UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
CURSO DE CIÊNCIA DA COMPUTAÇÃO

LORENZO PEZZI DAL' AQUA

Estimativa de profundidade usando uma única imagem esférica

Monografia apresentada como requisito parcial para a obtenção do grau de Bacharel em Ciência da Computação

Orientador: Prof. Dr. Claudio Rosito Jung

Porto Alegre
11 de Janeiro de 2018

AGRADECIMENTOS

Em especial ao Claudio pela orientação e companheirismo ao longo deste trabalho e de toda a minha graduação.

Ao Thiago por todo o apoio e contribuições essenciais para a realização do trabalho.

À Nicole por estar sempre ao meu lado, me apoiando e motivando para cada nova etapa.

Aos meus pais João Carlos e Simone, por me tornarem quem sou e me apoiarem sempre, sem eles nada disso seria possível.

Por fim, a todos os professores, funcionários e colegas do Instituto de Informática da UFRGS, por sua parte na experiência maravilhosa que foi ser aluno do curso de graduação em Ciência da Computação.

RESUMO

A estimativa de profundidade é um componente essencial de diversas aplicações de visão computacional, e um dos assuntos mais extensivamente estudados na área. Recentemente, houveram avanços na utilização de métodos de aprendizagem de máquina para realizar a estimativa a partir de uma única imagem, diferentemente do método tradicional de casamento estéreo, que utiliza duas ou mais imagens. Imagens esféricas, ou omnidirecionais, possuem um campo de visão de $360°$ e oferecem informação contextual muito maior da cena em relação a imagens planares. A estimativa de profundidade de cenas esféricas pode ser de grande utilidade para diversas aplicações, como navegação, compreensão de cena e realidade virtual. Tradicionalmente, no entanto, são necessárias múltiplas câmeras esféricas ou câmeras especializadas para a estimativa de profundidade na esfera, e com a popularização de câmeras $360°$ e o fácil acesso a métodos de geração de panoramas $360°$, é de interesse poder aplicar as técnicas de estimativa de profundidade utilizando um única imagem ao domínio das imagens esféricas. Este trabalho propõe um método para estimar profundidades utilizando apenas uma única imagem esférica a partir da divisão da esfera e projeção em planos. São estimadas as profundidades no domínio planar utilizando métodos já existentes, e então projeta-se as estimativas de volta para a esfera, combinando as estimativas de cada divisão da esfera em um único mapa de profundidades para toda a esfera.

Palavras-chave: Imagens esféricas. imagens omnidirecionais. imagens 360. estimativa de profundidade. visão computacional.

Depth estimation using a single spherical image

ABSTRACT

Depth estimation is an essential component in many computer vision applications, and one of the most extensively studied subjects in the field. Recently, advancements were made in applying machine learning methods to estimate depths from a single image, differently from traditional stereo matching methods, which need two or more images. Spherical, or omnidirectional, images, have a 360° field of view and provide higher contextual information about a scene in comparison to planar images. Depth estimation of spherical scenes can be an asset for many applications, such as navigation, scene understanding and virtual reality. Normally, however, multiple spherical cameras, or specialized camera configurations are needed to perform depth estimation on spherical images, and with the rise in ease of access to 360° cameras and panorama generation tools, it is of interest to be able to apply single image depth estimation method to spherical images. This work proposes a method for estimating depth from a single spherical image by dividing the sphere and projecting each section onto a plane. Depths are estimated on the planar domain using existing methods, and these estimates are projected back to the sphere, combining each sections' estimates into a single depth map for the whole sphere.

Keywords: spherical images, omnidirectional images, 360 images, depth estimation, computer vision.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CCD *Charge-coupled device*

MRF *Markov Random Field*

CNN *Convolutional Neural Network*

SVD *Singular Value Decomposition*

NCC *Normalized Cross Correlation*

SUMÁRIO

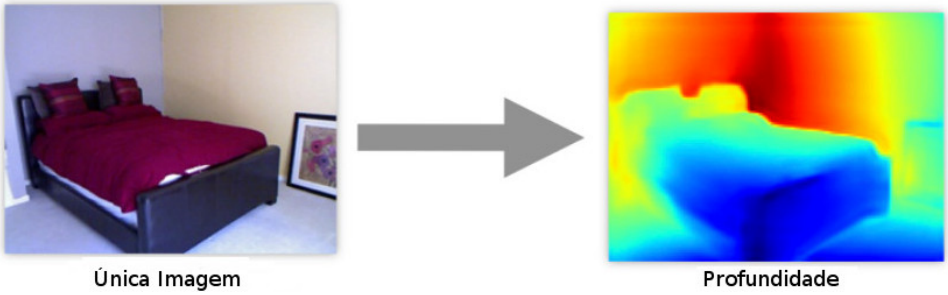
1 INTRODUÇÃO

1.1 Motivação

A estimativa de profundidade é um componente crucial para a compreensão da geometria tridimensional de uma cena. Imagens esféricas, ou omnidirecionais, possuem um campo de visão de $360°$ e oferecem informação contextual muito maior sobre a cena em relação à imagens planares comuns. Logo, a estimativa de profundidade nestas imagens pode ser de utilidade para uma gama de aplicações, desde compreensão de cenas, detecção de objetos, navegação, reconstrução 3D, realidade virtual e realidade aumentada.

O método tradicional para obtenção de mapas de profundidade é o casamento estéreo, que utiliza as correspondências entre um par de imagens para estimar a profundidade em uma cena, mas requer múltiplas imagens e/ou configurações de câmeras especializadas. Por outro lado, há diversos resultados promissores na obtenção de mapas de profundidade a partir de uma única imagem, como os trabalhos de Eigen e Fergus (2015) e Liu, Shen e Lin (2015) (Figura 1.1), que utilizam redes neurais convolucionais. Estas redes, no entanto, são treinadas a partir de imagens planares que possuem valores de profundidades conhecidos. Na medida em que a utilização de câmeras omnidirecionais se popularizam em aplicações como realidade virtual e aumentada, robótica, etc. seria proveitoso utilizar estes métodos de estimativa a partir de imagens monoculares para obter mapas de profundidade a partir de uma única imagem esférica. Contudo, são escassas as bases de dados de imagens esféricas com valores de profundidades conhecidos que se assemelhem às utilizadas para o treinamento das redes com imagens planares, portanto, seria interessante um método que adapte as redes já treinadas com imagens planares para utilização com imagens esféricas.

Figura 1.1: Estimativa de profundidade monocular



Fonte: (EIGEN; FERGUS, 2015), modificado

1.2 Objetivo

O objetivo do trabalho é desenvolver e testar um método que utilize as técnicas já existentes de estimativa de profundidade utilizando uma única imagem planar para obter mapas de profundidade a partir de imagens esféricas.

1.3 Estrutura do Texto

O trabalho é estruturado da seguinte forma:

- **Conceitos básicos:** Apresentar os principais conceitos abordados, de modo a contextualizar a discussão posterior do desenvolvimento do método.
- **Trabalhos relacionados:** Relacionar a pesquisa já realizada tanto em estimativa de profundidade quanto com imagens esféricas, e contextualizar o método proposto.
- **O método proposto:** Descrever o método desenvolvido, suas etapas e funcionamento.
- **Resultados:** Avaliação do método e apresentação de resultados qualitativos, quantitativos e visualizações dos mapas de profundidade obtidos.
- **Conclusão:** Discutir a contribuição do método e possibilidades de trabalho futuro.

2 CONCEITOS BÁSICOS

2.1 Imagens planares vs. esféricas

Uma imagem é uma representação 2D do espaço tridimensional, capturada por um meio sensível à luz, no caso de imagens digitais, sensores, por exemplo, os CCD (*Charge-coupled device*) (GONZALEZ; WOODS, 2006). Nesta seção detalharemos como são obtidas imagens planares, e como estas diferem das imagens omnidirecionais que são o foco deste trabalho.

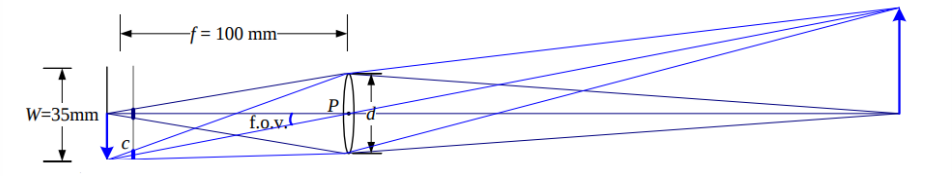
2.1.1 Imagens planares

Uma câmera comum mapeia o espaço tridimensional para um plano através da projeção perspectiva. Como descrito em Szeliski (2011), um modelo simples para o funcionamento de uma câmera é o modelo de lentes finas. Como descrito na figura (2.1), o campo de visão de uma imagem obtida por uma câmera comum é caracterizado por um ângulo $f.o.v.$ (*field of fiew*) e depende da relação entre o tamanho W do sensor de imagem e da distância focal f da lente. Exceto usando lentes especiais que distorcem a projeção perspectiva (e.g. *fisheye*), o campo de visão será limitado em no máximo $180°$.

2.1.2 Imagens esféricas

Imagens esféricas, também conhecidas como imagens omnidirecionais, idealmente capturam a luz de todas as direções, e possuem um campo de visão de $360°$. Como explicado na seção anterior, a projeção perspectiva é limitada a $180°$, então, para mapear todos os pontos da esfera de raio unitário, com duas dimensões, é necessário utilizar ou-

Figura 2.1: Modelo de lentes finas



O campo de visão ($f.o.v.$) é dependente da proporção entre o tamanho do sensor W e a distância focal f. Fonte: (SZELISKI, 2011)

tras projeções. Há uma míriade de projeções para projetar esferas, a maioria desenvolvida com mapas da Terra em mente, (WEISSTEIN, 2018b), por exemplo, cilíndricas, cônicas e azimutais. Ao longo do trabalho será utilizada somente a projeção equiretangular, por ser um formato comumente utilizado por câmeras que obtém imagens esféricas, e pela disponibilidade de bases de dados como a proposta em Xiao et al. (2012) para testes.

2.1.2.1 Projeção equiretangular

A projeção equiretangular é um caso específico da projeção cilíndrica equidistante. De acordo com Weisstein (2018a), as equações de mapeamento para latitude ϕ e longitude λ na esfera para coordenadas horizontais x e verticais y no plano (analógico) se dão por:

$$x = \lambda, \;\; y = \phi, \tag{2.1}$$

ou seja, a projeção equiretangular é simplesmente uma conversão direta de longitude na esfera para coordenada horizontal na imagem e latitude na esfera para coordenada vertical na imagem. No entanto, ângulos são contínuos, a latitude varia de 0 a 2π radianos e a longitude de 0 a π radianos, mas as coordenadas de pixel são discretas e sua variação depende da resolução da imagem. Dada uma imagem com W pixels na horizontal e H pixels na vertical como na figura 2.2, temos as seguintes relações:

$$\frac{x}{W} = \frac{\lambda}{2\pi}, \quad \text{e} \quad \frac{y}{H} = \frac{\phi}{\pi}, \tag{2.2}$$

e a partir destas relações obtemos a latitude ϕ e longitude λ na esfera para qualquer pixel com coordenadas (x, y) na imagem projetada através de:

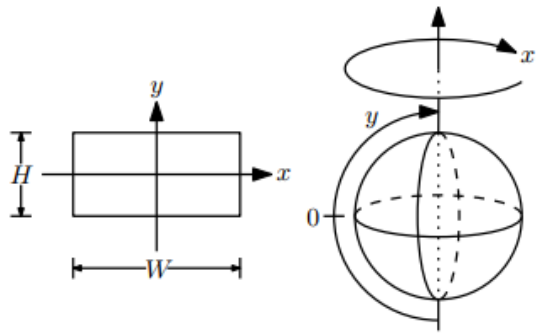
$$\lambda = \frac{2\pi x}{W}, \quad \phi = \frac{\pi y}{H}. \tag{2.3}$$

2.1.2.2 Captura de imagens esféricas

Como explicado anteriormente, câmeras comuns possuem um limite de campo de visão, e ainda não foi projetada uma câmera omnidirecional ideal. Porém imagens omnidirecionais podem ser obtidas de múltiplas maneiras, utilizando uma ou mais câmeras tradiconais ou especializadas:

- É possível alterar a projeção da imagem obtida por uma única câmera através de lentes ou espelhos especiais (cônicos, esféricos, hiperbólicos, etc.). As lentes ou

Figura 2.2: Imagem esférica em projeção equiretangular



Fonte: Jianxiong Xiao, 3D Geometry for panorama

espelhos permitem que a câmera possa capturar toda a cena, ou pelo menos aumentam seu campo de visão. Exemplos são descritos em Nayar (1997) e Onoe et al. (1998).

- Combinando informação obtida de múltiplas imagens em perspectiva, obtidas por múltiplas câmeras ou por uma câmera em movimento, e colando as imagens em uma única imagem panorâmica. Conhecido também como mosaico, há exemplos em Mann e Picard (1994), Szeliski (1996) e Peleg e Herman (1997).

2.2 Estimativa de profundidade

A estimativa de profundidade consiste em estimar a distância entre objetos contidos em uma ou mais imagens e uma câmera de referência. Há múltiplas aplicações em robótica, como navegação e reconhecimento de objetos, além de ser um componente essencial para a reconstrução 3D. No caso de imagens esféricas, a reconstrução 3D de uma imagem $360°$ tem aplicações diretas em realidade virtual e realidade aumentada.

Tradicionalmente, a estimativa de profundidade é realizada através do casamento estéreo, ou *structure from motion*, onde a profundidade é estimada utilizando a informação de correspondência entre pixels de duas ou mais imagens. Para tanto, é preciso conhecer ou estimar a pose e os parâmetros de calibração da(s) câmera(s) utilizadas. Este é uma das áreas mais antigas e mais extensivamente estudadas na área de visão computacional (SZELISKI, 2011).

No entanto, no contexto deste trabalho o foco é em métodos de estimativa de profundidade a partir de uma única imagem, que são um desenvolvimento mais recente (Ver Seção 3.1.2). A estimativa de profundidade a partir de uma única imagem é um

problema desafiador, pois não se tem as indicações de profundidade da visão binocular, e não é possível se basear apenas na informação local, sendo necessário levar em conta o contexto global da imagem. Diversos métodos foram bem-sucedidos ao tratar a estimativa de profundidade como uma tarefa de aprendizado, especialmente utilizando redes neurais convolucionais.

A Seção 3.1.2 faz uma revisão de alguns dos métodos já propostos para a estimativa a partir de uma única imagem, porém neste trabalho não será detalhado o funcionamento destes, pois a proposta é a extensão para imagens esféricas independente do método de estimativa de profundidade, como será detalhado na Seção 4.

3 TRABALHOS RELACIONADOS

Este capítulo faz uma breve revisão de estudos relacionados ao método proposto no trabalho. Inicialmente são abordados métodos para estimativa de profundidade em imagens planares, para em seguida discutir os métodos já existentes para a estimativa a partir de imagens esféricas.

3.1 Estimativa de profundidade em imagens planares

Os métodos mais tradicionais para estimativa de profundidade utilizam duas ou mais vistas de uma cena, logo estes serão abordados inicialmente como referência. Na seção posterior serão apresentados os diversos métodos para a estimativa utilizando uma única imagem, que são o foco e inspiração deste trabalho.

3.1.1 Múltiplas imagens

A estimativa de profundidade a partir de múltiplas imagens, especificamente o casamento estéreo, é um problema que já foi extensivamente abordado. Há múltiplas revisões da literatura, desde Scharstein, Szeliski e Zabih (2001), que propuseram uma taxonomia dos algoritmos de estéreo, métricas de avaliação, e uma base de dados para teste, até mais recentemente Hamzah e Ibrahim (2016), que na sua revisão avaliam dezenas de algoritmos, propõem uma taxonomia de cada etapa dos algoritmos de estéreo, e faz uma análise das diversas revisões da literatura já feitas.

Além do casamento de pares estéreo tradicional, mais focados na reconstrução 3D das cenas, há variações utilizando mais de duas imagens (FURUKAWA; HERNáNDEZ, 2015), ou estimando a profundidade a partir de correspondências entre imagens obtidas de uma câmera em movimento (*structure from motion*) (RANFTL et al., 2016). Há também métodos que estimam a profundidade sem variar a orientação da câmera, mas sim a iluminação da cena (WOODHAM, 1989), (ABRAMS; HAWLEY; PLESS, 2012), ou os parâmetros da câmera, tipicamente o foco (PENTLAND, 1987), (WEI; WU, 2015).

3.1.2 Única imagem

Este trabalho pretende estender técnicas de estimativa de profundidade a partir de uma única imagem para o domínio das imagens esféricas. Desta forma, esta seção fornece uma visão geral dos métodos já propostos para tal.

Sem as informações de profundidades fornecidas pela visão binocular, a estimativa de profundidade pode ser vista como um problema de aprendizado, como exemplificado por Saxena, Chung e Ng (2005) e Saxena, Chung e Ng (2008), que abordaram o problema como um problema de aprendizado supervisionado. Através de campos aleatórios de Markov (*Markov Random Fields*, ou MRF), treinados para reconhecer sinais monoculares de profundidade, como diferença de texturas, oclusão, tamanho dos objetos, etc., obtiveram mapas de profundidade razoavelmente coerentes, porém dependentes de alinhamento horizontal. Já o trabalho de Liu, Gould e Koller (2010) realiza uma segmentação da imagem em classes semânticas para utilizar informação contextual e conhecimento prévio (ex: o céu está normalmente longe da câmera) para estimar as profundidades.

Recentemente, métodos baseados em redes neurais convolucionais (CNNs) têm alcançado novas marcas em diversas aplicações de visão computacional. Eigen, Puhrsch e Fergus (2014) descrevem um método baseado em duas redes profundas, que estimam profundidades globais, para então refinar estas em uma maior resolução, o que foi estendido para incorporar outras informações da geometria tridimensional da cena (EIGEN; FERGUS, 2015), os vetores normais das superfícies e categorias semânticas, obtendo resultados no estado da arte. Liu, Shen e Lin (2015) propuseram uma alternativa que combina campos aleatórios com redes neurais profundas em o que eles chamam de campos neurais convolucionais profundos (*deep convolutional neural fields*).

Visto o sucesso dos métodos de aprendizado supervisionados, e visto a dificuldade de obtenção de grandes quantidades de imagens com dados reais de profundidade para treinamento, já foram propostos métodos utilizando aprendizado não-supervisionado, ou semi-supervisionado. Entre eles, Kuznietsov, Stückler e Leibe (2017) propuseram um método semi-supervisionado que utiliza mapas de profundidade reais esparsos para o treinamento, utilizando redes neurais para produzir os mapas densos. Godard, Mac Aodha e Brostow (2016) exploram o uso de pares estéreo para o treinamento ao invés de dados reais de profundidade, e Zhou et al. (2017) utilizam redes neurais para estimativa de profundidade e pose de câmera, atrelando as duas utilizando a síntese de vistas como objetivo de treinamento. Estes métodos produzem resultados comparáveis ou superiores ao estado

da arte em estimativa de profundidade com métodos de aprendizado supervisionado.

3.2 Estimativa de profundidade em imagens esféricas

Esta seção discute e referencia métodos encontrados na literatura que já realizam a estimativa de profundidade em imagens esféricas e sua relação com o método proposto. Não foram encontrados na literatura métodos que estimem a profundidade a partir de uma única imagem, somente com configurações especiais de câmeras e/ou luz.

Diversos estudos na área de robótica propõem métodos para a estimativa de profundidade e reconstrução 3D para navegação e reconhecimento de objetos utilizando configurações especiais de câmera, sempre com múltiplas imagens. Koyasu, Miura e Shirai (2001) utilizam um par de câmeras omnidirecionais baseadas em espelhos, e utilizando também a correspondência temporal entre imagens, computa o casamento estéreo em tempo real. Li, Tang e Shum (2001) utilizam uma única câmera rotacionada ao longo do tempo para formar imagens esféricas multiperspectiva, aproveitando a redundância entre imagens. Zhu (2001) faz uma análise de diversos métodos de obtenção de imagens omnidirecionais para uso em estéreo, e a qualidade da estimativa de profundidade de cada método.

O método proposto por Orghidan, Mouaddib e Salvi (2005) utiliza apenas uma imagem esférica para fazer a estimativa de profundidade, porém faz uso de técnicas de projeção de luz estruturada para evitar o casamento estéreo, propondo um sensor que combina uma câmera omnidirecional com um projetor para uso em navegação. Apesar de não utilizar múltiplas imagens, este tipo de método é limitado ao sensor específico que projeta a luz, e é suscetível à iluminação da cena.

Também buscando estender algoritmos treinados para imagens planares para o domínio de imagens esféricas, Su e Grauman (2017) propõem uma rede convolucional no domínio esférico que traduza as características de filtros planares levando em conta as distorções na vista esférica. Citam uma subdivisão da esfera em planos como proposto neste trabalho, porém a consideram muito computacionalmente intensiva. No entanto, o método é proposto para realizar extração de características, o que diferentemente da estimativa de profundidade, não impõe necessariamente coêrencia global nos resultados.

4 O MÉTODO PROPOSTO

4.1 Visão geral do método

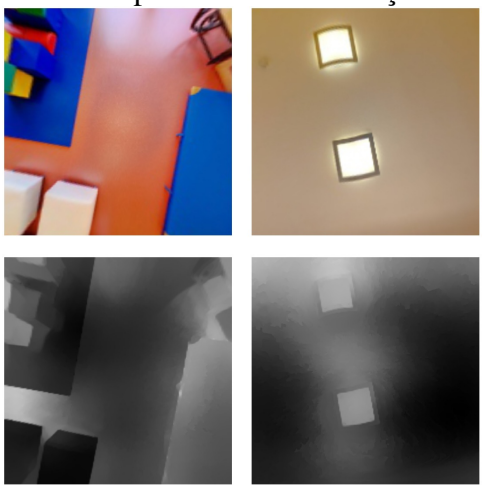
Recentemente, a aplicação de redes neurais convolucionais mostrou resultados impressionantes na estimativa de profundidade a partir de uma única imagem. No entanto, a utilização destes algoritmos com imagens 360° não produz a mesma qualidade de resultados, pois o treinamento destes é feito utilizando imagens planares, e a distorção introduzida pela projeção equiretangular torna as informações visuais de produndidade diferentes nas imagens esféricas, além de representar uma área muito maior da cena do que as imagens utilizadas para o treinamento. Como mostra a figura 4.2, a rede neural tem dificuldade em estimar a profundidade de regiões homogêneas, devido à falta de informação contextual. Na projeção equiretangular, o piso e o teto de cenas internas ocupam uma grande região (pois ficam próximas dos pólos da esfera), degradando a eficácia destas quando utilizadas diretamente nas imagens equiretangulares. É possível ver exemplos de estimativas nos pólos da esfera na 4.1, e os mapas de profundidade gerados mostram múltiplas profundidades diferentes em regiões homogêneas, e diferenças grandes nas estimativas entre objetos próximos, como os objetos no teto da cena.

O método proposto neste trabalho pretende estender as técnicas de estimativa de profundidade a partir de uma única imagem planar, sem modificação, para o domínio das imagens esféricas. Para tanto, é necessário converter a informação da imagem esférica para o domínio planar. Como não é possível projetar todos os 360° da esfera para única imagem em perspectiva sem perder informação, a abordagem proposta é seccionar a esfera de modo que o campo de visão de cada seção seja projetável para uma imagem planar sem introduzir distorções significativas. As profundidades de cada seção planificada são estimadas através de técnicas como as mencionadas na Seção 3.1.2, e então projetadas novamente para o domínio esférico. Finalmente, mapa de profundidades da imagem esférica é construído através da re-projeção das seções. Entretanto, há duas questões que precisam ser abordadas para a utilização do método.

Primeiramente, seções da esfera com pouca informação contextual, como uma parede homogênea, não fornecem informação visual suficiente para a estimativa de profundidade. Além disso, o piso e o teto de cenas internas capturadas com câmeras esféricas tipicamente diferem bastante das cenas utilizadas para treinamento das redes planares.

Um segundo problema é a “colagem” das estimativas de profundidade estimadas

Figura 4.1: Estimativas de profundidade em seções nos pólos da esfera



Topo: Seções nos pólos da esfera. Embaixo: Mapa de profundidades estimadas. Fonte: O Autor

Figura 4.2: Estimativas de profundidade preliminares

Esquerda: Imagem original, Centro: Profundidades estimadas diretamente na esfera, Direita: Descontinuidades ao estimar as profundidades em seções da esfera. Fonte: O Autor

em cada imagem planar. Se usarmos somente seções disjuntas da esfera, na maioria dos casos são geradas descontinuidades entre as seções, como é possível ver na Figura 4.2 (os pólos da esfera são omitidos da estimativa nesta figura). Isso ocorre pois a estimativa é local à cada seção, não há informação contextual da seção adjacente.

Para gerar um mapa de disparidades mais suave, a alternativa proposta é utilizar seções da esfera com sobreposição entre si. As profundidades nas sobreposições entre as seções devem idealmente ser iguais, fornecendo uma informação contextual que permite fazer uma ponderação das seções para minimizar as diferenças na região sobreposta, e propagando a ponderação para o resto de cada seção de modo a manter as profundidades de cada seção também contínuas. Como as imagens planares nos pólos tipicamente correspondem ao teto e ao chão (cenários que tipicamente não foram treinados nas redes planares existentes, e assim tendem a produzir estimativas de profundidade ruins), foi decidido excluir os pólos da esfera da estimativa final. Além disso, as regiões comumente representadas nos polos da esfera (e.g. piso, teto) são de pouco interesse para diversas aplicações, por serem normalmente homogêneas.

A figura 4.3 ilustra o método descrito, e as seções a seguir detalham cada etapa do

processo.

4.2 Seccionamento e projeção para o plano

Nesta seção será detalhada a primeira etapa do método, onde a esfera é seccionada e cada seção é projetada para um plano.

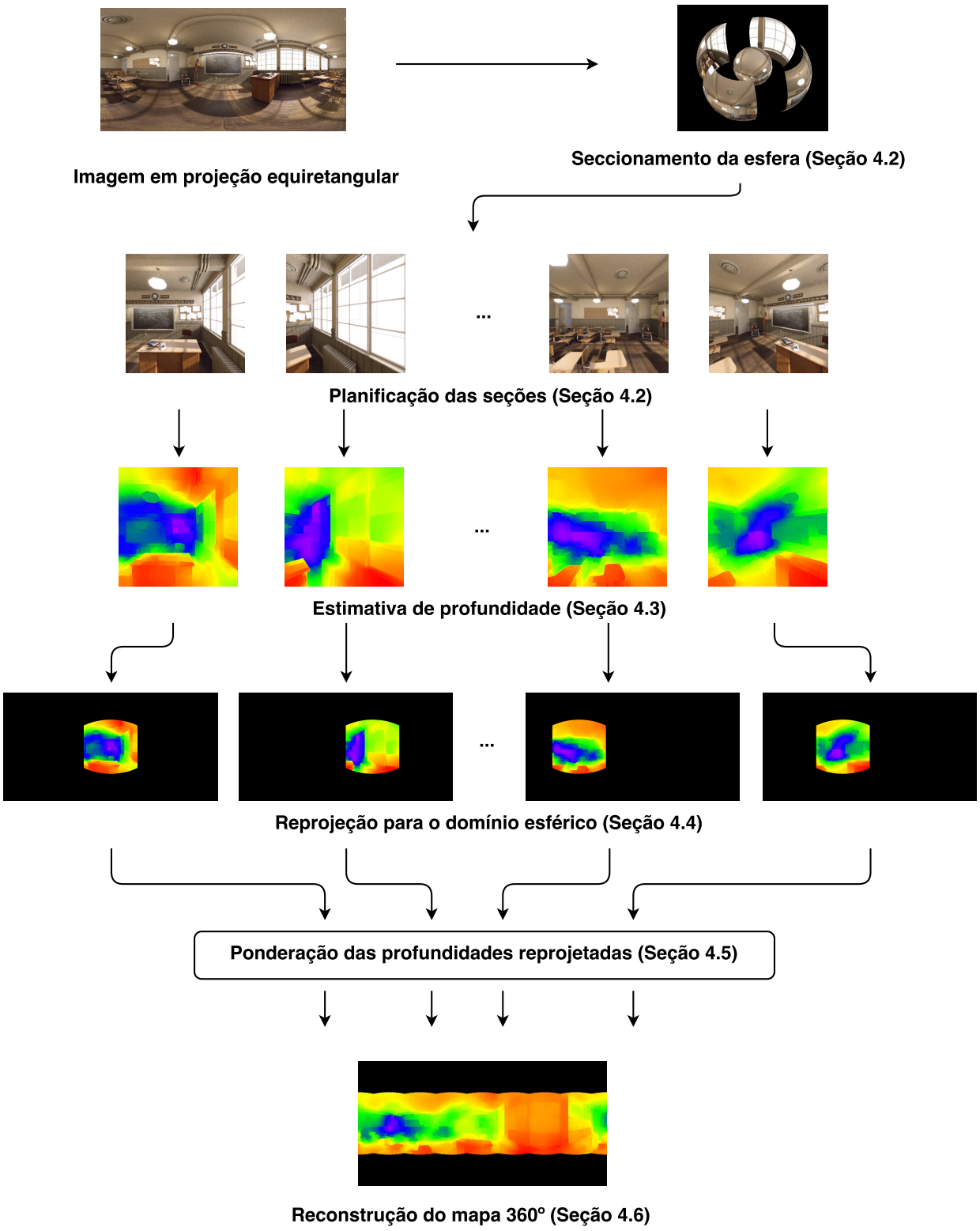
4.2.1 Seccionamento da esfera

A primeira etapa do método é definir as seções da esfera cujas profundidades serão estimadas. Como os pólos da esfera serão omitidos da estimativa, o objetivo é dividir os 360° do campo de visão horizontal da esfera, por simplicidade, em seções iguais. O campo de visão de cada seção é dado por $\theta = 360^\circ/N$, onde N é o número de seções. São escolhidas $2N$ seções, com campo de visão θ, e deslocamento de $\theta/2$ ao longo do eixo horizontal (equador) da esfera entre si, de modo que aproximadamente metade de cada seção está sobreposta a cada uma de suas seções vizinhas. A Figura 4.4 ilustra seções com $N = 4$ e $\theta = 90^\circ$, tanto disjuntas e com sobreposição no espaço tridimensional, com translações e escala para permitir a visualização.

4.2.2 Projeção para o plano

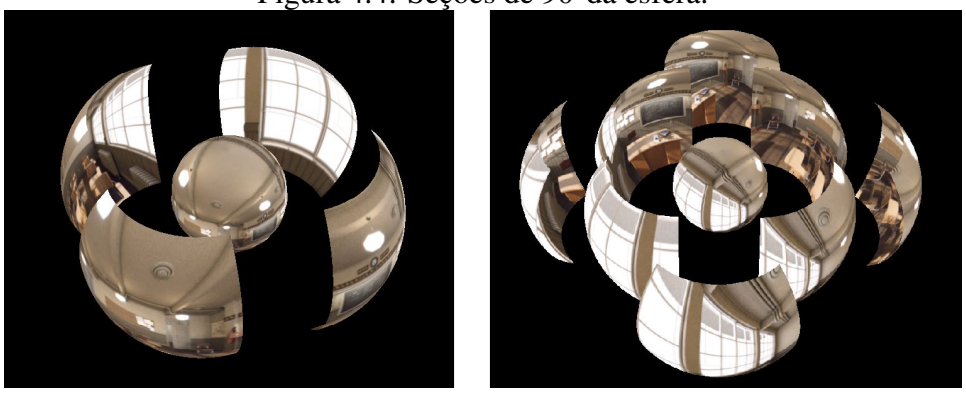
Dada uma seção da esfera de campo de visão θ, tanto latitudinal quanto longitudinal, centro de projeção com ângulo longitudinal λ_0, e latitudinal ϕ_0 na esfera, se deseja projetá-la para uma imagem planar com P linhas e P colunas, com campo de visão igual nos eixos horizontal e vertical. É necessário primeiro calcular os ângulos correspondentes na esfera para cada pixel no plano. Considerando o plano tangente à esfera, isto é somente encontrar as coordenadas angulares da interseção da superfície da esfera com as linhas que conectam cada pixel no plano até o centro da esfera. Logo, cada pixel (x, y) é mapeado para ângulos λ, ϕ na esfera, e como explicado na seção 2.1.2.1, dividindo os ângulos pelo número de pixels na dimensão, temos o índice do pixel na imagem equiretangular. Estes índices podem não ser exatos, no entanto, e nesse caso é simplesmente realizada uma interpolação linear entre os valores discretos da imagem equiretangular para obter o valor do pixel no plano. A implementação das projeções da esfera para o

Figura 4.3: Ilustração do método proposto



Fonte: O Autor

Figura 4.4: Seções de 90°da esfera.



Esquerda: seções disjuntas. Direita: seções com sobreposição. Fonte: O Autor

plano e do plano para a esfera foram baseadas em código MATLAB proposto em Xiao et al. (2012).

4.3 Estimativa de profundidade

O método proposto pretende ser genérico com relação à técnica de estimativa de profundidade em imagens planares utilizada. Logo, as técnicas são utilizadas sem modificação. Foram estudadas as técnicas com código disponível e redes neurais já treinadas, e entre estas foi escolhido o método descrito em Liu, Shen e Lin (2015) para a implementação, devido ao suporte a maiores resoluções. No entanto, a etapa de estimativa de profundidade é realizada independentemente das outras, e a técnica utilizada pode ser substituída sem alterações no método proposto.

4.4 Projetando a seção de volta para a esfera

Após estimar as profundidades, o caminho inverso da seção 4.2.2 deve ser realizado. No entanto, a projeção da esfera para um plano é de apenas uma seção, logo, não é possível reconstruir toda a esfera, somente um fragmento da imagem equiretangular reprojetada possui informação. O processo consiste em converter as latitudes e longitudes na esfera tridimensional, para os índices correspondentes dos pixels no plano. Porém, como só um fragmento da esfera será reprojetado, é necessário escolher somente os ângulos que fazem parte da seção. A conversão de ângulo para pixel é feita de maneira similar à de pixel para ângulo. Dado um plano tangente à esfera cuja projeção sobre a

esfera cobre um ângulo igual ao campo de visão da imagem, para cada ângulo discreto da imagem equiretangular é possível traçar uma linha até um ponto no plano. A cor do pixel na imagem equiretangular reprojetada será a cor do pixel na intersecção, e novamente, no caso de valores não exatos, é feita a interpolação linear nos valores dos pixels do plano. O resultado é uma imagem equiretangular na qual somente uma seção é preenchida, o resto da imagem tendo valor zero.

4.5 Ponderação dos mapas de profundidade de cada seção

A partir do conjunto de estimativas de profundidade das seções da esfera, armazenados em projeção equiretangular, queremos minimizar as diferenças nas regiões de sobreposição entre as seções. Inicialmente, pode-se pensar em ponderar cada seção da esfera por um valor que minimize a diferença nas profundidades onde as seções se sobrepôe. Porém, ao invés de ponderar toda a seção com o mesmo peso, o método proposto utiliza um valor de ponderação para cada linha de cada região de sobreposição.

Sejam $I(x, y)$ uma imagem esférica S_L e S_R duas seções da esfera armazenadas em imagen na projeção equiretangular, $D_L(x, y)$ e $D_R(x, y)$ estimativas de profundidade de S_L e S_R, também na projeção equiretangular, P o conjunto dos pixels região de sobreposição, ou seja, pixels em comum entre as seções, e L o conjunto das linhas da região de sobreposição entre S_L e S_R. Queremos calcular pesos w_{kL} e w_{kR} para cada linha k, com $k \in L$, para ponderar as linhas de S_L e S_R. Para isso, queremos escolher pesos que minimizem a diferença entre as estimativas de profundidade em cada pixel onde há sobreposição, ou seja, minimizar o somatório:

$$W_D = \sum_{(i,j)\in P} (w_{iL} D_L(i, j) - w_{iR} D_R(i, j))^2 \tag{4.1}$$

No entanto, selecionar os pesos w_{kL} e w_{kR} de forma independente para cada linha k pode gerar pesos muito diferentes entre linhas adjacentes nas imagens, gerando descontinuidades indesejadas no mapa de profundidade. Logo é interessante adicionar um termo de regularização ao problema de minimização para manter a coerência entre linhas adjacentes:

$$W_C = \sum_{(i,j)\in P} T(i, j)((w_{iL} - w_{i+1L})^2 + (w_{iR} - w_{i+1R})^2), \tag{4.2}$$

onde

$$T(i,j) = \begin{cases} \alpha, \text{ se } |I(i,j) - I(i+1,j)| < \gamma \\ \beta, \text{ se } |I(i,j) - I(i+1,j)| \geq \gamma \end{cases} \tag{4.3}$$

é o peso do termo de regularização com base nas cores da imagem esférica. Os parâmetros α, β e γ são constantes: α é um fator de escala para linhas com cor similar, e β o fator para linhas com cor não similar; γ define o valor limite da diferença de cor entre dois pixels. Inicialmente foi testada a utilização de uma função $T(i,j)$ que variasse linearmente com a diferença entre a intensidade dos pixels. Porém, por tentativa e erro, a utilização de um limiar demonstrou resultados melhores na ponderação.

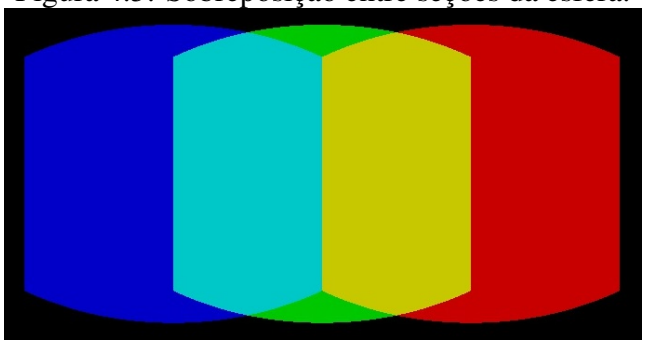
A função de custo total a ser minimizada é então dada por $W_T = W_D + W_C$. Para evitar a solução trivial $w_{kL} = w_{kR} = 0$, uma alternativa é montar o sistema linear abaixo

$$\begin{bmatrix} D_L(1,j) & 0 & \dots & 0 & -D_R(1,j) & 0 & \dots & 0 \\ 0 & D_L(2,j) & \dots & 0 & 0 & -D_R(2,j) & \dots & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & D_L(n,j) & 0 & 0 & \dots & -D_R(n,j) \\ \alpha & -\alpha & \dots & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & \alpha & -\alpha & \dots & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \end{bmatrix} \begin{bmatrix} w_{1L} \\ w_{2L} \\ \vdots \\ w_{nL} \\ w_{1R} \\ w_{2R} \\ \vdots \\ w_{nR} \end{bmatrix} = 0 \tag{4.4}$$

e selecionar como solução o vetor singular correspondente ao menor valor singular, obtidos pela decomposição em valores singulares (*Singular Value Decomposition*, ou SVD). Tal solução corresponde ao vetor de norma unitária que minimiza W_T . É importante salientar que o fator de escala dos pesos (e dos mapas de disparidade) é arbitrário, visto que as técnicas de estimativa a partir de uma única imagem fornecem apenas distâncias relativas.

Calculados os pesos que minimizam a função de custo, se quer ponderar as profundidades de cada seção da esfera. Como mostra a figura 4.5, dadas três seções da esfera representadas pelas cores vermelha, azul e verde, as sobreposições da seção verde com as azul e vermelha estão representadas pelas cores ciano e amarelo, respectivamente. As regiões sobrepostas tiveram pesos calculados para cada uma de suas linhas. Para obter pesos para todos os pixels da seção em verde onde não há sobreposição, a cada linha é feita a interpolação entre o peso da sobreposição em ciano e da sobreposição em ama-

Figura 4.5: Sobreposição entre seções da esfera.



Verde, vermelho e azul: seções da esfera. Amarelo e ciano: sobreposição entre as seções. Fonte: O Autor

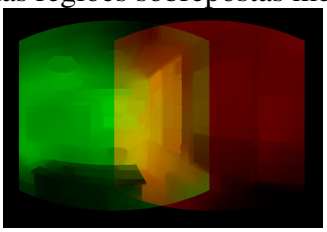
relo, de modo que pixels mais próximos da sobreposição à esquerda tenham pesos mais próximos dos pesos desta região de sobreposição, e os mais à direita tenham pesos próximos dos pesos da região de sobreposição à direita. Há ainda linhas em verde que não possuem linhas correspondentes das regiões de sobreposição em ciano e amarelo. Para estas linhas, os valores dos pesos da última e primeira linha das regiões de sobreposição são extrapolados.

É feito este processo para cada mapa de profundidades de cada seção da esfera, para então poder ser feita a reconstrução do mapa da esfera completa.

4.6 Reconstrução do mapa de profundidade completo

Idealmente, a ponderação faria com que a diferença entre as profundidades em regiões de sobreposição entre as seções fosse zero. No entanto, mesmo que seja possível minimizar as diferenças entre um único par de seções S_0 e S_1 adjacentes, é possível que as diferenças entre as linhas de S_1 e as da próxima seção adjacente S_2 necessitem de pesos completamente diferentes das entre S_1 e S_0 para serem minimizadas. As diferenças entre uma seção e sua sobreposição à esquerda podem ser negativas e entre sua sobreposição a direita positivas. A natureza cíclica das sobreposições entre as seções torna este problema ainda mais grave. Dito isto, dificilmente a ponderação elimina as diferenças entre as profundidades estimadas nas regiões com sobreposição. Além disso, o método utilizado para ponderação, linha a linha, é ainda mais sucetível a isto, pois minimiza a diferença ao longo de toda a sobreposição. Logo, as diferenças entre pixels individuais ainda podem ser significativas, como é possível ver na figura 4.6, onde duas seções sobrepostas são mapeadas para dois canais de cor diferentes. Uma sobreposição sem diferenças mostraria

Figura 4.6: Diferenças nas regiões sobrepostas mesmo após a ponderação.



Duas seções sobrepostas mapeadas para canais de cor diferentes. Fonte: O Autor

as profundidades na sobreposição somente em tons de amarelo, apenas com variação de intensidade e não de matiz, o que não é o caso, pois é possível ver regiões com matiz mais vermelha e mais verde na região, que demonstra que os valores das profundidades nestes pontos ainda são diferentes entre as seções. Para poder combinar as seções de volta em uma esfera completa, é preciso uma estratégia para combinar a informação destoante nas regiões onde há duas estimativas diferentes.

O método tem como objetivo remover as descontinuidades entre profundidades das seções da esfera e gerar um mapa suave. Para combinar suavemente as estimativas diferentes em pixels em região de sobreposição, é possível fazer uma combinação convexa dos valores de profundidade nas seções que formam a sobreposição. Para cada pixel p com coordenadas (x, y) em uma região de sobreposição P entre as seções já ponderadas da esfera S_{wL} e S_{wR}, dados x_R e x_L a maior e menor coordenadas horizontais de P, respectivamente, se tem que os valores do mapa de disparidade final F da esfera completa de cada $p \in P$ são

$$F(x, y) = (1 - \alpha(x))\, S_{wL}(x, y) + \alpha(x) S_{wR}(x, y), \tag{4.5}$$

com

$$\alpha(x) = (x - x_L)/(x_R - x_L) \tag{4.6}$$

sendo o peso que varia linearmente entre 0 e 1 ao longo de cada linha horizontal na imagem equiretangular.

Para as regiões onde não há sobreposição, somente há um valor de profundidade a ser colocado no mapa final, logo somente são coladas as profundidades obtidas de cada seção nestes casos. Como veremos na próxima seção, a desvantagem de se fazer esta interpolação dos valores nas regiões sobrepostas produz alguns artefatos nos mapas de profundidade nas bordas das regiões de sobreposição.

5 RESULTADOS

5.1 Análise quantitativa

Nesta seção é apresentada uma análise quantitativa do método proposto, são definidas a métrica utilizada para análise, o conjunto de imagens com profundidades conhecidas para teste e, por fim, são apresentados e avaliados os resultados.

5.1.1 Imagens para teste

Para realizar uma análise quantitativa do método proposto, devido à escassez de bases de dados de imagens esféricas com informação real de profundidade, foi escolhido utilizar imagens sintéticas que possuem mapas de profundidade associados, um conjunto de dados proposto em Silveira e Jung (2017). As imagens esféricas em diversos pontos de vista de uma cena sintética realista de uma sala de aula foram *renderizadas* com o software Blender.

5.1.2 Métrica

A métrica escolhida para avaliação do método foi a Correlação Cruzada Normalizada (*Normalized Cross-Correlation*, ou NCC) (ZHAO; HUANG; GAO, 2006), devido à sua invariância à escala. Conforme mencionado na seção anterior, a escala das profundidades é arbitrária, de modo que métricas de comparação direta de valores (como soma dos quadrados das diferenças) não são adequadas. A NCC entre duas imagens de profundidade, f e g, é descrita pela fórmula a seguir:

$$NCC(f,g) = \frac{1}{n\sigma_f\sigma_g}\sum_{x=0}^{n}(f(x)-\bar{f})(g(x)-\bar{g}), \tag{5.1}$$

onde n é o número de pixels das imagens, σ_h é o desvio padrão da imagem h e $\bar{h}$ é a média dos valores da imagem h. A métrica varia entre $[-1, 1]$, sendo -1 a diferença máxima entre as imagens e 1 a diferença mínima.

5.1.3 Resultados quantitativos

As tabelas 5.1 e 5.2 apresentam os resultados da aplicação da métrica entre o mapa de disparidade obtido utilizando o método proposto e a profundidade real (coluna NCC_N). Cada tabela apresenta os resultados para um campo de visão θ diferente, mas os parâmetros da ponderação α, β e γ são os mesmos para as duas tabelas. Os parâmetros utilizados foram $\alpha = 200$, $\beta = 50$ e γ é 10% da diferença quadrática máxima entre dois pixels, como pixels de imagens coloridas tem 3 canais com valores entre 0 e 255, $\gamma = (3 * 255^2)/10$. Para fins de comparação, foi aplicada a rede neural de Liu, Shen e Lin (2015) diretamente à imagem esférica, e a coluna NCC_D representa a respectiva correlação cruzada normalizada . Além disso, é exibida a diferença entre as colunas NCC_N e NCC_D para cada imagem do conjunto, e a média da NCC entre todas as imagens do conjunto para cada uma das técnicas de estimativa. A tabela 5.1 representa a execução do método proposto utilizando seções com campo de visão $\theta = 90°$, e a tabela 5.2 com campo de visão $\theta = 120°$. Como o mapa gerado pelo método não cobre os pólos da esfera, o NCC calculado ignora estas regiões na estimativa diretamente pela rede neural para fazer uma comparação justa.

Analisando os resultados, é possível perceber que a estimativa ainda difere bastante das profundidades reais pela métrica utilizada. Entretanto, na maioria dos casos há uma melhora significativa em relação à aplicação direta da estimativa de profundidade para imagens planares.

5.2 Resultados Qualitativos

Além da análise quantitativa, é interessante fazer uma inspeção qualitativa (visual) dos mapas de profundidade obtidos. Em conjunto com as cenas sintéticas utilizadas para a análise quantitativa, foram escolhidas cenas reais da base de dados SUN360 proposta em Xiao et al. (2012) para visualização dos mapas de profundidade. As figuras 5.1 e 5.2 exibem imagens de entrada, seguidas da estimativa feita pelo método proposto, e estimativas feitas com as redes neurais diretamente na imagem esférica. Os mapas de profundidade são apresentados como imagens em escala de cinza onde os valores mais escuros são profundidades menores e os mais claros profundidades maiores. Como os pólos da esfera são excluídos da estimativa pelo método apresentado, o preto nos mapas de profundidade indica a falta de informação nestes pontos da esfera.

Tabela 5.1: Métricas dos resultados para o método com $\theta = 90°$, $\alpha = 200$, $\beta = 50$, $\gamma = 10\%$

Imagem	NCC_N	NCC_D	*Diferença*
classroom000	-0.2523	0.1052	-0.3576
classroom001	0.4114	0.0554	0.3560
classroom002	0.2762	0.1538	0.1224
classroom003	0.3796	0.1030	0.2765
classroom004	0.1358	0.3563	-0.2205
classroom005	0.2578	0.1613	0.0966
classroom006	0.1758	0.0906	0.0852
classroom007	0.4354	0.1353	0.3001
classroom008	0.2341	0.1946	0.0396
classroom009	0.3967	0.0725	0.3242
classroom010	0.2897	0.1298	0.1599
classroom011	-0.1006	0.1435	-0.2441
classroom012	0.3929	0.1714	0.2214
classroom013	0.0139	0.1761	-0.1622
classroom014	0.2079	0.2344	-0.0264
classroom075	0.2761	0.1701	0.1060
classroom085	0.2632	0.2312	0.0321
classroom095	0.3495	0.2218	0.1277
classroom105	0.3500	0.2258	0.1242
classroom115	0.3363	0.2333	0.1029
classroom125	0.3448	0.2137	0.1311
classroom135	0.3534	0.2018	0.1517
classroom145	0.3197	0.2108	0.1089
classroom150	0.3048	0.1574	0.1475
classroom155	0.3232	0.1934	0.1299
classroom165	0.3431	0.1539	0.1892
classroom175	0.3319	0.1730	0.1589
classroom185	0.3280	0.1887	0.1393
classroom195	0.3170	0.1734	0.1436
Média	0.2688	0.1735	0.0953

Fonte: O Autor

Tabela 5.2: Métricas dos resultados para o método com $\theta = 120°$, $\alpha = 200$, $\beta = 50$, $\gamma = 10\%$

Imagem	NCC_N	NCC_D	*Diferença*
classroom000	-0.1673	0.1502	-0.3175
classroom001	0.4135	0.0757	0.3379
classroom002	0.3020	0.1549	0.1472
classroom003	0.3555	0.1173	0.2382
classroom004	0.2006	0.2995	-0.0989
classroom005	0.3320	0.1529	0.1791
classroom006	0.1950	0.1172	0.0778
classroom007	0.4722	0.1375	0.3347
classroom008	0.2772	0.1963	0.0809
classroom009	0.4405	0.0833	0.3573
classroom010	0.3223	0.1185	0.2038
classroom011	-0.0118	0.1412	-0.1531
classroom012	0.4105	0.1635	0.2470
classroom013	0.1279	0.1647	-0.0367
classroom014	0.3051	0.2234	0.0816
classroom075	0.3200	0.1518	0.1682
classroom085	0.3204	0.2085	0.1119
classroom095	0.3093	0.2005	0.1087
classroom105	0.3330	0.2082	0.1248
classroom115	0.3413	0.2187	0.1226
classroom125	0.3600	0.2008	0.1592
classroom135	0.3599	0.1953	0.1645
classroom145	0.3538	0.2001	0.1537
classroom150	0.3279	0.1642	0.1637
classroom155	0.3372	0.1878	0.1494
classroom165	0.3649	0.1529	0.2120
classroom175	0.3622	0.1640	0.1982
classroom185	0.3640	0.1748	0.1892
classroom195	0.3899	0.1646	0.2254
Média	0.3041	0.1686	0.1355

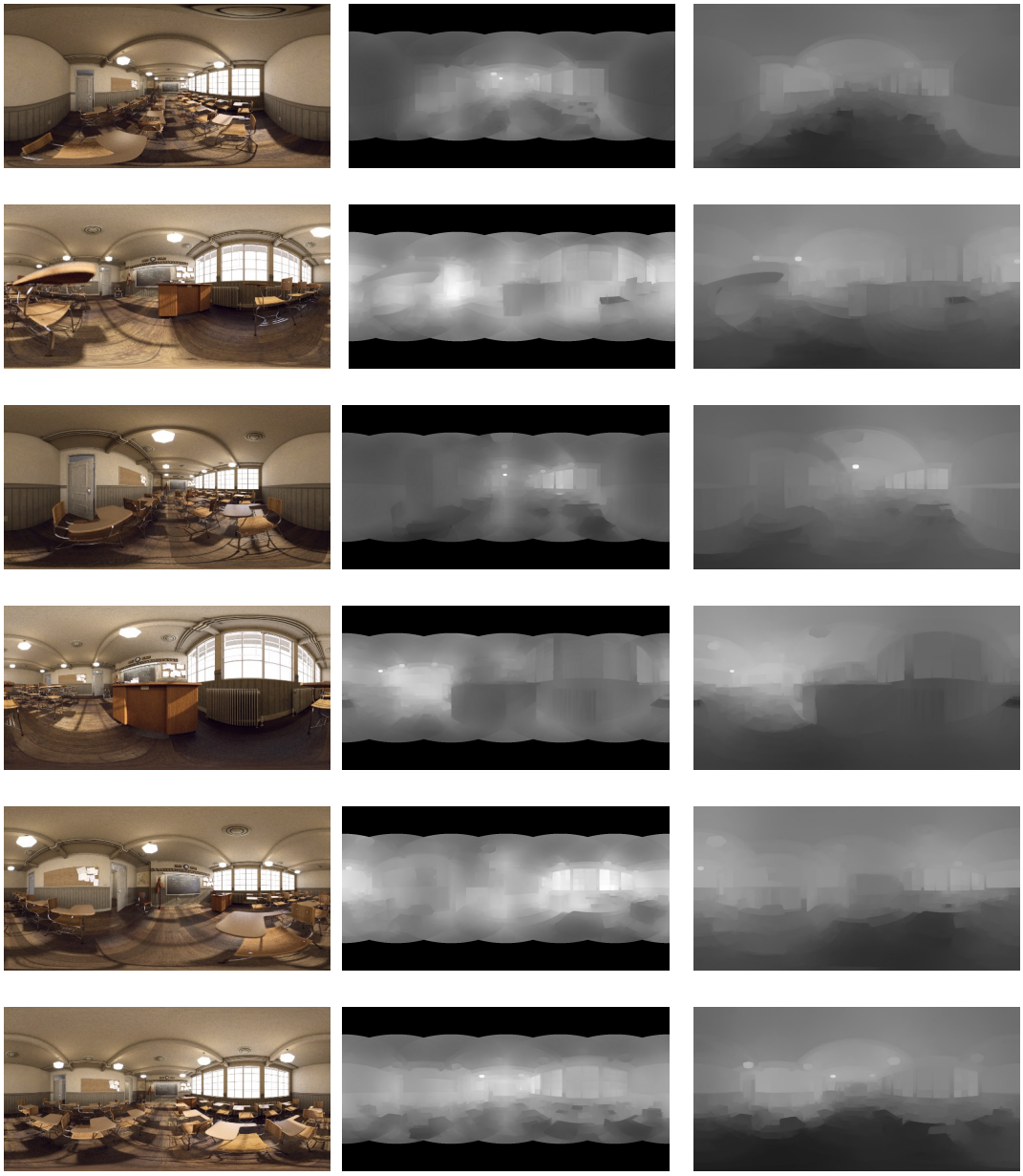
Fonte: O Autor

É possível notar que ainda não foi possível obter mapas de profundidade completamente suaves, e que a ponderação e posterior interpolação das seções sobrepostas ainda gera variações nas profundidades que formam artefatos na visualização. Analisando as imagens sintéticas na figura 5.1, é visível em alguns casos que o método proposto identifica profundidades diferentes para estruturas próximas ao solo que a aplicação direta não identifica, o que pode ser explicado pela ausência de regiões homogêneas como piso e teto do ambiente, que realmente geram estimativas visualmente piores pela rede. Também é possível ver que a maior dimensão da sala (parede mais longe da câmera fica mais clara) em alguns casos é identificada corretamente pelo método mas não pela aplicação direta.

O método proposto neste trabalho é claramente altamente dependente dos resultados da rede neural. Por exemplo, na segunda imagem sintética é possível perceber objetos identificados com profundidade visualmente errada em ambos os mapas, além de estruturas como janelas apresentarem uma ambiguidade inerente (a profundidade estimada deve ser a da janela ou do que é possível ver através desta?).

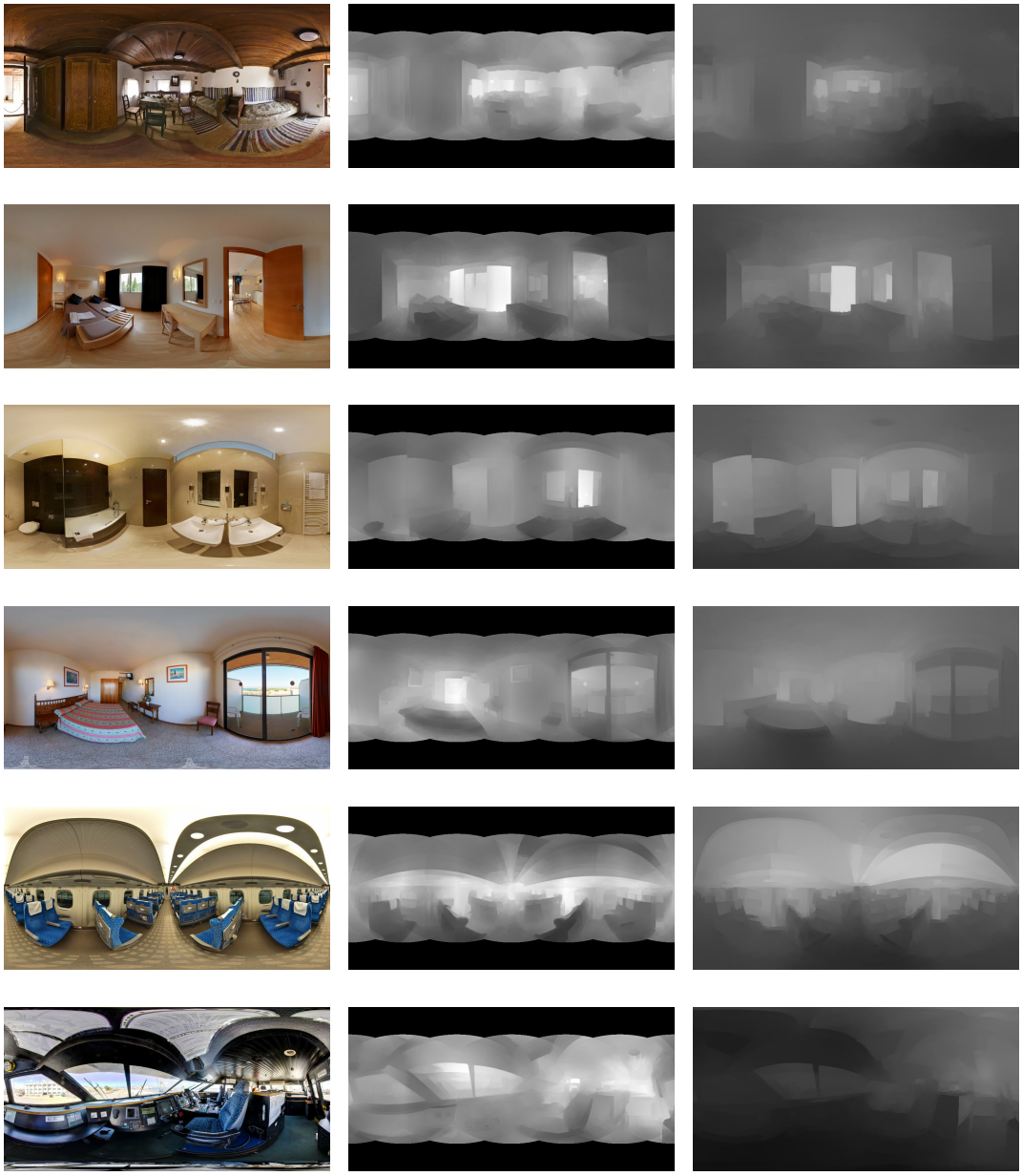
Analisando a aplicação em imagens reais na figura 5.2 é possível perceber estruturas muito mais interessantes para uma inspeção visual dos mapas de profundidade gerados. As imagens escolhidas para visualização são divididas em dois grupos, ambientes domésticos internos, e duas imagens de ambientes internos menores, no caso uma aeronave. É possível perceber as mesmas questões levantadas no parágrafo anterior, que as regiões dos polos da esfera afetam bastante a estimativa da rede, e além disso, regiões homogêneas escuras (como na segunda imagem na figura) são especialmente desafiadoras para a rede, e isso nos leva à sensibilidade do método proposto à que regiões fazem parte de cada seção, por exemplo, se em uma seção da esfera não há informação contextual ao redor de uma região homogênea, e na seção adjacente há, haverá uma diferença muito grande entre as profundidades estimadas. Percebe-se, como nas imagens sintéticas que nas cenas domésticas internas há estruturas que parecem ter suas profundidades estimadas melhor pelo método proposto, por não se confundirem com regiões homogêneas como piso e teto, e também que em vários casos a escal global da cena parece melhor estimada (superfícies mais longe com cor mais clara no mapa.), ainda que com os artefatos gerados pela interpolação. Na segunda metade do conjunto, as cenas da aeronave, a inspeção visual do método proposto permite observar profundidades significativamente mais coerentes com as esperadas do que a aplicação direta da rede, por serem casos onde a distorção introduzida pela projeção equiretangular é muito mais significativa, o que demonstra o valor da técnica de projeção da esfera para casos como estes.

Figura 5.1: Comparação visual dos mapas de profundidade obtidos de imagens sintéticas



Esquerda: Imagem de entrada, centro: método proposto, direita: CNN direto na imagem equiretangular. Fonte: O Autor

Figura 5.2: Comparação visual dos mapas de profundidade obtidos de imagens reais



Esquerda: Imagem de entrada, centro: método proposto, direita: CNN direto na imagem equiretangular. Fonte: O Autor

Figura 5.3: Visualização de nuvem de pontos das profundidades estimadas

1. Imagem original 2. Mapa de profundidade estimado 3. Visualização da nuvem de pontos de parte da cena 4. Visualização da nuvem de pontos de outra parte da cena 5. Nuvem de pontos da cena inteira vista de fora da esfera. Fonte: O Autor

A figura 5.3 mostra uma visualização possível dos mapas de profundidades obtidos, que é gerar uma nuvem de pontos no espaço 3D, onde cada pixel da imagem esférica é associado à sua profundidade em relação ao centro da esfera, que é considerada a coordenada $(0, 0, 0)$ no espaço tridimensional. Pode-se observar nas seções da esfera apresentadas que, nesta cena, estruturas como camas e outros móveis são reconhecíveis na reconstrução tridimensional, e ainda que as extremidades da cena como paredes e teto estejam distorcidas, a estrutura geral da cena é estimada razoavelmente.

Todos os testes realizados utilizaram a técnica de estimativa de profundidade de Liu, Shen e Lin (2015), e a implementação utiliza scripts em MATLAB e Python para cada uma das etapas do método. Scripts MATLAB para projeção da esfera para o plano e vice-versa foram adaptados dos disponibilizados por Xiao et al. (2012), a estimativa de profundidade utiliza a rede já treinada disponibilizada por Liu, Shen e Lin (2015), e para as etapas de ponderação e reconstrução do mapa, além de execução do método completo, foram desenvolvidos scripts em Python. O tempo de execução do método em uma CPU Intel Corei7-6700, 8GB de memória RAM e no sistema operacional Windows 10 foi de,

em média, 3 minutos para cada imagem, e todas as etapas do método.

6 CONCLUSÕES

Neste trabalho foi proposta uma técnica para a estimativa de profundidades em imagens esféricas utilizando apenas uma única imagem. Utilizando métodos projetados para imagens planares já existentes, sem modificação, o método proposto consiste em fazer divisões da esfera e projetar as imagens para o plano, estimar as profundidades no plano, e converter as estimativas de volta para a esfera. Uma estratégia de ponderação das profundidades obtidas de cada plano foi desenvolvida para atenuar descontinuidades na construção do mapa de profundidades esférico.

Inspecionando os resultados obtidos, é possível concluir que há potencial no método proposto, visto que é possível observar algumas melhorias relativas à aplicação direta de técnicas de estimativa de profundidade existentes à imagens esféricas, tanto em uma análise quantitativa quanto qualitativa. No entanto, a qualidade dos mapas de profundidade obtidos ainda deixa muito a desejar, o que pode ser atribuído à incapacidade do método em sua forma atual de forçar completamente a coêrencia entre profundidades obtidas das seções sobrepostas da esfera. A ponderação das seções linha a linha minimiza a soma das diferenças nas disparidades ao longo da linha, o que pode gerar resultados insatisfatórios para linhas onde há muita variação. Além disso, a alta sensibilidade do método à rotação da esfera é outra questão a ser resolvida, pois se no seccionamento um plano é muito homogêneo, as estimativas ruins geradas pela falta de informação contextual ainda são propagadas pela ponderação. A utilização de grandes regiões de sobreposição podem mitigar isto, mas ainda assim o resultado pode ser muito diferente se for feito uma rotação da esfera antes de seccionar.

No futuro pretende-se aprimorar a etapa de ponderação, inclusive foi testada uma formulação similar à da seção 4.5, porém com um peso por pixel na região de sobreposição. Entretanto, a complexidade polinomial da solução por SVD tornou a solução imprática para imagens com as resoluções utilizadas neste trabalho. É possível fazer um *downscaling* das profundidades para se ter um peso a cada N pixels, sendo N o fator de escala, mas a ponderação com esta formulação gerou resultados menos satisfatórios que a formulação por linhas devido à presença de maiores artefatos. Portanto, será estudado como aprimorar a ponderação a cada conjunto de pixels, ou ainda propor uma abordagem utilizando superpixels, de modo a ponderar regiões homogêneas com o mesmo peso, mas visando manter as relações de profundidade entre cada região na estimativa. Em relação à sensibilidade à rotação, uma melhoria pode ser aplicar o método a várias rotações da

esfera original, e escolher a melhor entre estas, ou combinar as estimativas obtidas em um único mapa de profundidades.

Um método alternativo que pode ser explorado seria uma adaptação da rede neural proposta em Su e Grauman (2017), elaborar um modelo que traduza o conhecimento das redes neurais para estimativa em imagens planares para o domínio da esfera. Ou ainda pode-se buscar o treinamento das técnicas utilizadas utilizando imagens esféricas, porém a ausência de bases de dados de imagens equiretangulares com informação real de profundidade comprometem tal alternativa.

Por fim, conclui-se que a estimativa de profundidade a partir de uma única imagem esférica é um assunto relativamente pouco abordado, que há muito espaço para trabalhos futuros. O método proposto, ainda que com resultados não muito satisfatórios, abre várias possibilidades de aprimoramento, além de demonstrar a viabilidade da aplicação de técnicas de visão computacional para imagens planares em imagens esféricas através da abordagem de seção da esfera em planos. No entanto, garantir a coerência quando agregados os resultados de cada plano ainda é um desafio.

REFERÊNCIAS

ABRAMS, A.; HAWLEY, C.; PLESS, R. Heliometric stereo: Shape from sun position. In: ____. **ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part II**. Springer Berlin Heidelberg, 2012. p. 357–370. ISBN 978-3-642-33709-3. Disponível em: <https://doi.org/10.1007/978-3-642-33709-3_26>.

EIGEN, D.; FERGUS, R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: **The IEEE International Conference on Computer Vision (ICCV)**. [S.l.: s.n.], 2015.

EIGEN, D.; PUHRSCH, C.; FERGUS, R. Depth map prediction from a single image using a multi-scale deep network. In: **Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2**. Cambridge, MA, USA: MIT Press, 2014. (NIPS'14), p. 2366–2374. Disponível em: <http://dl.acm.org/citation.cfm?id=2969033.2969091>.

FURUKAWA, Y.; HERNáNDEZ, C. Multi-view stereo: A tutorial. **Foundations and Trends® in Computer Graphics and Vision**, v. 9, n. 1-2, p. 1–148, 2015. ISSN 1572-2740. Disponível em: <http://dx.doi.org/10.1561/0600000052>.

GODARD, C.; Mac Aodha, O.; BROSTOW, G. J. Unsupervised monocular depth estimation with left-right consistency. **CoRR**, abs/1609.03677, 2016. Disponível em: <http://arxiv.org/abs/1609.03677>.

GONZALEZ, R. C.; WOODS, R. E. **Digital Image Processing (3rd Edition)**. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 2006. ISBN 013168728X.

HAMZAH, R. A.; IBRAHIM, H. Literature survey on stereo vision disparity map algorithms. In: **Journal of Sensors, vol. 2016**. [S.l.: s.n.], 2016. v. 2016.

KOYASU, H.; MIURA, J.; SHIRAI, Y. Real-time omnidirectional stereo for obstacle detection and tracking in dynamic environments. In: **Proceedings 2001 IEEE/RSJ International Conference on Intelligent Robots and Systems. Expanding the Societal Role of Robotics in the the Next Millennium (Cat. No.01CH37180)**. [S.l.: s.n.], 2001. v. 1, p. 31–36 vol.1.

KUZNIETSOV, Y.; STÜCKLER, J.; LEIBE, B. Semi-supervised deep learning for monocular depth map prediction. **CoRR**, abs/1702.02706, 2017. Disponível em: <http://arxiv.org/abs/1702.02706>.

LI, Y.; TANG, C.-K.; SHUM, H.-Y. Efficient dense depth estimation from dense multiperspective panoramas. In: IEEE INTERNATIONAL CONFERENCE ON COMPUTER VISION, 2001, Vancouver, BC, Canada. **Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on**. [S.l.]: IEEE, 2001.

LIU, B.; GOULD, S.; KOLLER, D. Single image depth estimation from predicted semantic labels. In: **2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2010. p. 1253–1260. ISSN 1063-6919.

LIU, F.; SHEN, C.; LIN, G. Deep convolutional neural fields for depth estimation from a single image. In: **The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [s.n.], 2015. Disponível em: <http://arxiv.org/abs/1411.6387>.

MANN, S.; PICARD, R. W. Virtual bellows: constructing high quality stills from video. In: **Proceedings of 1st International Conference on Image Processing**. [S.l.: s.n.], 1994. v. 1, p. 363–367 vol.1.

NAYAR, S. K. Catadioptric omnidirectional camera. In: **Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 1997. p. 482–488. ISSN 1063-6919.

ONOE, Y. et al. Telepresence by real-time view-dependent image generation from omnidirectional video streams. **Comput. Vis. Image Underst.**, Elsevier Science Inc., New York, NY, USA, v. 71, n. 2, p. 154–165, ago. 1998. ISSN 1077-3142. Disponível em: <http://dx.doi.org/10.1006/cviu.1998.0705>.

ORGHIDAN, R.; MOUADDIB, E. M.; SALVI, J. Omnidirectional depth computation from a single image. In: **Proceedings of the 2005 IEEE International Conference on Robotics and Automation**. [S.l.: s.n.], 2005. p. 1222–1227. ISSN 1050-4729.

PELEG, S.; HERMAN, J. Panoramic mosaics by manifold projection. In: **Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 1997. p. 338–343. ISSN 1063-6919.

PENTLAND, A. P. A new sense for depth of field. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, PAMI-9, n. 4, p. 523–531, July 1987. ISSN 0162-8828.

RANFTL, R. et al. Dense monocular depth estimation in complex dynamic scenes. In: **2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.: s.n.], 2016. p. 4058–4066.

SAXENA, A.; CHUNG, S. H.; NG, A. Y. Learning depth from single monocular images. In: **Proceedings of the 18th International Conference on Neural Information Processing Systems**. Cambridge, MA, USA: MIT Press, 2005. (NIPS'05), p. 1161–1168. Disponível em: <http://dl.acm.org/citation.cfm?id=2976248.2976394>.

SAXENA, A.; CHUNG, S. H.; NG, A. Y. 3-d depth reconstruction from a single still image. **International Journal of Computer Vision**, v. 76, n. 1, p. 53–69, Jan 2008. ISSN 1573-1405. Disponível em: <https://doi.org/10.1007/s11263-007-0071-y>.

SCHARSTEIN, D.; SZELISKI, R.; ZABIH, R. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. In: **Proceedings IEEE Workshop on Stereo and Multi-Baseline Vision (SMBV 2001)**. [S.l.: s.n.], 2001. p. 131–140.

SILVEIRA, T. L. T. D.; JUNG, C. R. Evaluation of keypoint extraction and matching for pose estimation using pairs of spherical images. In: **2017 30th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)**. [S.l.: s.n.], 2017. p. 374–381.

SU, Y.; GRAUMAN, K. Flat2sphere: Learning spherical convolution for fast features from 360°imagery. **CoRR**, abs/1708.00919, 2017. Disponível em: <http://arxiv.org/abs/1708.00919>.

SZELISKI, R. Video mosaics for virtual environments. **IEEE Computer Graphics and Applications**, v. 16, n. 2, p. 22–30, Mar 1996. ISSN 0272-1716.

SZELISKI, R. **Computer Vision: Algorithms and Applications**. London: Springer-Verlag London Limited, 2011.

WEI, Y.; WU, C. Fast depth reconstruction with a defocus model on micro scale. In: **2015 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER)**. [S.l.: s.n.], 2015. p. 947–952.

WEISSTEIN, E. W. **Cylindrical Equidistant Projection. From MathWorld—A Wolfram Web Resource**. 2018. Acesso em Janeiro de 2018. Disponível em: <http://mathworld.wolfram.com/CylindricalEquidistantProjection.html>.

WEISSTEIN, E. W. **Map Projection. From MathWorld—A Wolfram Web Resource**. 2018. Acesso em Janeiro de 2018. Disponível em: <http://mathworld.wolfram.com/MapProjection.html>.

WOODHAM, R. J. Shape from shading. In: HORN, B. K. P.; BROOKS, M. J. (Ed.). Cambridge, MA, USA: MIT Press, 1989. cap. Photometric Method for Determining Surface Orientation from Multiple Images, p. 513–531. ISBN 0-262-08183-0. Disponível em: <http://dl.acm.org/citation.cfm?id=93871.93888>.

XIAO, J. et al. Recognizing scene viewpoint using panoramic place representation. In: **2012 IEEE Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2012. p. 2695–2702. ISSN 1063-6919.

ZHAO, F.; HUANG, Q.; GAO, W. Image matching by normalized cross-correlation. In: **2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings**. [S.l.: s.n.], 2006. v. 2, p. II–II. ISSN 1520-6149.

ZHOU, T. et al. Unsupervised learning of depth and ego-motion from video. **CoRR**, abs/1704.07813, 2017. Disponível em: <http://arxiv.org/abs/1704.07813>.

ZHU, Z. Omnidirectional stereo vision. In: **Proc. of ICAR'01**. [S.l.: s.n.], 2001. p. 22–25.