Capítulo 2

Processamento de Imagens Omnidirecionais e Aplicações

Thiago L. T. da Silveira e Cláudio R. Jung

Programa de Pós-Graduação em Computação (PPGC) - UFRGS

Abstract

Omnidirectional images and videos have been widely disseminated due to the popularization of devices for capture and visualization. Unlike images captured with perspective projection, omnidirectional media are defined on the surface of a sphere, having a field of view of $360^\circ \times 180^\circ$. *Thus, they store the light intensities in the entire region around the capture point, with high potential for use in applications involving immersive augmented, mixed and virtual reality experiences. Although defined in the spherical domain, omnidirectional images are often mapped to a (multi)planar representation, which results in distorted images and degrades the performance of most traditional visual computing algorithms designed to work in the plane. This chapter reviews the spherical camera model, the most common capture devices, and popular (multi)planar representations of omnidirectional media. It also lists the main challenges of omnidirectional visual computing, focusing on the deep learning paradigm, and discusses potential applications.*

Resumo

Imagens e vídeos omnidirecionais têm sido amplamente difundidos devido à popularização de dispositivos para captura e visualização. Ao contrário das imagens capturadas com projeção em perspectiva, as mídias omnidirecionais são definidas sobre a superfície de uma esfera, tendo um campo de visão de $360^\circ \times 180^\circ$. *Assim, elas armazenam as intensidades de luz em toda região em torno do ponto de captura, com alto potencial de uso em aplicações que envolvem experiências imersivas de realidade aumentada, mista e virtual. Embora definidas no domínio esférico, as imagens omnidirecionais muitas vezes são mapeadas para uma representação (multi) planar, o que resulta em imagens distorcidas*

Vídeo com a apresentação do capítulo: `https://youtu.be/rqLWrSRm-Y0`

e degrada o desempenho da maioria dos algoritmos tradicionais de computação visual projetados para funcionar no plano. Este capítulo revisa o modelo de câmera esférica, os dispositivos de captura mais comuns e as representações (multi) planares populares de mídias omnidirecionais. Ele também elenca os principais desafios da computação visual omnidirecional, com foco no paradigma de aprendizado profundo, e aborda potenciais aplicações.

2.1. Introdução

Imagens omnidirecionais – também conhecidas como imagens esféricas, panorâmicas ou em 360° – são populares nos dias de hoje graças à acessibilidade e portabilidade dos dispositivos de captura lançados nos últimos anos [J. Huang et al. 2017, da Silveira and Jung 2019b]. Imagens e vídeos em 360° aproximam-se do modelo de imagem ideal chamado modelo de imagem plenóptica, onde toda a informação visual da cena é capturada a partir de todos os pontos de vista possíveis ao longo do tempo [Ebrahimi et al. 2016]. Além das aplicações clássicas, as mídias esféricas ajudam a proporcionar experiências imersivas ao usuário em novas aplicações de realidade aumentada, mista e virtual (AR/MR/VR) quando visualizadas em dispositivos de visualização montados na cabeça (HMDs) [Serrano et al. 2019]. Em particular, a edição de panoramas permite a manipulação de imagens e vídeos, o que pode melhorar a experiência do usuário [Zhang et al. 2022b, Zhang et al. 2021]. Ao contrário das imagens regulares baseadas no modelo de projeção em perspectiva, definidas em um plano, as imagens omnidirecionais são definidas na superfície da esfera unitária [Li 2008, Fujiki et al. 2007]. As imagens esféricas têm um campo de visão (FoV, ou *Field-of-View*) de $360° \times 180°$ [da Silveira and Jung 2019b] que captura as intensidades de luz de toda a cena.

Para exemplificar as diferenças visuais entre imagens em perspectiva e panorâmicas, considere a Figura 2.1. Mais precisamente, a Figura 2.1(a) ilustra uma captura usando uma câmera em perspectiva com FoV limitado, enquanto a Figura 2.1(a) mostra uma imagem esférica capturada do mesmo ponto de vista[1]. Percebe-se claramente que a topologia de ambas imagens são bastante distintas.

Embora as imagens panorâmicas sejam definidas no domínio esférico, elas são comumente representadas em formato planar ou multi-planar [Yang et al. 2018, Zelnik-Manor et al. 2005, da Silveira et al. 2022]. Muitas funções de mapeamento da esfera para um ou mais planos podem ser usadas para gerar a representação planar, mas todas elas introduzem distorções [Su and Grauman 2017, Azevedo et al. 2020]. Um panorama pode ser representado no plano (como em formato de "mapa-múndi"), mas o algoritmo de computação visual que o utiliza como entrada ainda precisa considerar as deformações introduzidas para ser preciso em sua tarefa [Cruz-Mota et al. 2012, da Silveira et al. 2021].

Comparado com a evolução dos algoritmos projetados para imagens em perspectiva, a computação visual *omnidirecional* ainda está em estágio embrionário, e apenas alguns problemas clássicos são abordados sob esta ótica renovada. Este artigo lança luz sobre como se pode esperar que a computação visual omnidirecional difira da tradicional

[1]As imagens foram geradas artificalmente, a partir do modelo *Classroom* 3D, disponível sob licença CC0 license em `https://www.blender.org`.

(a) (b)

Figura 2.1: Duas capturas da mesma cena 3D com poses idênticas, mas com câmeras diferentes. A primeira captura foi feita por uma (a) câmera em perspectiva de FoV estreito, e a segunda (b) veio de uma câmera 360°, ilustrada no plano por projeção ortográfica.

e quais esforços podem ser empregados para lidar com essas discrepâncias. Ele traz uma discussão aprofundada sobre o modelo de câmera esférica, *pipelines* padrão de aquisição de imagens em 360° e representações planares ou multi-planares. Além disso, o artigo ataca deficiências do uso de redes convolucionais em panoramas, e apresenta alternativas recentes.

2.2. Uma visão geral sobre imagens 360°

Esta seção abrange aspectos técnicos envolvendo a base matemática do imageamento omnidirecional, sistemas comuns para aquisição de imagem 360° e representações planares padrão de panoramas. Mais detalhes sobre criação de conteúdo omnidirecional podem ser encontrados em pesquisas como [Wang et al. 2020b].

2.2.1. O modelo de câmera esférica

Uma câmera *pinhole* é modelada por projeções centrais e em perspectiva, onde um raio vem de um ponto tridimensional (3D) em coordenadas de mundo, passa por seu centro de projeção e toca o plano da imagem [Hartley and Zisserman 2003]. As particularidades do mapeamento 3D–2D subjacente – como a cobertura da cena na imagem resultante – dependem da matriz da câmera, que combina parâmetros intrínsecos e extrínsecos [Hartley and Zisserman 2003].

Por outro lado, o modelo de câmera esférica deriva das projeções central e esférica [S. Li and Fukumori 2005]. Abstrai-se a câmera como uma *esfera unitária* [da Silveira and Jung 2019b] localizada e orientada no espaço. Uma vez que uma câmera omnidirecional cobre todo o FoV, todos os pontos do mundo 3D ao redor da câmera são capturados em uma única projeção esférica [Akihiko et al. 2005], com exceção de regiões com oclusões. O modelo de câmera esférica não considera parâmetros intrínsecos e assume que a câmera é totalmente representada por seus seis graus de liberdade (6 DoF, ou *Degrees of Freedom*) extrínsecos [Guan and Smith 2017, Krolla et al. 2014].

Primeiro precisamos definir um sistema de coordenadas de mundo (3D) para entender o modelo de câmera esférica. Dado esse sistema de coordenadas, podemos centralizar a câmera em uma dada posição $\mathbf{C} \in \mathbb{R}^3$ e orientá-la usando uma dada matriz de

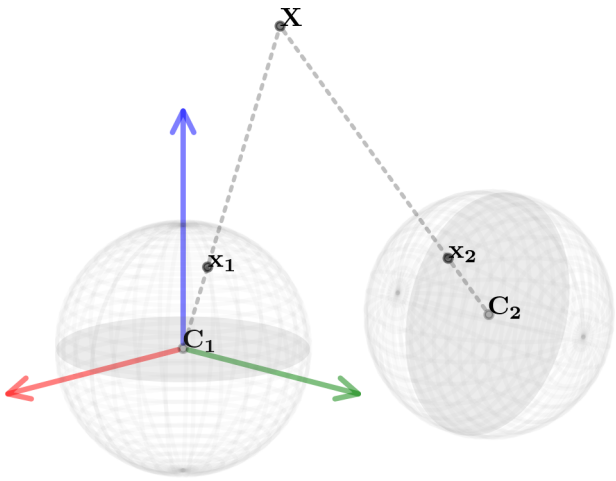

Figura 2.2: Projeção de um ponto 3D $\mathbf{X}$ em duas câmeras esféricas com diferentes parâmetros. As câmeras são descritas pelos extrínsecos$[\mathbf{R}_1 = \mathbf{I}|\mathbf{t}_1 = -\mathbf{R}_1\mathbf{C}_1 = \mathbf{0}]$ e $[\mathbf{R}_2 \neq \mathbf{I}|\mathbf{t}_2 = -\mathbf{R}_2\mathbf{C}_2 \neq \mathbf{0}]$.

rotação $\mathbf{R} \in SO(3)$. Assim, podemos caracterizar a câmera através de seus parâmetros extrínsecos $[\mathbf{R}|\mathbf{t}]$, onde $\mathbf{t} = -\mathbf{RC} \in \mathbb{R}^3$ é chamado de "vetor de translação" [da Silveira et al. 2022].

Um ponto 3D $\mathbf{X} \in \mathbb{R}^3$ em coordenadas de mundo, parametrizado de acordo com o sistema de coordenadas definido, é então projetado na câmera definida por $[\mathbf{R}|\mathbf{t}]$ usando [Akihiko et al. 2005]

$$\mathbf{x} = \frac{\mathbf{RX} + \mathbf{t}}{\|\mathbf{RX} + \mathbf{t}\|_2}, \tag{1}$$

onde $\|\cdot\|_2$ é a norma ℓ_2. Observa-se que o ponto $\mathbf{x}$ resultante da imagem está na superfície de uma esfera unitária, *i.e.,* $\mathbf{x} \in S^2 \subset \mathbb{R}^3$ [S. Li and Fukumori 2005].

A Figura 2.2 ilustra um ponto tridimensional $\mathbf{X}$ do mundo projetado em duas câmeras esféricas distintas. Uma das câmeras está posicionada na origem e alinhada ao sistema de coordenadas pré-definido, tendo extrínsecos $[\mathbf{R}_1 = \mathbf{I}|\mathbf{t}_1 = -\mathbf{R}_1\mathbf{C}_1 = \mathbf{0}]$. A outra câmera não está na origem e tem uma orientação diferente, com extrínsecos $[\mathbf{R}_2 \neq \mathbf{I}|\mathbf{t}_2 = -\mathbf{R}_2\mathbf{C}_2 \neq \mathbf{0}]$. Note que os pontos de imagem $\mathbf{x}_1$ e $\mathbf{x}_2$ são descritos em coordenadas locais da imagem, ou seja, em relação a cada câmera. Sozinhos, eles não codificam informações explícitas sobre as posições originais das câmeras no sistema de coordenadas pré-definido.

2.2.2. Aquisição de imagens esféricas

As estratégias existentes para aquisição de imagens e vídeos omnidirecionais envolvem o uso de uma ou mais câmeras regulares, potencialmente equipadas com componentes ópticos especiais [da Silveira et al. 2022]. Na verdade, ao contrário do que sugere o modelo de imagem esférica, não há um dispositivo de sensor único para capturar todas as informações da cena de uma só vez [Adarve and Mahony 2017]. Abaixo, é apresentada uma breve revisão dos três principais sistemas de captura: catadióptrico, polidióptrico e dispositivos de imagem de 360° com duas lentes olho de peixe.

(a) (b) (c)

Figura 2.3: Versões planificadas da imagem esférica ilustrada na Figura 2.1b: (a) Formato ERP, (b) CMP and (c) representação icosahedral.

Sistemas de imagem catadióptricos combinam uma câmera regular com um espelho de formato convexo (cônico, esférico, parabólico ou hiperbólico) e permitem capturar informações visuais de todo o campo de visão *horizontal* de uma cena [Nayar 1997]. Esse método sofre de auto-oclusão do sensor/espelho e normalmente gera imagens representadas em formato cilíndrico [Cruz-Mota et al. 2012]. Devido ao campo de visão vertical restrito e aos componentes frágeis do espelho, os dispositivos catadióptricos são raros em pesquisas e aplicações industriais recentes [Aggarwal et al. 2016].

Por outro lado, dispositivos de captura polidióptricos consistem em um número (variável) de câmeras regulares apontando para fora em uma estrutura (*rig*). Cada câmera captura uma parte estreita da cena (ou seja, tem um campo de visão estreito), e todas as visualizações são combinadas em um procedimento baseado em *software* chamado costura de imagens (*stitching* ou *mosaicking*) [Im et al. 2016]. Dispositivos de imageamento polidióptricos costumam ser volumosos e caros, mas podem produzir panoramas de alta resolução com campo de visão personalizado [Fangi et al. 2018].

Um tipo mais recente de dispositivo de captura, suportado por muitos fabricantes, combina dois sensores localizados em posições opostas equipados com lentes olho de peixe [Shan and Li 2018]. Cada sensor captura uma imagem "hemisférica" (com campo de visão ultra-grande) adequada para a costura esférica de duas visualizações [Lo et al. 2018]. Esses dispositivos portáteis e baratos simplificaram e democratizaram a aquisição de conteúdo real em 360° e impulsionaram a indústria e a pesquisa em áreas relacionadas à realidade aumentada/mista/virtual [Jung et al. 2019, da Silveira et al. 2022].

2.2.3. Representação de mídia esférica

Imagens e vídeos esféricos podem ser representados como um mapeamento na esfera unitária [Akihiko et al. 2005]. A imagem digital subjacente é obtida por meio de um procedimento de amostragem, que é normalmente realizado após a aplicação de uma ou mais funções de mapeamento de esfera para plano.

Uma função popular de mapeamento de esfera para plano é a chamada projeção equirretangular (ERP, abreviada de *EquiRectangular Projection*). A ERP - também conhecida como mapeamento latitude-longitude [Gava et al. 2018] - é considerada a representação planar padrão da esfera [Eder et al. 2019, da Silveira and Jung 2019b], e permite um relação simples entre *pixels* no plano e pontos amostrados na esfera.

Uma vez que um determinado ponto **x** está sobre a superfície de uma esfera unitária, ele pode ser reescrito em coordenadas esféricas usando dois parâmetros angulares (θ, ϕ) [Akihiko et al. 2005]:

$$\mathbf{x} = [x\ y\ z]^\top = [\cos(\theta)\sin(\phi)\ \sin(\theta)\sin(\phi)\ \cos(\phi)]^\top, \tag{2}$$

onde $\theta \in [0, 2\pi)$ e $\phi \in [0, \pi)$.

Além disso, cabe lembrar que uma câmera omnidirecional captura todo o conteúdo da cena e, portanto, há informações associadas a cada posição (θ, ϕ) na superfície esférica. Como tal, uma imagem omnidirecional pode ser representada em um plano de $[0, 2\pi) \times [0, \pi)$. A *imagem ERP* é, então, gerada a partir de uma discretização em (θ, ϕ), de modo que a intensidade de luz associada a um ponto de imagem **x** mapeia para a posição do pixel **p** dada por

$$\mathbf{p} = [u\ v]^\top = \left[\left\lfloor \frac{\theta w}{2\pi} \right\rfloor \ \left\lfloor \frac{\phi h}{\pi} \right\rfloor \right]^\top, \tag{3}$$

onde w e h são a largura e altura da imagem ERP em *pixels*, respectivamente. As imagens ERP frequentemente têm uma razão de aspecto 2:1, o que significa que a variação angular em θ e ϕ é a mesma.

Os parâmetros θ e ϕ são recuperados de **x** usando a relação inversa apresentada na Eq. (2):

$$\theta = \tan^{-1}(y, x) \tag{4}$$

e

$$\phi = \cos^{-1}(z), \tag{5}$$

onde $\tan^{-1}(\cdot, \cdot)$ representa a função arco-tangente sensível a quadrante.

A ERP é direta e simples de calcular, mas possui uma amostragem não-uniforme que distorce os objetos da cena dependendo de sua localização na imagem [Cruz-Mota et al. 2012], que se torna mais intensa próximo aos polos norte e sul [Ferreira et al. 2017]. Muitos outros mapeamentos de esfera para plano podem ser considerados, mas nenhum é livre de distorção [Zelnik-Manor et al. 2005, Su and Grauman 2017]. Como essas deformações dependem da magnitude do FoV usado na projeção [da Silveira et al. 2018], alguns autores propõem mapear a esfera em não apenas um, mas vários planos. Por exemplo, mapear a esfera em um cubo circunscrito resulta em seis imagens com FoV mais estreito e faces equi-angulares, conhecidas como mapeamento em cubo (CMP, de *Cube Map Projection*) [Dai et al. 2019, da Silveira et al. 2018]. A CMP reduz as distorções, mas o FoV de 90° de cada face ainda é maior do que o comumente encontrado em imagens de perspectiva [Su and Grauman 2017, Wang et al. 2018a]. Além disso, a conectividade das faces deve ser considerada ao processar *imagens CMP*. Representações emergentes baseadas em divisões sucessivas de uma forma geométrica 3D tentam mitigar ainda mais as distorções. Abordagens proeminentes incluem a icosfera/projeção em planos tangentes [Eder et al. 2020], que deriva de um icosaedro, e aquelas baseadas em um octaedro [Lee et al. 2020].

A Figura 2.3 ilustra o mapeamento da imagem omnidirecional da Figura 2.1b para suas representações ERP, mapa de cubo e baseadas em icosaedro (desdobrados). Deve-se mencionar que a troca de formatos de representação pode levar à perda de informações e introduzir artefatos [Azevedo et al. 2020], uma vez que esses mapeamentos requerem transformações *subpixel* [Coors et al. 2018].

Como uma nota final sobre a representação de imagens, artefatos de compressão também estão presentes em conteúdos esféricos armazenados digitalmente e podem gerar degradações visuais adicionais. O leitor é encaminhado ao [Xu et al. 2020] para uma análise detalhada sobre compressão de imagens esféricas e avaliação de qualidade.

2.3. Imagens esféricas a aprendizado profundo

Seguindo a tendência de aplicações que exploram imagens em perspectiva, técnicas de aprendizado profundo também têm sido exploradas para processar mídias esféricas. Entretanto, a amostragem não-uniforme gerada pelo processo de planificação exige cuidado quando se deseja aproveitar arquiteturas planejadas para imagens em perspectiva.

Este capítulo foca a análise no formato ERP por ser a representação planar padrão de imagens em 360°, amplamente empregada na indústria e na pesquisa [Su and Grauman 2017], mas considerações serão feitas sobre outras representações. É importante observar que, embora tais representações possam ser obtidas diretamente por amostragem de um domínio esférico, a maioria dos panoramas é armazenada no formato ERP. Assim, os mapeamentos multi-planares envolvem a interpolação do ERP para a esfera e, em seguida, a amostragem para a representação desejada.

2.3.1. Desafios com a representação ERP

Como discutido na Seção 2.2.3, a ERP amostra a esfera unitária de maneira *não-uniforme*. Esse procedimento resulta em um "efeito de alongamento" que se acentua nas proximidades dos polos. De fato, os polos são superamostrados, pois as primeiras e últimas linhas da imagem colapsam nos polos norte e sul da esfera, respectivamente (como nos paralelos com latitudes muito altas nos dois polos do globo terrestre). Portanto, essas linhas replicam as informações em todas as colunas. Em geral, o espaçamento entre pontos adjacentes ao longo de uma linha no formato ERP é proporcional a $\sin\phi$ [De Simone et al. 2017], resultando em um desequilíbrio acentuado entre a linha do equador ($\phi = \frac{\pi}{2}$) e os polos ($\phi = 0$ e $\phi = \pi$).

Além das distorções induzidas pela amostragem não-uniforme, as imagens ERP têm uma propriedade *cíclica* (ou *circular*) [da Silveira et al. 2022, Lee et al. 2020], o que significa que as bordas esquerda e direita se conectam. Portanto, os objetos podem ser divididos nas porções esquerda e direita de uma imagem ERP. A Figura 2.3a apresenta uma ilustração das questões mencionadas acima.

O amplo uso de imagens ERP decorre de sua simplicidade e porque elas contêm todas as informações da cena em um único plano. Explorar todo o contexto da cena a partir de uma única imagem com domínio retangular é muito atraente, especialmente na era do aprendizado profundo e o uso generalizado de redes neurais convolucionais (CNNs, de *Convolutional Neural Networks*) [Goodfellow et al. 2016]. A ideia central de uma

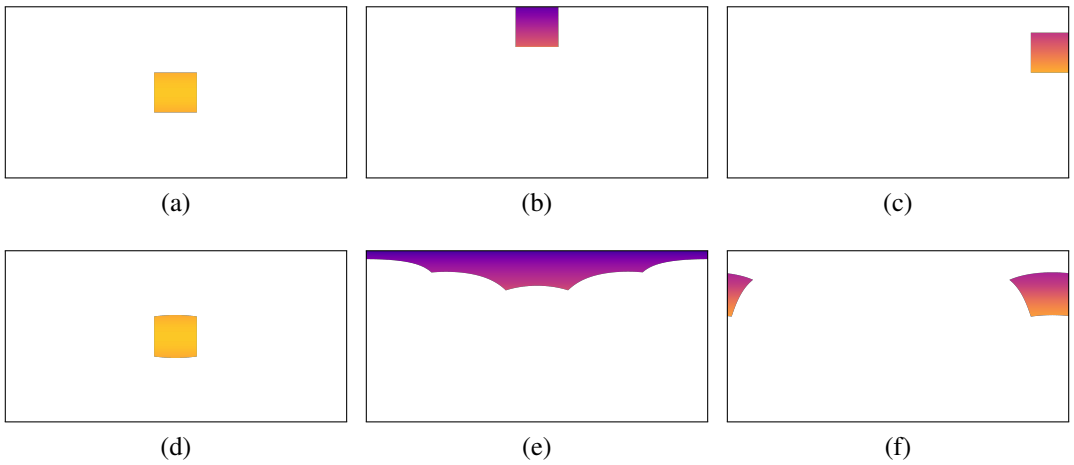

Figura 2.4: Suporte de um *kernel* (amplificado para fins de visualização) posicionado em posições diferentes de uma imagem ERP: (a)–(c) *kernels* retangulares, e (d)–(f) *kernels* ideais.

camada convolucional é que ela contém filtros com suporte espacialmente invariante – seu campo receptivo – e pesos [Goodfellow et al. 2016]. Os *kernels* convolucionais padrão são retangulares ou, mais comumente, quadrados, e são aplicados em toda a imagem usando um mecanismo de janela deslizante [Goodfellow et al. 2016]. Devido às distorções relacionadas à amostragem, aplicar esses filtros regulares a uma imagem ERP faz com que regiões desiguais da superfície da esfera sejam cobertas, dependendo da posição do filtro [Fernandez-Labrador et al. 2020]. Se quisermos que o suporte do *kernel* cubra a mesma área na superfície da esfera, ele deve ser ajustado dependendo da latitude ϕ da imagem em que está centrado [Su and Grauman 2017].

A Figura 2.4 ilustra a forma do suporte de um *kernel* regular e de um kernel "ideal" (maior do que o usual para fins de visualização) em diferentes partes de uma imagem ERP. A área relativa da superfície da esfera relacionada a cada pixel coberto pelo *kernel* é mostrada em um mapa de cores: cores arroxeadas representam valores pequenos e cores amareladas representam valores maiores (independentemente da resolução da imagem ERP, eles somam 4π, que é a área da superfície da esfera). Quando o centro do filtro ideal está na região do equador ($\phi = \frac{\pi}{2}$), como mostrado na Fig2.4d, seu suporte é pouco distorcido (compare com a Fig 2.4a, por exemplo). As distorções ficam aparentes à medida que movemos o filtro em direção aos polos. A Figura 2.4e ilustra o ajuste necessário para que o filtro cubra a mesma área da superfície da esfera que na Fig2.4d, quando se aproxima do polo norte ($\phi \to 0$) da imagem. Como podemos observar, a forma do filtro já não é mais retangular, com suporte mais amplo próximo ao polo (primeira linha). Por outro lado, a Figura 2.4b mostra o efeito de um kernel convolucional padrão, que é fixo e cobre uma região menor da superfície da esfera em comparação com a Fig2.4a. Note que a área em destaque na Figura 2.4b atinge o mesmo número de *pixels* que aquela na Figura 2.4a, mas a área coberta na superfície da esfera é menor. Finalmente, a Figura 2.4c retrata o que acontece quando um filtro regular toca uma borda lateral da imagem. O filtro é frequentemente aplicado após preenchimento com zeros ou extrapolação de dados em uma convolução regular. No caso esférico, um filtro ideal realiza uma convolução circular,

como mostrado na Figura 2.4f. A área da superfície da esfera coberta pelo *kernel* ideal em diferentes posições, ilustradas nas Figuras 2.4d, 2.4e e 2.4f, é fixa.

Mitigar o problema de circularidade em imagens ERP é simples. Isso pode ser implementado através de uma estratégia de preenchimento circular horizontal, que mantém a continuidade das informações espaciais ao longo dos paralelos da esfera nas bordas horizontais. Esses preenchimentos podem ser usados como etapa de pré-processamento, onde o panorama de entrada é preenchido circularmente antes de usar a CNN e, em seguida, recortado de volta ao tamanho original, conforme usado para o cálculo de fluxo óptico em [da Silveira and Jung 2019a]. Alternativamente, pode ser incorporado a uma camada convolucional circular, observando que o tamanho do preenchimento deve ser ajustado com base nas dimensões do *kernel* [Sun et al. 2019, Zioulis et al. 2021, Wang et al. 2018b, Zhuang et al. 2022].

A deformação espacial do *kernel* é mais difícil de lidar. Alguns trabalhos propõem ajustar as convoluções (e às vezes as operações de *pooling*) para lidar com as distorções induzidas pelas mapeamentos da esfera para o plano [Zioulis et al. 2018, Tateno et al. 2018, Su and Grauman 2017, Fernandez-Labrador et al. 2020]. Por exemplo, Su e Grauman [Su and Grauman 2017] propõem aprender pesos que ajustam as respostas de um filtro regular para acomodar as distorções do ERP. Convoluções sensíveis à distorção, propostas por Tateno *et al.* [Tateno et al. 2018], deformam seus campos receptivos para amostrar pontos dentro do suporte ideal (conforme discutido anteriormente). Uma ideia semelhante é explorada em [Fernandez-Labrador et al. 2020], onde convoluções deformáveis são usadas para a amostragem induzida pelo ERP. Uma desvantagem dessas abordagens é a sobrecarga computacional e a complexidade do código, uma vez que o uso de *kernels* que amostram irregularmente tende a ser mais lento do que convoluções 2D regulares que exploram diretamente o paralelismo em placas gráficas (GPUs).

Convoluções dilatadas foram introduzidas no contexto do aprendizado profundo para imagens planares por Fisher e Koltun [Yu and Koltun 2015], e são uma solução atraente para lidar com a amostragem não-uniforme de imagens ERP. De fato, elas foram exploradas por Zioulis e colegas [Zioulis et al. 2018] para ajustar o campo receptivo horizontal da convolução, dependendo da latitude, sendo maior próximo aos polos e menor próximo à linha do equador. As redes de convoluções dilatadas combinadas de forma adaptativa (ACDNets), propostas recentemente por Zhuang e colegas [Zhuang et al. 2022], consistem em aplicar um conjunto de convoluções dilatadas paralelas combinadas de forma adaptativa em relação aos canais através de pesos aprendíveis. Soluções baseadas em convoluções dilatadas são mais rápidas do que os *kernels* deformáveis. No entanto, eles não podem cobrir regiões exatamente de área igual na esfera, uma vez que cada linha requer uma deformação separada, como mostrado na Figura 2.4.

Outra estratégia atraente para mitigar a amostragem não-uniforme de imagens ERP é usar operadores não-locais. Ao contrário das CNNs, que apresentam um campo receptivo local com base no tamanho do *kernel*, os operadores não-locais potencialmente exploram todas as características espaciais de uma determinada camada. O representante mais conhecido de operador não local é o *Transformer*, inicialmente introduzido para processamento de texto [Vaswani et al. 2017] e posteriormente estendido para imagens. A extensão mais popular para imagens planares é o *Visual Transformer* (ViT) [Dosovitskiy

et al. 2020], que explora pequenas regiões (*patches*) não sobrepostas de uma imagem como *tokens* no *framework* original dos *Transformers*. No contexto de panoramas, Sun e colegas[Sun et al. 2021] propuseram uma arquitetura que gera primeiro características latentes na direção vertical e depois explora um *Transformer multi-head* na direção horizontal. O *Panorama Transformer* (Panoformer) [Shen et al. 2022] trabalha diretamente na representação ERP e usa planos tangentes para extrair os *tokens* baseados em *patches*. Ele também explora um *embedding* posicional relativo baseada em projeções ERP-esfera-ERP para levar em conta a localização espacial dos *tokens*. O *Parallel Convolutional Transformer* (PCFormer)[Xu et al. 2022] explora um ramo convolucional para extrair características locais e um ramo semelhante ao ViT para extrair interações de longo prazo, e, em seguida, explora um módulo de fusão de atenção dupla para fusão de características em múltiplas escalas. O modelo Trans4PASS [Zhang et al. 2022a] explora uma rede de pirâmide de características e introduz deslocamentos relativos dependentes dos dados usados para incorporar um agrupamento de *patches* deformáveis, com o objetivo de mitigar a amostragem não-uniforme da ERP. Em teoria, o suporte não local das abordagens baseadas em *Transformers* pode mitigar os problemas de circularidade e amostragem não-uniforme em representações ERP. No entanto, arquiteturas baseadas exclusivamente em *Transformers* geralmente requerem conjuntos de dados maiores e mais recursos computacionais. Dai e colegas [Dai et al. 2021] mostraram que a combinação de camadas convolucionais e *Transformers* pode ser usada para "combinar" conjuntos de dados de diferentes tamanhos para imagens de perspectiva, e o mesmo pode ser verdadeiro para panoramas.

2.3.2. Desafios com representações multi-plano e híbridas

A principal causa de distorção na representação ERP é o mapeamento de toda a esfera em um único plano. Outros mapeamentos de esfera para plano, como CMP ou baseados em icosaedro, aliviam as distorções, pois extraem projeções em planos tangentes com FoVs mais estreitos. No entanto, há um compromisso entre o FoV da imagem e a informação contextual sendo representada [da Silveira et al. 2018]: por um lado, representações multi-plano aliviam os problemas de distorção usando FoVs menores; por outro lado, cada plano contém apenas informações parciais sobre o conteúdo completo. Além disso, o problema de circularidade horizontal presente em formatos ERP é potencializado em representações multi-plano, uma vez que o mesmo conteúdo esférico pode se espalhar pelas bordas das projeções planares adjacentes. O leitor pode voltar às Figuras 2.3b e2.3c e perceber o quão intrincadas são as conexões das faces em representações esféricas multi-plano. Lidar com imagens multi-plano requer tratamento adequado, como preenchimento de faces [Wang et al. 2020a, Eder et al. 2020] ou costura [da Silveira et al. 2018, Rey-Area et al. 2022] para mitigar problemas de descontinuidade quando os planos são processados independentemente.

Outra estratégia para lidar com as bordas da representação multi-plano envolve ajustar o operador convolucional para o domínio desejado. Por exemplo, Lee *et al.* [Lee et al. 2019, Lee et al. 2020] geram uma representação icosaédrica da esfera, chamada SpherePHD, e definem *kernels* convolucionais e operadores de *pooling* que trabalham nos triângulos da tesselação que já levam em conta a conectividade das bordas. O uso conjunto de convoluções e operadores de *pooling* adaptados também expande o campo

Tabela 2.1: Análise de representações esféricas comuns para aprendizado profundo: prós, contras e estratégias para mitigação de problemas

Representação	Prós	Contras	Mitigação
ERP	Conteúdo completo (informação global) no mesmo plano	Deformações relacionadas à amostragem	Convoluções deformáveis; abordagens baseadas em *Transformers*
	Única imagem	Circularidade horizontal	Preenchimento horizontal adaptativo; convolução horizontal
Multi-plano	Menores distorções por plano (menor conforme o número de planos aumenta)	Informação local por plano (menos conteúdo conforme o número de planos aumenta)	Processamento conjunto dos planos; convoluções e *pooling* adaptados à representação
	Uso potencial de CNNs para cada plano	Conexão de bordas inter-plano	Preenchimento de bordas; pós-processamento baseado em costura; *embeddings* posicionais em *Transformers*
Híbridos	Potencial aprendizagem do melhor de cada representação	Modelos maiores e mais complexos	Destilação de conhecimento

receptivo dos *kernels* convolucionais em camadas mais profundas, o que ajuda a propagar as informações de um plano para os vizinhos. No entanto, essas abordagens geralmente não permitem transferir os pesos da rede treinada em imagens de perspectiva para o domínio esférico.

Como discutido para o formato ERP, *Transformers* também têm sido adotados para lidar com representações multi-plano. O *Cube-map Vision Transformer* (CViT) [Bai et al. 2022] visa aprender implicitamente as conexões das faces do CMP usando uma abordagem baseada em ViT. O CViT extrai *patches* planares das seis faces da CMP e usa incorporações posicionais aprendidas para manter implicitamente as informações espaciais dos *patches* e as conexões entre as faces. Li e colegas [Li et al. 2023] exploram uma amostragem *Hierarchical Equal Area isoLatitude Pixelization* (HEALPix), que é baseada em quadriláteros curvilíneos em vez de projeções planares, para gerar *tokens*. Em seguida, eles exploram uma incorporação posicional feita manualmente com base nas coordenadas esféricas do *patch* (vetor unitário na superfície da esfera) para lidar com o problema de conectividade das bordas.

Alguns autores propõem o uso de mais de um esquema de projeção para lidar melhor com a amostragem não-uniforme de imagens ERP. A rede *BiFuse* [Wang et al. 2020a] explora representações ERP e CMP através de um codificador-decodificador profundo em paralelo, com interconexões e um módulo de fusão de duas projeções. Li *et al.* [Li et al. 2022] propõem uma combinação de representações ERP, tri-cilíndrica e modificada de CMP (por meio do preenchimento das faces do cubo) no contexto de cálculo de fluxo óptico e usam um esquema de fusão profunda baseado em U-Net. O uso de várias projeções parece uma direção interessante para tirar o melhor proveito de cada uma delas. No entanto, o uso de várias representações e múltiplos ramos também aumenta o tamanho e a complexidade do modelo, o que tende a aumentar os requisitos de memória da GPU. Uma solução potencial para esse problema é o uso de técnicas de destilação de conhecimento [Gou et al. 2021], que transferem um modelo professor maior para um modelo aluno menor.

Em resumo, os desafios para explorar abordagens de aprendizado profundo dependem da representação escolhida. A Tabela 2.1 lista os prós e contras das representações ERP, multi-plano, híbridas e estratégias de mitigação representativas.

2.4. Algumas Aplicações

Embora praticamente qualquer aplicação possa se beneficiar do uso de imagens esféricas, essa seção foca em algumas aplicações que naturalmente requerem um FoV mais amplo.

2.4.1. Correção de orientação

Embora uma imagem esférica capture todo ambiente ao redor da câmera, a visualização é normalmente feita em dispositivos convencionais (como celulares e monitores), que possuem FoV limitado. Assim, é necessário recortar uma porção da imagem esférica e realizar uma projeção planar antes da visualização [da Silveira and Jung 2023]. Uma etapa de pré-processamento usual consiste em rotacionar o conteúdo da imagem esférica para que o horizonte da imagem se alinhe com o horizonte no mundo, obtendo-se uma orientação canônica de captura. Tal processo de correção de orientação é normalmente

(a) (b)

Figura 2.5: A (a) tilted 360° capture (in ERP format) of the same scene as in Figura 2.1b and (b) its upright aligned version using the estimates from the method in [Bergmann et al. 2021].

chamado de *gravity alignment*, *horizon alignment* ou *upright adjustment*.

O objetivo de uma abordagem correção de orientação é estimar uma matriz de rotação $\mathbf{R}^{\uparrow} \in SO(3)$ que alinha o plano do solo com o equador e posiciona os objetos da cena verticalmente. A correção da orientação de uma imagem ERP, por exemplo, é tipicamente realizada pelos seguintes passos. Primeiro, precisamos projetar a imagem (suas intensidades de luz) na esfera unitária usando a Eq.(2). Em seguida, temos que girar a esfera (ou seja, todos os pontos em sua superfície individualmente) usando $\mathbf{R}^{\uparrow\top}$. Por fim, precisamos projetar as intensidades de luz associadas a esses pontos de volta ao plano usando a Eq.(3). Vale ressaltar que a imagem de entrada, com orientação arbitrária, sofrerá diferentes operações de reamostragem para ser mapeada na imagem corrigida verticalmente de saída. Por exemplo, as informações nos polos da imagem de entrada podem ser reduzidas para se ajustarem às latitudes centrais da imagem de saída. *Pixels* originalmente no equador também podem ser mapeados para os polos da imagem de saída, perdendo grande parte das informações de alta frequência [Murrugarra-Llerena et al. 2022]. Também cabe notar que a rotação desejada envolve dois graus de liberdade, visto que qualquer rotação em torno do eixo vertical para uma imagem já alinhada segue gerando uma imagem alinhada. O terceiro grau de liberdade pode ser ajustado para colocar conteúdo de interesse à frente do panorama. Isso pode ser feito manualmente, onde o usuário ajusta a rotação de acordo com seus interesses pessoais, ou até de modo automático, através do uso de técnicas que identificam regiões "interessantes" na imagem esférica usando técnicas de saliência visual [Bernal-Berdun et al. 2022].

A tendência atual para abordar o problema de correção de orientação é usar técnicas de aprendizado de máquina, tipicamente usando redes profundas, para estimar o vetor de orientação (*upright vector*) da imagem de entrada. Com isso, se pode gerar a matriz de rotação e gerar uma versão canônica da imagem de entrada. Para tal, são necessárias bases de dados contendo imagens na orientação canônica, a partir das quais se pode gerar rotações arbitrárias para construir os dados de treinamento anotados: a imagem rotacionada é o dado de entrada, e o *upright vector* é o dado que a rede precisa inferir. Por exemplo, a base SUN360 [J. Xiao et al. 2012] forece cerca de 57,000 panoramas em ambientes internos e externos, e pode ser usada nessa tarefa.

Outro ponto importante se refere à avaliação dos resultados de correção de orientação. Como discutido em Jung *et al.* [Jung et al. 2019], discrepâncias angulares menores

que 5° são consideradas muito satisfatórias pelos seres humanos, enquanto aquelas menores que 12° são consideradas satisfatórias. Por exemplo, considere a imagem ERP mostrada na Figura 2.5a, que corresponde a uma captura inclinada do ambiente ilustrado na Figura 2.3a. Como se pode perceber, a imagem apresenta fortes distorções visuais, sendo inclusive difícil a identificação dos elementos da cena (cadeiras, quadro, teto, etc.). Aplicando o algoritmo de correção de orientação proposto em [Bergmann et al. 2021] gera a imagem mostrada na Figura 2.5b, que visualmente alinha o plano horizontal do mundo (sala de aula) com o equador da imagem esférica.

2.4.2. Reconstrução 3D

A reconstrução tridimensional de cenas desempenha um papel fundamental em uma ampla gama de aplicações, abrangendo campos que vão desde a realidade virtual/aumentada até a robótica, a análise de cenas forenses e a indústria do entretenimento [da Silveira and Jung 2023]. A habilidade de transformar uma cena bidimensional capturada em uma representação 3D precisa e detalhada é essencial para compreender e interagir com o ambiente de maneira mais rica e imersiva. Além disso, a reconstrução 3D é valiosa na preservação do patrimônio cultural, permitindo a digitalização precisa de artefatos históricos e locais arquitetônicos, o que facilita a documentação, a restauração e a disseminação do conhecimento. Na visão computacional tradicional, usando imagens em perspectiva, são necessárias diversas capturas com diferentes pontos de vista para que se tenha uma cobertura completa do ambiente. Por outro lado, uma única captura esférica já fornece toda informação visual em torno da câmera.

Uma imagem esférica armazena a informação visual (cor) ao longo dos raios emitidos em torno da câmera. Para gerar a reconstrução 3D da cena, é necessário estimar a distância de cada um desses raios, gerando uma representação RGB-D (cor + distância) que pode ser diretamente mapeada para uma nuvem de pontos colorida. Do ponto de vista físico, não é possível obter a informação de profundidade a partir de uma única captura esférica, visto que qualquer ponto ao longo de um raio 3D é mapeado na mesma posição da superfície da esfera unitária. Dessa forma, seriam necessárias duas ou mais capturas da cena para inferir a profundidade da cena, como na estereoscopia clássica [Hartley and Zisserman 2003]. Em particular, considerar múltiplas vistas (mais de duas) adiciona robustez na estimativa da profundidade [da Silveira and Jung 2019a]. Por outro lado, o uso de múltiplas imagens requer a obtenção de múltiplas capturas da mesma cena, o que pode ser um fator complicador sobretudo se houver objetos em movimento na cena: capturas em instantes de tempos distintos não geram vistas da mesma cena devido ao movimento relativo dos objetos dinâmicos.

Apesar de ser fisicamente implausível, se pode estimar a profundidade a partir de uma única captura. No caso de imagens em perspectiva, os seres humanos conseguem estimar distâncias a partir de uma única imagem. Para tal, usam relações entre objetos cujos tamanhos são conhecidos no mundo real, e inferem uma noção de distância a partir do tamanho projetado do objeto na imagem: quanto mais distante, menor será a projeção do objeto. Inspirados por essa característica, vários algoritmos de visão computacional baseados em aprendizado de máquina têm sido propostos para estimar a profundidade a partir de uma única imagem, problema comumente chamado de *single-image stereo* ou *monocular depth estimation* [Masoumian et al. 2022]. Seguindo essa estratégia baseada

em aprendizado de máquina, alguns autores têm atacado o problema de estimativa de profundidade usando uma única captura panorâmica.

A maioria das abordagens de estimativa de profundidade a partir de um único panorama adota imagens ERP [Tateno et al. 2018, Sun et al. 2021, Albanis et al. 2021], que contêm todas as informações contextuais, mas apresentam distorções acentuadas, ou imagens de múltiplos planos [da Silveira et al. 2018, Rey-Area et al. 2022], que aliviam as distorções, mas exigem que as projeções individuais sejam unidas de volta à esfera. Conforme discutido brevemente na Seção 2.3.2, alguns trabalhos utilizam duas representações planares da esfera e incorporam o mapeamento delas no processo de aprendizado para aliviar os problemas de cada uma individualmente [Wang et al. 2020a, Jiang et al. 2021]. Outras abordagens consideram representações de imagem baseadas em icosaedro e arquiteturas de rede especializadas [Lee et al. 2020].

As abordagens que exploram uma única representação ERP geralmente estão restritas a panoramas de baixa resolução (512×256 ou 1024×512) devido às limitações de memória da GPU, o que pode não ser adequado para aplicações de realidade virtual que envolvem HMDs de alta qualidade [Liu et al. 2021]. Para esses cenários, o uso de representações de múltiplos planos [da Silveira et al. 2018, Rey-Area et al. 2022] é uma solução possível, já que cada projeção planar contém uma imagem com um FoV estreito que pode ser tratado individualmente. No entanto, tais abordagens devem lidar com descontinuidades entre representações planares adjacentes e falta de informação contextual para processar cada projeção planar. Outro desafio na estimativa de profundidade monocular diz respeito a cenas capturadas em ambientes externos. A maioria dos métodos pode não generalizar para esses cenários devido à falta de conjuntos de dados anotados (necessários para o treinamento supervisionado), e eventual incapacidade para lidar com valores de profundidade infinitos, como no céu. Felizmente, novas abordagens começaram a abordar esse aspecto relevante [Bhanushali et al. 2022].

A Figura 2.6a mostra a nuvem de pontos colorida (vista de fora da sala) associada à estimativa de mapa de profundidade pelo modelo de aprendizado baseado em U-Net de[Albanis et al. 2021] usando a imagem na Figura 2.3a como entrada. Embora a estrutura externa do ambiente tenha sido capturado, percebe-se que que as paredes laterais e o teto não são exatamente planares.

Técnicas de estimativa de profundidade densas são capazes de estimar a posição 3D de cada ponto do panorama. Por um lado, permitem uma granularidade fina da cena; por outro lado, podem não considerar alguma informação geométrica pré-existente na cena e gerar modelos 3D com menor precisão, como ilustrado na Figura 2.6a (o fato de que paredes em um ambiente interno são normalmente planares não foi levado em consideração, e gerou "ondulações" na nuvem de pontos estimada).

Para algumas aplicações específicas, como a modelagem de ambientes internos cuja geometria segue padrões pré-definidos, se pode fazer o uso de técnicas de estimativa de *layout*. Os métodos de estimativa de *layout* têm como objetivo recuperar uma representação tridimensional esparsa a partir de um panorama capturado no interior de um ambiente, gerando informações sobre a geometria do mesmo (quintas e junções entre paredes, teto e piso, por exemplo).

(a) (b)

Figura 2.6: Duas maneiras de extração de um modelo 3D a partir da ERP ilustrada na Figura 2.3a: (a) estimativa de profundidade densa (ponto-a-ponto) usando [Albanis et al. 2021] e (b) estimativa do *layout* usando [Wang et al. 2021].

Métodos pioneiros para estimativa de *layout* eram semi-automáticos ou usavam explicitamente primitivas geométricas ou pontos/linhas de fuga [Jia and Li 2015]. Já abordagens mais recentes [Sun et al. 2019, Sun et al. 2021, Fernandez-Labrador et al. 2020, Wang et al. 2021, Jiang et al. 2022, Pintore et al. 2020a] atacam o problema a partir de uma perspectiva de aprendizado de máquina, e usam uma única imagem omnidirecional como entrada. As técnicas existentes exploram diferentes maneiras de representar o panorama (ERP, CMP, etc.) e adotam arquiteturas de rede compatíveis com o dado de entrada, como discutido na Seção 2.3. Destacamos que alguns métodos [Zou et al. 2018, Zhang et al. 2014] incluem uma etapa de pré-processamento que alinha as imagens esféricas verticalmente antes da inferência do *layout*, ressaltando a importância dos métodos discutidos na Seção 2.4.1. A maioria dos métodos de estimativa de *layout* opta por fazer a regressão das junções de paredes/teto/piso [Sun et al. 2019, Sun et al. 2021, Fernandez-Labrador et al. 2020] de ambientes internos, embora também seja possível inferir suas bordas [Pintore et al. 2020b].

Abordagens de estimativa de *layout* se baseiam em restrições geométricas que guiam o processo de otimização [da Silveira et al. 2022]. O modelo geométrico mais simples usado na estimativa de *layout* é um cuboide, que implica um *layout* em forma de caixa (também chamada de cuboide) [Zhang et al. 2014]. A forma de cuboide é um caso particular para a hipótese de "Mundo Manhattan", no qual o *layout* da sala tem paredes perpendiculares entre si [Fernandez-Labrador et al. 2020, Wang et al. 2021], ou seja, o plano de chão do ambiente é uma região poligonal com segmentos de reta ortogonais entre si. Assim, os modelos Manhattan compreendem ambientes mais complexos, como salas em forma de "L". Mundos Manhattan aumentados podem ter paredes que não são perpendiculares entre si [Pintore et al. 2018, Fernandez-Labrador et al. 2020]. Por fim, a restrição de *layout* mais genérica é chamada de suposição de Mundo Atlanta, onde até paredes curvas podem existir, desde que o teto e o piso sejam paralelos [Pintore et al. 2020a].

A Figura 2.6b mostra um exemplo de estimativa de *layout* a partir de um único panorama, onde o método de[Wang et al. 2021] utiliza a representação ERP da cena *Classroom* mostrada na Figura 2.3a como entrada. Embora o método de[Wang et al. 2021]

seja capaz de lidar com layouts genéricos de Manhattan, ele consegue detectar corretamente o *layout* em forma de cuboide da imagem de entrada, gerando paredes planares e ortogonais, contrastando com a técnica densa mostrada na Figura 2.6a. Por outro lado, objetos não-planares (como as mesas no chão) estão distorcidos na representação final e planificados na face do cuboide que representa o chão.

Estabelecer um *benchmark* para estimativas de profundidade e *layout* de panoramas é um desafio devido à diversidade de abordagens listadas nesta seção. O leitor pode consultar os trabalhos [da Silveira and Jung 2023] e [da Silveira et al. 2022] que compilam bases de dados e métricas de avaliação adequadas às variantes desses dois problemas.

2.5. Considerações Finais

Este capítulo tem como objetivo fornecer uma introdução sólida à computação visual omnidirecional. Inicialmente, ele revisou o modelo de imageamento esférico, *pipelines* de aquisição existentes e formatos de representação (multi-)planar proeminentes usados para armazenar e processar mídia omnidirecional. Em seguida, o artigo apresentou os principais desafios das representações omnidirecionais, focando nas distorções inerentes a essas representações e em seu impacto nas arquiteturas de aprendizado profundo, que são a abordagem atual para processar panoramas. Este capítulo também apresentou tendências recentes para mitigar esses desafios e mostrou os avanços em três cenários de aplicação que exploram plenamente imagens omnidirecionais.

Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Referências

[Adarve and Mahony 2017] Adarve, J. D. and Mahony, R. (2017). Spherepix: A data structure for spherical image processing. *IEEE Robotics and Automation Letters*, 2(2):483–490.

[Aggarwal et al. 2016] Aggarwal, R., Vohra, A., and Namboodiri, A. M. (2016). Panoramic Stereo Videos with a Single Camera. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3755–3763.

[Akihiko et al. 2005] Akihiko, T., Atsushi, I., and Ohnishi, N. (2005). Two-and three-view geometry for spherical cameras. *Workshop on Omnidirectional Vision, Camera Networks and Non-classical Cameras*, 105:29–34.

[Albanis et al. 2021] Albanis, G., Zioulis, N., Drakoulis, P., Gkitsas, V., Sterzentsenko, V., Alvarez, F., Zarpalas, D., and Daras, P. (2021). Pano3d: A holistic benchmark and a solid baseline for 360° depth estimation. In *Conference on Computer Vision and Pattern Recognition Workshops*, pages 3722–3732.

[Azevedo et al. 2020] Azevedo, R. G. d. A., Birkbeck, N., De Simone, F., Janatra, I., Adsumilli, B., and Frossard, P. (2020). Visual Distortions in 360-degree Videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(8):2524–2537.

[Bai et al. 2022] Bai, J., Lai, S., Qin, H., Guo, J., and Guo, Y. (2022). Glpanodepth: Global-to-local panoramic depth estimation. *arXiv preprint arXiv:2202.02796*.

[Bergmann et al. 2021] Bergmann, M. A., Pinto, P. G. L., da Silveira, T. L. T., and Jung, C. R. (2021). Gravity alignment for single panorama depth inference. In *Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 1–8. IEEE.

[Bernal-Berdun et al. 2022] Bernal-Berdun, E., Martin, D., Gutierrez, D., and Masia, B. (2022). Sst-sal: A spherical spatio-temporal approach for saliency prediction in 360° videos. *Computers & Graphics*, 106:200–209.

[Bhanushali et al. 2022] Bhanushali, J., Chakravarthula, P., and Muniyandi, M. (2022). OmniHorizon: In-the-wild outdoors depth and normal estimation from synthetic omnidirectional dataset.

[Coors et al. 2018] Coors, B., Condurache, A. P., and Geiger, A. (2018). SphereNet: Learning spherical representations for detection and classification in omnidirectional images. *European Conference on Computer Vision*, pages 525–541.

[Cruz-Mota et al. 2012] Cruz-Mota, J., Bogdanova, I., Paquier, B., Bierlaire, M., and Thiran, J. P. (2012). Scale invariant feature transform on the sphere: Theory and applications. *International Journal of Computer Vision*, 98(2):217–241.

[da Silveira and Jung 2023] da Silveira, T. L. and Jung, C. R. (2023). Omnidirectional visual computing: Foundations, challenges, and applications. *Computers & Graphics*, 113:89–101.

[da Silveira et al. 2018] da Silveira, T. L. T., Dalaqua, L. P., and Jung, C. R. (2018). Indoor Depth Estimation from Single Spherical Images. In *IEEE International Conference on Image Processing*, pages 2935–2939.

[da Silveira et al. 2021] da Silveira, T. L. T., de Oliveira, A. Q., Walter, M., and Jung, C. R. (2021). Fast and accurate superpixel algorithms for 360° images. *Signal Processing*, 189:108277.

[da Silveira and Jung 2019a] da Silveira, T. L. T. and Jung, C. R. (2019a). Dense 3D Scene Reconstruction from Multiple Spherical Images for 3-DoF+ VR Applications. In *IEEE Conference on Virtual Reality and 3D User Interfaces*, pages 9–18.

[da Silveira and Jung 2019b] da Silveira, T. L. T. and Jung, C. R. (2019b). Perturbation Analysis of the 8-Point Algorithm: A Case Study for Wide FoV Cameras. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11757–11766.

[da Silveira et al. 2022] da Silveira, T. L. T., Pinto, P. G. L., Murrugarra-Llerena, J., and Jung, C. R. (2022). 3d scene geometry estimation from 360° imagery: A survey. *ACM Comput. Surv.*, 55(4).

[Dai et al. 2019] Dai, F., Zhu, C., Ma, Y., Cao, J., Zhao, Q., and Zhang, Y. (2019). Freely Explore the Scene with 360° Field of View. In *IEEE Conference on Virtual Reality and 3D User Interfaces*, pages 888–889.

[Dai et al. 2021] Dai, Z., Liu, H., Le, Q. V., and Tan, M. (2021). Coatnet: Marrying convolution and attention for all data sizes. *Advances in Neural Information Processing Systems*, 34:3965–3977.

[De Simone et al. 2017] De Simone, F., Frossard, P., Wilkins, P., Birkbeck, N., and Kokaram, A. (2017). Geometry-driven quantization for omnidirectional image coding. *2016 Picture Coding Symposium, PCS 2016*.

[Dosovitskiy et al. 2020] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

[Ebrahimi et al. 2016] Ebrahimi, T., Foessel, S., Pereira, F., and Schelkens, P. (2016). JPEG Pleno: Toward an Efficient Representation of Visual Reality. *IEEE Multimedia*, 23(4):14–20.

[Eder et al. 2019] Eder, M., Moulon, P., and Guan, L. (2019). Pano Popups: Indoor 3D Reconstruction with a Plane-Aware Network. In *2019 International Conference on 3D Vision (3DV)*, pages 76–84. IEEE.

[Eder et al. 2020] Eder, M., Shvets, M., Lim, J., and Frahm, J.-M. (2020). Tangent images for mitigating spherical distortion. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[Fangi et al. 2018] Fangi, G., Pierdicca, R., Sturari, M., and Malinverni, E. S. (2018). Improving spherical photogrammetry using 360° OMNI-Cameras: Use cases and new applications. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42(2):331–337.

[Fernandez-Labrador et al. 2020] Fernandez-Labrador, C., Facil, J. M., Perez-Yus, A., Demonceaux, C., Civera, J., and Guerrero, J. (2020). Corners for layout: End-to-end layout recovery from 360 images. *IEEE Robotics and Automation Letters*, pages 1–1.

[Ferreira et al. 2017] Ferreira, L. S., Sacht, L., and Velho, L. (2017). Local Moebius transformations applied to omnidirectional images. *Computers & Graphics*, 68:77–83.

[Fujiki et al. 2007] Fujiki, J., Torii, A., and Akaho, S. (2007). Epipolar Geometry Via Rectification of Spherical Images. In *Computer Vision/Computer Graphics Collaboration Techniques*, volume 4418, pages 461–471. Springer Berlin Heidelberg.

[Gava et al. 2018] Gava, C. C., Stricker, D., and Yokota, S. (2018). Dense Scene Reconstruction from Spherical Light Fields. In *IEEE International Conference on Image Processing*, pages 4178–4182.

[Goodfellow et al. 2016] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. The MIT Press.

[Gou et al. 2021] Gou, J., Yu, B., Maybank, S. J., and Tao, D. (2021). Knowledge distillation: A survey. *International Journal of Computer Vision*, 129:1789–1819.

[Guan and Smith 2017] Guan, H. and Smith, W. A. P. (2017). Structure-From-Motion in Spherical Video Using the von Mises-Fisher Distribution. *IEEE Transactions on Image Processing*, 26(2):711–723.

[Hartley and Zisserman 2003] Hartley, R. and Zisserman, A. (2003). *Multiple View Geometry in Computer Vision*. Cambridge.

[Im et al. 2016] Im, S., Ha, H., Rameau, F., Jeon, H.-G., Choe, G., and Kweon, I. S. (2016). All-around depth from small motion with a spherical panoramic camera. In *European Conference on Computer Vision*, pages 156–172.

[J. Huang et al. 2017] J. Huang, Z. Chen, D. Ceylan, and H. Jin (2017). 6-DoF VR videos with a single 360-camera. In *IEEE Virtual Reality*, pages 37–44.

[J. Xiao et al. 2012] J. Xiao, E., K. A., Oliva, A., and Torralba, A. (2012). Recognizing scene viewpoint using panoramic place representation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2695–2702.

[Jia and Li 2015] Jia, H. and Li, S. (2015). Estimating structure of indoor scene from a single full-view image. In *IEEE International Conference on Robotics and Automation*, pages 4851–4858.

[Jiang et al. 2021] Jiang, H., Sheng, Z., Zhu, S., Dong, Z., and Huang, R. (2021). Unifuse: Unidirectional fusion for 360° panorama depth estimation. *IEEE Robotics and Automation Letters*, 6(2):1519–1526.

[Jiang et al. 2022] Jiang, Z., Xiang, Z., Xu, J., and Zhao, M. (2022). LGT-net: Indoor panoramic room layout estimation with geometry-aware transformer network. In *Conference on Computer Vision and Pattern Recognition*.

[Jung et al. 2019] Jung, R., Lee, A. S. J., Ashtari, A., and Bazin, J.-C. (2019). Deep360Up: A Deep Learning-Based Approach for Automatic VR Image Upright Adjustment. In *IEEE Conference on Virtual Reality and 3D User Interfaces*, pages 1–8.

[Krolla et al. 2014] Krolla, B., Diebold, M., Goldlücke, B., and Stricker, D. (2014). Spherical light fields. *British Machine Vision Conference*, (67.1-67.12).

[Lee et al. 2019] Lee, Y., Jeong, J., Yun, J., Cho, W., and Yoon, K.-J. (2019). Spherephd: Applying cnns on a spherical polyhedron representation of 360 images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9181–9189.

[Lee et al. 2020] Lee, Y., Jeong, J., Yun, J., Cho, W., and Yoon, K.-J. (2020). Spherephd: Applying cnns on 360° images with non-euclidean spherical polyhedron representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1.

[Li et al. 2023] Li, M., Wang, S., Yuan, W., Shen, W., Sheng, Z., and Dong, Z. (2023). $\mathscr{S}^2$net: Accurate panorama depth estimation on spherical surface. *IEEE Robotics and Automation Letters*, 8(2):1053–1060.

[Li 2008] Li, S. (2008). Binocular spherical stereo. *IEEE Transactions on Intelligent Transportation Systems*, 9(4):589–600.

[Li et al. 2022] Li, Y., Barnes, C., Huang, K., and Zhang, F.-L. (2022). Deep 360° optical flow estimation based on multi-projection fusion. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, pages 336–352. Springer.

[Liu et al. 2021] Liu, R., Peng, C., Zhang, Y., Husarek, H., and Yu, Q. (2021). A survey of immersive technologies and applications for industrial product development. *Computers & Graphics*, 100:137–151.

[Lo et al. 2018] Lo, I., Shih, K., and Chen, H. H. (2018). Image stitching for dual fisheye cameras. In *IEEE International Conference on Image Processing*, pages 3164–3168.

[Masoumian et al. 2022] Masoumian, A., Rashwan, H. A., Cristiano, J., Asif, M. S., and Puig, D. (2022). Monocular depth estimation using deep learning: A review. *Sensors*, 22(14):5353.

[Murrugarra-Llerena et al. 2022] Murrugarra-Llerena, J., da Silveira, T. L. T., and Jung, C. R. (2022). Pose estimation for two-view panoramas based on keypoint matching: A comparative study and critical analysis. In *Conference on Computer Vision and Pattern Recognition Workshops*, pages 5202–5211.

[Nayar 1997] Nayar, S. K. (1997). Catadioptric Omnidirectional Camera*. In *Conference on Computer Vision and Pattern Recognition*, pages 482–488.

[Pintore et al. 2020a] Pintore, G., Agus, M., and Gobbetti, E. (2020a). AtlantaNet: Inferring the 3D indoor layout from a single 360 image beyond the Manhattan world assumption. In *European Conference on Computer Vision*.

[Pintore et al. 2020b] Pintore, G., Mura, C., Ganovelli, F., Fuentes-Perez, L., Pajarola, R., and Gobbetti, E. (2020b). State-of-the-art in automatic 3d reconstruction of structured indoor environments. *Computer Graphics Forum*, 39(2).

[Pintore et al. 2018] Pintore, G., Pintus, R., Ganovelli, F., Scopigno, R., and Gobbetti, E. (2018). Recovering 3d existing-conditions of indoor structures from spherical images. *Computers & Graphics*, 77:16–29.

[Rey-Area et al. 2022] Rey-Area, M., Yuan, M., and Richardt, C. (2022). 360monodepth: High-resolution 360° monocular depth estimation. In *Conference on Computer Vision and Pattern Recognition*, pages 3762–3772.

[S. Li and Fukumori 2005] S. Li and Fukumori, K. (2005). Spherical stereo for the construction of immersive vr environment. In *IEEE Virtual Reality*, pages 217–222.

[Serrano et al. 2019] Serrano, A., Kim, I., Chen, Z., DIVerdi, S., Gutierrez, D., Hertzmann, A., and Masia, B. (2019). Motion parallax for 360° RGBD video. *IEEE Transactions on Visualization and Computer Graphics*, 25(5):1817–1827.

[Shan and Li 2018] Shan, Y. and Li, S. (2018). Descriptor Matching for a Discrete Spherical Image With a Convolutional Neural Network. *IEEE Access*, 6:20748–20755.

[Shen et al. 2022] Shen, Z., Lin, C., Liao, K., Nie, L., Zheng, Z., and Zhao, Y. (2022). Panoformer: Panorama transformer for indoor 360° depth estimation. In *European Conference on Computer Vision*, pages 195–211. Springer.

[Su and Grauman 2017] Su, Y.-C. and Grauman, K. (2017). Learning Spherical Convolution for Fast Features from 360° Imagery. In *Conference on Neural Information Processing Systems*, pages 529–539.

[Sun et al. 2019] Sun, C., Hsiao, C.-W., Sun, M., and Chen, H.-T. (2019). HorizonNet: Learning Room Layout with 1D Representation and Pano Stretch Data Augmentation. pages 1047–1056.

[Sun et al. 2021] Sun, C., Sun, M., and Chen, H.-T. (2021). Hohonet: 360 indoor holistic understanding with latent horizontal features. In *Conference on Computer Vision and Pattern Recognition*, pages 2573–2582.

[Tateno et al. 2018] Tateno, K., Navab, N., and Tombari, F. (2018). Distortion-Aware Convolutional Filters for Dense Prediction in Panoramic Images. *European Conference on Computer Vision*, pages 732–750.

[Vaswani et al. 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30:5998–6008.

[Wang et al. 2018a] Wang, F.-E., Hu, H.-N., Cheng, H.-T., Lin, J.-T., Yang, S.-T., Shih, M.-L., Chu, H.-K., and Sun, M. (2018a). Self-supervised Learning of Depth and Camera Motion from 360° Videos. volume 11364, pages 53–68. Asian Conference on Computer Vision.

[Wang et al. 2020a] Wang, F.-E., Yeh, Y.-H., Sun, M., Chiu, W.-C., and Tsai, Y.-H. (2020a). Bifuse: Monocular 360 depth estimation via bi-projection fusion. In *Conference on Computer Vision and Pattern Recognition*.

[Wang et al. 2021] Wang, F.-E., Yeh, Y.-H., Sun, M., Chiu, W.-C., and Tsai, Y.-H. (2021). LED2-Net: Monocular 360° layout estimation via differentiable depth rendering. pages 12956–12965.

[Wang et al. 2020b] Wang, M., Lyu, X.-Q., Li, Y.-J., and Zhang, F.-L. (2020b). VR content creation and exploration with deep learning: A survey. *Computational Visual Media*, 6(1):3–28.

[Wang et al. 2018b] Wang, T.-H., Huang, H.-J., Lin, J.-T., Hu, C.-W., Zeng, K.-H., and Sun, M. (2018b). Omnidirectional CNN for visual place recognition and navigation. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2341–2348. IEEE.

[Xu et al. 2022] Xu, C., Yang, H., Han, C., and Zhang, C. (2022). Pcformer: A parallel convolutional transformer network for 360° depth estimation. *IET Computer Vision*.

[Xu et al. 2020] Xu, M., Li, C., Zhang, S., and Callet, P. L. (2020). State-of-the-art in 360° video/image processing: Perception, assessment and compression. *IEEE Journal of Selected Topics in Signal Processing*, 14(1):5–26.

[Yang et al. 2018] Yang, W., Qian, Y., Kamarainen, J. K., Cricri, F., and Fan, L. (2018). Object Detection in Equirectangular Panorama. *International Conference on Pattern Recognition*, pages 2190–2195.

[Yu and Koltun 2015] Yu, F. and Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*.

[Zelnik-Manor et al. 2005] Zelnik-Manor, L., Peters, G., and Perona, P. (2005). Squaring the circle in panoramas. In *IEEE International Conference on Computer Vision*, volume 2, pages 1292–1299 Vol. 2.

[Zhang et al. 2022a] Zhang, J., Yang, K., Ma, C., Reiß, S., Peng, K., and Stiefelhagen, R. (2022a). Bending reality: Distortion-aware transformers for adapting to panoramic semantic segmentation. In *Conference on Computer Vision and Pattern Recognition*, pages 16917–16927.

[Zhang et al. 2014] Zhang, Y., Song, S., Tan, P., and Xiao, J. (2014). PanoContext: A whole-room 3D context model for panoramic scene understanding. In *European Conference on Computer Vision*.

[Zhang et al. 2021] Zhang, Y., Zhang, F.-L., Lai, Y.-K., and Zhu, Z. (2021). Efficient propagation of sparse edits on 360° panoramas. *Computers & Graphics*, 96:61–70.

[Zhang et al. 2022b] Zhang, Y., Zhang, F.-L., Zhu, Z., Wang, L., and Jin, Y. (2022b). Fast edit propagation for 360 degree panoramas using function interpolation. *IEEE Access*, 10:43882–43894.

[Zhuang et al. 2022] Zhuang, C., Lu, Z., Wang, Y., Xiao, J., and Wang, Y. (2022). Acdnet: Adaptively combined dilated convolution for monocular panorama depth estimation. In *AAAI Conference on Artificial Intelligence*, volume 36, pages 3653–3661.

[Zioulis et al. 2021] Zioulis, N., Alvarez, F., Zarpalas, D., and Daras, P. (2021). Single-shot cuboids: Geodesics-based end-to-end manhattan aligned layout estimation from spherical panoramas.

[Zioulis et al. 2018] Zioulis, N., Karakottas, A., Zarpalas, D., and Daras, P. (2018). OmniDepth: Dense Depth Estimation for Indoors Spherical Panoramas. In *European Conference on Computer Vision*, pages 453–471.

[Zou et al. 2018] Zou, C., Colburn, A., Shan, Q., and Hoiem, D. (2018). LayoutNet: Reconstructing the 3D Room Layout from a Single RGB Image. In *Conference on Computer Vision and Pattern Recognition*, pages 2051–2059.