teste. A resolução destes é de 640 × 480 pixels.
- Middlebury 2014: Foi introduzido em 2014 por Scharstein et al. (2014). Contém 33 pares de imagens de alta resolução. Os pares de imagens compreendem: imagens RGB e imagens de disparidade correspondentes às profundidades verdadeiras. A resolução de cada imagem é 2872 × 1984 pixels.
- Hypersim: Foi introduzido em 2021 por Roberts et al. (2021). É um conjunto de dados de imagens sintéticas de ambientes internos criados por artistas profissionais. É composto por 77400 imagens RGB, onde cada uma delas tem sua profundidade verdadeira, imagens com segmentações semânticas, etc.

#### 5.1.1.2 Conjuntos de Dados de Ambientes Externos

- KITTI: Foi introduzido em 2013 por Geiger et al. (2013). É um dos conjuntos de dados mais aplicados em pesquisas para ambientes externos relacionados ao fluxo óptico, odometria visual, segmentação semântica, entre outros. Ele contém 39810 imagens para treinamento, 4424 para validação e 697 para teste. As imagens têm uma resolução de 1024 × 320 pixels.
- Cityscapes: Foi introduzido em 2016 por Cordts et al. (2016). É um conjunto de dados amplamente utilizado na área de segmentação. Contém 22973 pares de imagens estéreo para treinamento, com uma resolução de 2048 × 1024 pixels cada imagem.
- Driving Stereo: Foi introduzido em 2019 por Yang et al. (2019). É um conjunto de dados de imagem estéreo de grande escala contendo 182 mil imagens. A resolução das imagens é de 1762 × 800 pixels.
- PreSIL: Foi introduzido em 2019 por Hurl, Czarnecki e Waslander (2019). É um conjunto de dados sintéticos criado a partir do jogo de vídeo Grand Thefth Auto V, onde um LiDAR foi simulado para capturar as imagens. Tem 50 mil pares de imagens de resolução 1920 × 1080 pixels.

#### 5.1.2 Conjunto de Dados Selecionado: NYU Depth-V2 Dataset

Depois de revisar a literatura sobre os conjuntos de dados usados para resolver o problema de estimativa de profundidade monocular, selecionamos o conjunto de dados NYU Depth-V2 por ser mais usado em várias pesquisas. Mesmo abordagens de última geração usam esse conjunto de dados e comparam seus resultados.

Este conjunto de dados foi introduzido por Silberman et al. (2012). O conjunto de dados consiste em pares de imagens coloridas e de profundidade de ambientes internos com  $640 \times 480$  pixels. Está dividido em conjuntos de 120K pares de imagens de treinamento e 654 pares de imagens de teste. Devido às restrições no poder de computação disponível, pegamos 8,0 mil e 1,8 mil imagens aleatórias do conjunto de treinamento original para compor nossos subconjuntos de treinamento e validação, mantendo as 654 imagens originais no subconjunto de teste.

Na Figura 23 alguns exemplos de pares de imagens do conjunto de dados selecionado são mostrados.



Figura 23: Exemplos de pares de imagens do conjunto de dados NYU Depth-V2. Linha 1: imagens RGB; linha 2: profundidades verdadeiras.

## 5.2 Ferramentas Escolhidas

Para implementar e testar nossa abordagem, selecionamos algumas ferramentas de acordo com nossas necessidades. Implementamos o DENN usando PyTorch 1.8.1 e executamos nossos experimentos em uma CPU Intel Xeon E5-2698 v4 com 2,2 Ghz, 512 GB de RAM e uma GPU NVIDIA Tesla P100-SXM2 de 16 GB.

## 5.3 Metodologia Experimental

Para a etapa de treinamento e validação do modelo, usamos uma abordagem de estimativa de profundidade monocular e uma abordagem de extração de albedo. Após ajuste do modelo, realizamos testes usando três abordagens de estimativa de profundidade monocular, onde um deles é a abordagem usado no treinamento e duas outras novas abordagens.

### 5.3.1 Seleção de Abordagens de Estimativa de Profundidade Monocular e Estimativa de Albedo para Treinamento

Inicialmente, selecionamos algumas técnicas de estimativa de profundidade monocular das quais tivemos acesso à sua implementação e reproduzimos algumas destas abordagens do estado da arte (Yin et al. (2019), Bhat, Alhashim e Wonka (2021), Song, Lim e Kim (2021)) e selecionamos o trabalho de Yin et al. (2019) porque apresentou melhor definição de bordas e detalhes de maior qualidade em comparação com outras técnicas. Os resultados produzidos por sua técnica podem ser vistos na Figura 12, linhas 2 e 3.

Para a abordagem para estimar o albedo da imagem, reproduzimos algumas técnicas de decomposição intrínseca da imagem monocular (Nestmeyer e Gehler (2017), Lettry, Vanhoey e Van Gool (2018), Yunfei Liu et al. (2020)) para verificar qual delas poderia separar melhor a cor característica dos objetos de outros componentes associados com sombreamento. Após análise empírica, selecionamos o trabalho de Lettry, Vanhoey e Van Gool (2018), pois o albedo obtido por sua técnica, comparado aos demais, contém qualitativamente mais informações sobre a cor dos objetos na cena. Os resultados produzidos por sua técnica podem ser vistos na Figura 13c.

Todos os modelos de terceiros foram treinados por seus respectivos autores, baixados da Internet e usados em nossa abordagem.

#### 5.3.2 Treinamento e Otimização dos Hiperparâmetros

Para a fase de pré-processamento das direções de luz, o  $\lambda$  da Figura 21 que representa o conjunto de direções de luz candidatas que tomamos inclui 162 direções. Utilizamos esse número de direções porque foi o maior possível considerando os recursos computacionais disponíveis.

Para a fase de treinamento, ajustamos os hiperparâmetros do DENN através do conjunto de ferramentas Weights & Biases, tendo como métrica objetiva o valor da perda. Os hiperparâmetros considerados foram: número de camadas convolucionais (de 1 a 8), função de ativação (ReLU, Tanh e Softsign), dropout (de 0,2 a 0,5), tamanho do lote (2 e 4), taxa de aprendizado (de 1e-2 a 1e-5) e otimizador (SGD e Adam).

Os melhores parâmetros encontrados foram: 4 camadas convolucionais, função de ativação Softsign, dropout de 0,5, tamanho de lote de 4, taxa de aprendizado de 1e-4 e otimizador Adam. Após do ajuste de hiperparâmetros, treinar nosso modelo com os melhores valores de hiperparâmetros levou  $\sim 75$  minutos por época e 80 épocas.

As convergências do nosso modelo em relação a um valor médio de perda para a fase de treinamento e para a fase de validação podem ser vistas na Figura 24.

#### 5.3.3 Teste

Após a fase de treinamento, testamos nosso modelo usando as abordagens de Yin et al. (2019), Bhat, Alhashim e Wonka (2021) e uma abordagem recente de Kim et al. (2022) como modelo para a estimativa de profundidade monocular e a abordagem de Lettry, Vanhoey e Van Gool (2018) como modelo extrator de albedo.

Avaliamos a qualidade dos mapas de profundidade produzidos por DENN realizando análises qualitativas (Seção 5.4) e quantitativas (Seção 5.5).

## 5.4 Avaliação Qualitativa

A Figura 25 compara alguns exemplos de mapas de profundidade preditos por Yin et al. (2019) (3<sup>a</sup> linha) com mapas de profundidade produzidos por DENN (4<sup>a</sup> linha). Como se pode ver, há mudanças notórias na definição de alguns objetos. Por exemplo, na área selecionada de (a), os apoios de braços e partes do assento do sofá apresentam uma melhor definição na profundidade produzida por DENN. Na área selecionada de (b), a planta no vaso mostra as bordas mais definidas no mapa de profundidade produzido por DENN e, por sua vez, pode ser melhor distinguida do plano de fundo. Na área selecionada de (c), os objetos nas laterais mostram mais detalhes em suas bordas no mapa de profundidade profundidade produzido por DENN. Também vemos o melhoramento das bordas nos mapas da área selecionada de (d), onde podemos ver bordas mais nítidas e superfícies mais planas, e na área selecionada de (e), onde o lixo parece não se misturar com a parede de fundo. Os detalhes mais exatos dessas melhorias nas áreas selecionadas em cada exemplo podem ser vistos nas ampliações das áreas selecionadas na 5<sup>a</sup> linha.



Figura 24: Perda média nas fases de treinamento (acima) e validação (abaixo) depois de 80 épocas.

Além disso, vale a pena mencionar que as imagens de profundidade produzidas por DENN se parecem visualmente mais com as imagens de profundidade verdadeira (2<sup>a</sup> linha) em comparação com as imagens de profundidade preliminar.

Algumas melhorias se tornam aparentes ao inspecionar os mapas normais. Por exemplo, em (b) e (d), é visto na parte ampliada na 9<sup>a</sup> linha que as normais dos móveis no mapa de profundidade produzido por DENN (parte inferior) são mais detalhadas em comparação com as do mapa de profundidade preliminar (topo), ou seja, eles se ajustam melhor



Figura 25: Exemplos de resultados produzidos pela abordagem proposta por Yin et al. (2019) (profundidade preliminar) e pela nossa abordagem (profundidade produzida por DENN).

que as normais das superfícies. Além disso, nos mapas de normais de (c), a pintura na parede se destaca porque as profundidades foram corrigidas e a moldura da pintura se



Figura 26: (a) Profundidade preliminar; (b) Profundidade produzida por DENN; (c) Diferença entre as profundidades.

torna aparente.

Na Figura 26c, são mostradas as diferenças entre as profundidades preliminares e as profundidades obtidas por DENN. Obtivemos essas diferenças encontrando o erro relativo por pixel entre as duas imagens de profundidade. Como visto, as áreas brancas da imagem resultante representam as superfícies e bordas onde foram feitas mais alterações.

## 5.5 Avaliação Quantitativa

A Tabela 2 resume a avaliação de nossa abordagem em mapas de profundidade previstos por três técnicas diferentes usando métricas padrão. Usamos o teste t de Welch para testar a hipótese nula de que os valores métricos produzidos para os mapas de profundidade preliminares e os produzidos por DENN possuem médias iguais. A hipótese alternativa é que as médias das distribuições são desiguais. Assumimos um nível de significância  $\alpha = 0.05$ . Portanto, rejeitamos a hipótese nula se o valor de p for menor ou igual a 0.05.

Método		Métricas padrão						
		Menor é melhor			Maior é melhor			
		Abs-Rel	Sq-Rel	RMSE	RMSE-Log	$\delta_1$	$\delta_2$	$\delta_3$
Yin	Preliminar	0.826	0.184	0.135	0.56	0.505	0.72	0.827
	DENN	0.562	0.103	0.121	0.764	0.494	0.7	0.794
Bhat	Preliminar	0.699	0.113	0.121	0.527	0.514	0.744	0.851
	DENN	0.419	0.062	0.12	0.675	0.49	0.719	0.818
Kim	Preliminar	0.479	0.082	0.168	0.547	0.365	0.614	0.776
	DENN	0.503	0.113	0.21	1.178	0.221	0.386	0.515

Tabela 2: Comparação entre os mapas de profundidade preliminares e os mapas de profundidade produzidos por DENN em métricas padrão. Fonte em negrito indica melhores resultados ou empate.

Tabela 3: Comparação entre os mapas de profundidade preliminares e os mapas de profundidade produzidos por DENN em métricas DBE. Fonte em negrito indica melhores resultados ou empate.

		Métricas DBE			
N	Nétodo	Menor é melhor			
		$\varepsilon^{acc}_{DBE}$	$\varepsilon_{DBE}^{comp}$		
Vin	Preliminar	3.955	1.433		
1 111	DENN	3.973	1.371		
Bhat	Preliminar	3.973	1.413		
Dilat	DENN	3.994	1.342		
Kim	Preliminar	3.954	1.324		
IXIIII	DENN	3.977	1.206		

A hipótese nula foi rejeitada em todas as comparações, exceto para as métricas  $\delta_1 \in \delta_2$  de Yin et al. (2019) ( $p = 0.433 \in p = 0.113$  respectivamente), as métricas  $\delta_1 \in \text{RMSE}$  de Bhat, Alhashim e Wonka (2021) ( $p = 0.073 \in p = 0.766$  respectivamente) e para a métrica Abs-Rel de Kim et al. (2022) (p = 0.146), o que significa que tivemos um empate nesses casos.

As métricas Abs-Rel (Diferença Relativa Absoluta), Sq-Rel (Erro Relativo Quadrado), RMSE, RMSE-log e  $\delta_t$  (precisão com um limite) correspondem ao conjunto de cinco métricas padrão propostas por Eigen, Puhrsch e Fergus (2014) para comparar os mapas de profundidade produzidos com as profundidades verdadeiras. De acordo com as métricas Abs-Rel, Sq-Rel e RMS, o DENN melhorou os mapas de profundidade previstos por Yin et al. (2019) e Bhat, Alhashim e Wonka (2021). No caso de Bhat, Alhashim e Wonka (2021), os mapas produzidos por DENN também superam os resultados de última geração de Kim et al. (2022) sem melhoramento. Nenhuma melhora foi observada considerando as métricas RMSE-Log e  $\delta_t$ . Acreditamos que esse resultado esteja relacionado a questões

Estimador de prof.		Estimador de	Inferência de	Tompo total (corra)	
preliminar (segs.)		albedo (Lettry) (segs.)	DENN (segs.)	rempo totar (segs.)	
Yin	2.113	1.829	0.164	4.106	
Bhat	4.963	1.829	0.164	6.956	
Kim	5.717	1.829	0.164	7.709	

Tabela 4: Tempos médios de execução de DENN, avaliados com três técnicas diferentes de estimativa de profundidade monocular e uma técnica de estimativa de albedo.

sobre a profundidade verdadeira, que não possui bordas de objetos bem definidas devido ao sistema de captura, levando a discrepâncias superestimadas em regiões corrigidas por DENN em função dos detalhes observados nas imagens coloridas. De acordo com as métricas padrão, o modelo DENN treinado com resultados de Yin et al. (2019) não produziu melhorias nos mapas de profundidade previstos por Kim et al. (2022). Acreditamos que a razão é que os mapas de profundidade previstos por essas abordagens são muito diferentes em termos de qualidade. Nosso modelo só teve acesso a mapas de qualidade inferior durante o treinamento, tornando melhores mapas fora das amostras de distribuição aprendida.

A Tabela 3 mostra métricas relacionadas a erros de contorno de profundidade (DBEs) obtidos a partir da aplicação de métricas propostas por Koch et al. (2018) para avaliar a nitidez de imagens de profundidade. Comparando o erro de precisão  $\varepsilon_{DBE}^{acc}$  dos mapas de profundidade previstos e aprimorados, observamos que eles estão bem próximos, apesar do teste de hipótese não mostrar que são estatisticamente iguais. Esses resultados sugerem que o DENN é capaz de preservar os limites atuais de profundidade. Os resultados do erro de completude  $\varepsilon_{DBE}^{comp}$  sugerem que o DENN pode aprimorar os mapas de profundidade previstos fornecidos como entrada. Na prática, essa métrica está dizendo que o DENN produz resultados com menos arestas ausentes para os três modelos comparados, adicionando arestas ausentes às imagens de profundidade.

### 5.6 Tempo de Execução

O tempo de execução da abordagem completa também foi medido. Foram medidos o tempo médio de execução de cada uma das técnicas de estimativa de profundidade monocular (Yin et al. (2019), Bhat, Alhashim e Wonka (2021) e Kim et al. (2022)), o tempo médio de execução da técnica para estimar o albedo (Lettry, Vanhoey e Van Gool (2018)) e o tempo médio de execução da etapa de inferência de DENN. A Tabela 4 mostra os tempos médios obtidos em segundos.

Método		Métricas padrão						
		Menor é melhor			Maior é melhor			
		Abs-Rel	Sq-Rel	RMSE	RMSE-Log	$\delta_1$	$\delta_2$	$\delta_3$
Yin	Preliminar	0.826	0.184	0.135	0.56	0.505	0.72	0.827
	DENN	0.808	0.179	0.134	0.559	0.51	0.723	0.828
Bhat	Preliminar	0.699	0.113	0.121	0.527	0.514	0.744	0.851
	DENN	0.677	0.109	0.12	0.519	0.519	0.747	0.854
Kim	Preliminar	0.479	0.082	0.168	0.547	0.365	0.614	0.776
	DENN	0.466	0.082	0.169	0.549	0.365	0.61	0.771

Tabela 5: Comparação entre os mapas de profundidade preliminares e os mapas de profundidade produzidos por DENN em métricas padrão.

Tabela 6: Comparação entre os mapas de profundidade preliminares e os mapas de profundidade produzidos por DENN em métricas DBE.

		Métricas DBE				
Ν	Aétodo	Menor é melhor				
		$\varepsilon_{DBE}^{acc}$	$\varepsilon_{DBE}^{comp}$			
Vin	Preliminar	3.955	1.433			
1 111	DENN	3.956	1.431			
Bhot	Preliminar	3.973	1.413			
Dilat	DENN	3.973	1.411			
Kim	Preliminar	3.954	1.324			
171111	DENN	3.955	1.32			

## 5.7 Discussão

Para os experimentos também treinamos o DENN usando conjuntamente os mapas de profundidade produzidos por três diferentes técnicas de estimativa de profundidade (Yin et al. (2019), Bhat, Alhashim e Wonka (2021) e Kim et al. (2022)) e seus correspondentes mapas de direções de luz. Avaliamos o desempenho da rede no conjunto de dados de teste usando as métricas padrão e as métricas DBE. Os resultados são mostrados nas Tabelas 5 e 6 respectivamente.

Como é visto nas Tabelas 5 e 6, os resultados obtidos para a estimativa de profundidade preliminar e a estimativa de profundidade por DENN são muito semelhantes. Analisando esses resultados podemos concluir que, aparentemente, DENN é mais adequado para melhorar qualquer técnica alvo específica.

## **6** Conclusões

Esta dissertação teve como objetivo geral apresentar um modelo para melhorar os mapas de profundidade produzidos por abordagens de estimativa de profundidade de imagem única. O modelo apresentado foi chamado DENN. Treinamos o DENN usando uma nova função de perda que compara uma versão renderizada da cena com a imagem colorida de entrada. Assim, o DENN está desenhado para adicionar detalhes ao mapa de profundidade preliminar estimado.

Nossa abordagem mostra que as imagens de profundidade utilizadas como entradas são melhoradas em aspectos de definição de bordas, planaridade da superfície e, consequentemente, melhor distinção de objetos na cena quando estão no mesmo plano. Vemos também que o modelo DENN, apesar de ter sido treinado com um conjunto de dados, apresenta um bom desempenho quando é avaliado com outros conjuntos de dados.

Não há restrições sobre como o mapa de profundidade preliminar é produzido porque o DENN é uma abordagem não invasiva para melhorar os mapas de profundidade. Portanto, qualquer técnica, incluindo abordagens de triangulação, pode ser escolhida em teoria. No entanto, neste trabalho, voltamos nossa atenção para a estimativa de profundidade monocular, pois ela tem se destacado nos últimos anos na literatura.

Finalmente, podemos ver que a ideia de usar um modelo de iluminação, como o modelo de Phong, foi bem aproveitada para melhorar a profundidade corrigindo as normais da cena.

## 6.1 Trabalhos Futuros

Durante o desenvolvimento desta dissertação, surgiram muitas ideias complementares à parte central da nossa proposta que foram tomadas como possíveis trabalhos futuros.

Por exemplo, sobre o uso de algoritmos e técnicas opcionais para determinadas etapas do modelo DENN, como no caso de usar como entrada algum mapa de profundidade gerado por outras técnicas que não sejam estimativa monocular, foram considerados os mapas de profundidade obtidos pelo Kinects, mapas de profundidade gerados por triangulação e até mesmo a opção de usar nuvens de pontos 3D. Com relação ao uso do modelo de Phong, para a geração do mapa de normais a partir do mapa de profundidade, outras técnicas alternativas poderiam ser utilizadas ao invés do produto interno de pixels vizinhos utilizados por nossa abordagem. Também consideramos a opção do treinamento ponta-a-ponta das redes, ou seja, após treinar DENN, descongelar a rede de estimativa de profundidade preliminar e retreinar os dois modelos simultaneamente.

Outras ideias para trabalhos futuros incluem o uso de conjuntos de dados onde as imagens coloridas são de melhor qualidade, como no caso de conjuntos de dados atuais. Com esta ideia, foi também considerada a utilização de conjuntos de dados sintéticos, que têm a particularidade de oferecer imagens coloridas de altíssima qualidade onde as saliências em superfícies e bordas de objetos são muito mais apreciadas. Também consideramos a ideia de substituir o modelo de Phong por um modelo de rendering baseado em aprendizado profundo que seja treinado a partir do mapa de profundidade e do mapa de albedo.

Finalmente, em relação à aplicação do nosso modelo, pensamos em avaliar seu desempenho em tempo real, ou seja, ao invés de usar apenas imagens para inferência, também utilizamos sequências de vídeo como dados de entrada e aplicamos em robôs para sua autonomia, em veículos autônomos, em sistemas de segurança e inspeção baseados na visão ou na indústria.

# REFERÊNCIAS

ALHASHIM, Ibraheem; WONKA, Peter. High quality monocular depth estimation via transfer learning. **arXiv preprint arXiv:1812.11941**, 2018.

ATAPOUR-ABARGHOUEI, Amir; BRECKON, Toby P. Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer. In: PROCEEDINGS of the IEEE Conference on Computer Vision and Pattern Recognition. [S. l.: s. n.], 2018. p. 2800–2810.

BARRON, Jonathan T; MALIK, Jitendra. Shape, illumination, and reflectance from shading. IEEE Transactions on Pattern Analysis and Machine Intelligence, v. 37, n. 8, p. 1670–1687, 2014.

BHAT, Shariq Farooq; ALHASHIM, Ibraheem; WONKA, Peter. AdaBins: Depth estimation using adaptive bins. In: PROCEEDINGS of the IEEE Conference on Computer Vision and Pattern Recognition. [S. l.: s. n.], 2021. p. 4009–4018.

BURT, Peter J; ADELSON, Edward H. The Laplacian pyramid as a compact image code. In: READINGS in Computer Vision. [S. l.: s. n.], 1987. p. 671–679.

CAO, Yuanzhouhan et al. Monocular depth estimation with augmented ordinal depth relationships. **IEEE Transactions on Circuits and Systems for Video Technology**, v. 30, n. 8, p. 2674–2682, 2020.

CASSER, Vincent et al. Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. In: 01. PROCEEDINGS of the AAAI Conference on Artificial Intelligence. [S. l.: s. n.], 2019. v. 33, p. 8001–8008.

CHAKRABARTI, Ayan; SHAO, Jingyu; SHAKHNAROVICH, Greg. Depth from a single image by harmonizing overcomplete local network predictions. Advances in Neural Information Processing Systems, v. 29, p. 2658–2666, 2016.

CHANG, Wenjie; ZHANG, Yueyi; XIONG, Zhiwei. Transformer-based Monocular Depth Estimation with Attention Supervision. In: 32ND British Machine Vision Conference. [S. l.: s. n.], 2021. p. 136. CHAWLA, Hemang et al. Multimodal scale consistency and awareness for monocular self-supervised depth estimation. In: IEEE International Conference on Robotics and Automation. [S. l.: s. n.], 2021. p. 5140–5146.

CHEN, Long et al. Self-supervised monocular image depth learning and confidence estimation. **Neurocomputing**, v. 381, p. 272–281, 2020.

CHEN, Tian et al. Improving monocular depth estimation by leveraging structural awareness and complementary datasets. In: ECCV. [S. l.: s. n.], 2020. p. 90–108.

CHOI, Hong-Tae et al. SSD-EMB: An improved SSD using enhanced feature map block for object detection. **Sensors**, v. 21, p. 2842, 2021.

CORDTS, Marius et al. The Cityscapes dataset for semantic urban scene understanding. In: IEEE Conference on Computer Vision and Pattern Recognition. [S. l.: s. n.], 2016. p. 3213–3223.

DUAN, Xiangyue et al. High quality depth estimation from monocular images based on depth prediction and enhancement sub-networks. In: IEEE International Conference on Multimedia and Expo. [S. l.: s. n.], 2018. p. 1–6.

EIGEN, David; PUHRSCH, Christian; FERGUS, Rob. Depth map prediction from a single image using a multi-scale deep network. Advances in Neural Information **Processing Systems**, v. 27, p. 2366–2374, 2014.

FU, Huan et al. Deep ordinal regression network for monocular depth estimation. In:PROCEEDINGS of the IEEE Conference on Computer Vision and Pattern Recognition.[S. l.: s. n.], 2018. p. 2002–2011.

GEIGER, Andreas et al. Vision meets robotics: The KITTI dataset. **The International Journal of Robotics Research**, v. 32, n. 11, p. 1231–1237, 2013.

GODARD, Clément; MAC AODHA, Oisin; BROSTOW, Gabriel J. Unsupervised monocular depth estimation with left-right consistency. In: PROCEEDINGS of the IEEE Conference on Computer Vision and Pattern Recognition. [S. l.: s. n.], 2017. p. 270–279.

GUIZILINI, Vitor et al. 3D packing for self-supervised monocular depth estimation. In: PROCEEDINGS of the IEEE Conference on Computer Vision and Pattern Recognition. [S. l.: s. n.], 2020. p. 2485–2494.

GURRAM, Akhil et al. Monocular depth estimation through virtual-world supervision and real-world SFM self-supervision. **IEEE Transactions on Intelligent Transportation Systems**, 2021.

#### REFERÊNCIAS

HARSÁNYI, Károly et al. A Hybrid CNN Approach for Single Image Depth Estimation: A Case Study. In: PROCEEDINGS of the 11th International Conference on Multimedia and Network Information Systems. [S. l.: s. n.], 2018. v. 833, p. 372–381.

HE, Lei; WANG, Guanghui; HU, Zhanyi. Learning depth from single images with deep neural network embedding focal length. **IEEE Transactions on Image Processing**, v. 27, n. 9, p. 4676–4689, 2018.

HU, Junjie et al. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. In: IEEE Winter Conference on Applications of Computer Vision. [S. l.: s. n.], 2019. p. 1043–1051.

HURL, Braden; CZARNECKI, Krzysztof; WASLANDER, Steven. Precise synthetic image and lidar (PreSIL) dataset for autonomous vehicle perception. In: IEEE Intelligent Vehicles Symposium. [S. l.: s. n.], 2019. p. 2522–2529.

JIAO, Jianbo et al. Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss. In: PROCEEDINGS of the European Conference on Computer Vision. [S. l.: s. n.], 2018. p. 53–69.

KIM, Doyeon et al. Global-local path networks for monocular depth estimation with vertical CutDepth. **ArXiv**, abs/2201.07436, 2022.

KLINGNER, Marvin et al. Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance. In: ECCV. [S. l.: s. n.], 2020. p. 582–600.

KOCH, Tobias et al. Evaluation of CNN-based single-image depth estimation methods.In: PROCEEDINGS of the European Conference on Computer Vision Workshops.[S. l.: s. n.], 2018. p. 331–348.

KUZNIETSOV, Yevhen; STUCKLER, Jorg; LEIBE, Bastian. Semi-supervised deep learning for monocular depth map prediction. In: PROCEEDINGS of the IEEE Conference on Computer Vision and Pattern Recognition. [S. l.: s. n.], 2017. p. 6647–6655.

LEE, Jae-Han; KIM, Chang-Su. Monocular depth estimation using relative depth maps. In: PROCEEDINGS of the IEEE Conference on Computer Vision and Pattern Recognition. [S. l.: s. n.], 2019. p. 9729–9738.

LEE, Jin Han; HAN, Myung-Kyu et al. From big to small: Multi-scale local planar guidance for monocular depth estimation. **ArXiv**, abs/1907.10326, 2019.

LETTRY, Louis; VANHOEY, Kenneth; VAN GOOL, Luc. Unsupervised deep single-image intrinsic decomposition using illumination-varying image sequences. In: 7. COMPUTER Graphics Forum. [S. l.: s. n.], 2018. v. 37, p. 409–419.

LI, Shuai et al. Hierarchical object relationship constrained monocular depth estimation. Pattern Recognition, v. 120, p. 108116, 2021.

LI, Yaoxin et al. Depth estimation from monocular image and coarse depth points based on conditional GAN. In: MATEC Web of Conferences. [S. l.: s. n.], 2018. v. 175, p. 03055.

LI, Yuanzhen; LUO, Fei; XIAO, Chunxia. Self-supervised coarse-to-fine monocular depth estimation using a lightweight attention module. **Computational Visual** Media, p. 1–17, 2022.

LIN, Xiao et al. Depth estimation and semantic segmentation from a single RGB image using a hybrid convolutional neural network. **Sensors**, v. 19, n. 8, p. 1795, 2019.

LIU, Beyang; GOULD, Stephen; KOLLER, Daphne. Single image depth estimation from predicted semantic labels. In: IEEE Conference on Computer Vision and Pattern Recognition. [S. l.: s. n.], 2010. p. 1253–1260.

LIU, Fayao; SHEN, Chunhua; LIN, Guosheng. Deep convolutional neural fields for depth estimation from a single image. In: PROCEEDINGS of the IEEE Conference on Computer Vision and Pattern Recognition. [S. l.: s. n.], 2015. p. 5162–5170.

LIU, Fayao et al. Learning depth from single monocular images using deep convolutional neural fields. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 38, n. 10, p. 2024–2039, 2015.

LIU, Yunfei et al. Unsupervised learning for intrinsic image decomposition from a single image. In: PROCEEDINGS of the IEEE Conference on Computer Vision and Pattern Recognition. [S. l.: s. n.], 2020. p. 3248–3257.

MAN, Yunze et al. GroundNet: Monocular ground plane normal estimation with geometric consistency. In: PROCEEDINGS of the 27th ACM International Conference on Multimedia. [S. l.: s. n.], 2019. p. 2170–2178.

MANCINI, Michele et al. Toward domain independence for learning-based monocular depth estimation. **IEEE Robotics and Automation Letters**, v. 2, n. 3, p. 1778–1785, 2017.

MASOUMIAN, Armin et al. GCNdepth: Self-supervised monocular depth estimation based on graph convolutional network. **arXiv preprint arXiv:2112.06782**, 2021.

MERN, John M et al. Visual depth mapping from monocular images using recurrent convolutional neural networks. In: AIAA Scitech 2019 Forum. [S. l.: s. n.], 2019. p. 1189.

MO, Donglin et al. Soft-aligned gradient-chaining network for height estimation from single aerial images. **IEEE Geoscience and Remote Sensing Letters**, v. 18, n. 3, p. 538–542, 2020.

NESTMEYER, Thomas; GEHLER, Peter V. Reflectance adaptive filtering improves intrinsic image estimation. In: PROCEEDINGS of the IEEE Conference on Computer Vision and Pattern Recognition. [S. l.: s. n.], 2017. p. 6789–6798.

PARIDA, Kranti Kumar; SRIVASTAVA, Siddharth; SHARMA, Gaurav. Beyond image to depth: Improving depth prediction using echoes. In: PROCEEDINGS of the IEEE Conference on Computer Vision and Pattern Recognition. [S. l.: s. n.], 2021. p. 8268–8277.

PHONG, Bui Tuong. Illumination for computer generated pictures. Communications of the ACM, v. 18, n. 6, p. 311–317, 1975.

QUEIROZ MENDES, Raul de et al. On deep learning techniques to boost monocular depth estimation for autonomous navigation. Robotics and Autonomous Systems, v. 136, p. 103701, 2021.

RANFTL, René; BOCHKOVSKIY, Alexey; KOLTUN, Vladlen. Vision transformers for dense prediction. In: PROCEEDINGS of the IEEE International Conference on Computer Vision. [S. l.: s. n.], 2021. p. 12179–12188.

ROBERTS, Mike et al. Hypersim: A Photorealistic Synthetic Dataset for Holistic Indoor Scene Understanding. In: PROCEEDINGS of the IEEE International Conference on Computer Vision. [S. l.: s. n.], 2021.

SAXENA, Ashutosh; CHUNG, Sung; NG, Andrew. Learning depth from single monocular images. Advances in Neural Information Processing Systems, v. 18, p. 1161–1168, 2005.

SAXENA, Ashutosh; CHUNG, Sung H; NG, Andrew Y. 3-D depth reconstruction from a single still image. International Journal of Computer Vision, v. 76, n. 1, p. 53–69, 2008.

SAXENA, Ashutosh; SCHULTE, Jamie; NG, Andrew Y et al. Depth estimation using monocular and stereo cues. In: PROCEEDINGS of the 20th International Joint Conference on Artificial Intelligence. [S. l.: s. n.], 2007. v. 7, p. 2197–2203.

SAXENA, Ashutosh; SUN, Min; NG, Andrew Y. Make3d: Learning 3D scene structure from a single still image. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 31, n. 5, p. 824–840, 2008.

SCHARSTEIN, Daniel et al. High-resolution stereo datasets with subpixel-accurate ground truth. In: GERMAN Conference on Pattern Recognition. [S. l.: s. n.], 2014. p. 31–42.

SILBERMAN, Nathan et al. Indoor segmentation and support inference from RGBD images. In: ECCV. [S. l.: s. n.], 2012. p. 746–760.

SONG, Minsoo; LIM, Seokjae; KIM, Wonjun. Monocular depth estimation using Laplacian Pyramid-based depth residuals. **IEEE Transactions on Circuits and Systems for Video Technology**, v. 31, n. 11, p. 4381–4393, 2021.

TABIAN, Iuliana; FU, Hailing; SHARIF KHODAEI, Zahra. A convolutional neural network for impact detection and characterization of complex composite structures. **Sensors**, v. 19, n. 22, p. 4933, 2019.

TAN, Daniel Stanley et al. Single-image depth inference using generative adversarial networks. **Sensors**, v. 19, n. 7, p. 1708, 2019.

TOSI, Fabio et al. Learning monocular depth estimation infusing traditional stereo knowledge. In: PROCEEDINGS of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S. l.: s. n.], 2019. p. 9799–9809.

WANG, Lijun et al. SDC-depth: Semantic divide-and-conquer network for monocular depth estimation. In: PROCEEDINGS of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S. l.: s. n.], 2020. p. 541–550.

WEDEL, Andreas et al. Realtime depth estimation and obstacle detection from monocular video. In: JOINT Pattern Recognition Symposium. [S. l.: s. n.], 2006. p. 475–484.

WOFK, Diana et al. FastDepth: Fast monocular depth estimation on embedded systems. In: INTERNATIONAL Conference on Robotics and Automation. [S. l.: s. n.], 2019. p. 6101–6108.

XIAN, Ke et al. Monocular relative depth perception with web stereo data supervision. In: PROCEEDINGS of the IEEE Conference on Computer Vision and Pattern Recognition. [S. l.: s. n.], 2018. p. 311–320. XUE, Feng et al. Toward hierarchical self-supervised monocular absolute depth estimation for autonomous driving applications. In: IEEE/RSJ International Conference on Intelligent Robots and Systems. [S. l.: s. n.], 2020. p. 2330–2337.

YANG, Guorun et al. Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios. In: IEEE Conference on Computer Vision and Pattern Recognition. [S. l.: s. n.], 2019. p. 899–908.

YE, Xinchen et al. DRM-SLAM: Towards dense reconstruction of monocular SLAM with scene depth fusion. **Neurocomputing**, v. 396, p. 76–91, 2020.

YIN, Wei et al. Enforcing geometric constraints of virtual normal for depth prediction.In: PROCEEDINGS of the IEEE International Conference on Computer Vision.[S. l.: s. n.], 2019. p. 5684–5693.

ZHANG, Ziyu et al. Monocular object instance segmentation and depth ordering with CNNs. In: PROCEEDINGS of the IEEE International Conference on Computer Vision. [S. l.: s. n.], 2015. p. 2614–2622.

ZHOU, Tinghui; KRAHENBUHL, Philipp; EFROS, Alexei A. Learning data-driven reflectance priors for intrinsic image decomposition. In: PROCEEDINGS of the IEEE International Conference on Computer Vision. [S. l.: s. n.], 2015. p. 3469–3477.
## APÊNDICE A - MUDANÇAS NA PROF. DURANTE O TREINAMENTO



Figura 27: Exemplos de mapas de profundidade melhorando durante o treinamento.



Figura 28: Mais exemplos de mudanças na profundidades durante o treinamento que mostram a melhoria dos mapas de profundidade.

## APÊNDICE B - MUDANÇAS NAS NORMAIS DURANTE O TREINAMENTO



Figura 29: Exemplos de mapas de normais melhorando durante o treinamento.



Figura 30: Mais exemplos de mudanças nos mapas de normais durante o treinamento que mostram a melhoria dos mapas de normais.