中图法分类号:TP391　　文献标识码:A　　文章编号:1006-8961(2024)10-2955-24

论文引用格式:Li Y Q, Hu X, Xu X, Xu Y N and Wang L. 2024. Deep learning-based foveated rendering in 3D space: a review. Journal of Image and Graphics, 29(10):2955-2978(李英群, 胡啸, 徐翔, 徐延宁, 王璐. 2024. 三维场景注视点渲染深度学习方法综述. 中国图象图形学报, 29(10):2955-2978)[DOI:10.11834/jig.230708]

三维场景注视点渲染深度学习方法综述

李英群[1],胡啸[1],徐翔[2],徐延宁[1],王璐[1*]

1. 山东大学软件学院,济南　250101; 2. 山东财经大学山东省区块链金融重点实验室,济南　250014

摘　要: 在大型高分辨显示器和头戴式显式设备中实现实时、逼真的渲染仍然是计算机图形学面临的主要挑战之一。注视点渲染(foveated rendering)利用人类视觉系统的局限性,根据注视点调整图像渲染质量,从而在不损失用户感知质量的前提下大大提高渲染速度。随着深度学习方法在渲染领域的广泛应用,涌现出大量基于深度学习的注视点渲染新方法。本文从深度学习的角度对注视点渲染领域的最新方法进行综述。首先,概述了人类视觉感知的背景知识。接着,简要介绍了注视点渲染中最具代表性的非深度学习方法,包括自适应分辨率、几何简化、着色简化和硬件实现,并总结了这些方法的优缺点。随后,描述了文中用于评估深度学习不同方法所使用的评估准则,包括常用的注视点渲染图像的评估指标和注视点预测评估指标。接下来,将注视点渲染中的深度学习方法细分为超分辨率、降噪、补全、图像合成、注视点预测和图像应用,对它们进行详细概述和总结。最后,提出了深度学习方法目前面临的问题和挑战。通过对注视点渲染领域的深度学习方法的讨论,可以更详细地展示深度学习在注视点渲染中的研究前景和发展方向,对后续研究人员在选择研究方向和设计网络架构等方面都有一定的参考价值。

关键词: 注视点渲染;深度学习;实时渲染;注视点预测;图像补全;超分辨率;光路追踪降噪

Deep learning-based foveated rendering in 3D space: a review

Li Yingqun[1], Hu Xiao[1], Xu Xiang[2], Xu Yanning[1], Wang Lu[1*]

1. School of Software, Shandong University, Jinan　250101, China;

2. Shandong Key Laboratory of Blockchain Finance, Shandong University of Finance and Economics, Jinan　250014, China

Abstract: The widespread adoption of virtual reality (VR) and augmented reality technologies across various sectors, including healthcare, education, military, and entertainment, has propelled head-mounted displays with high resolution and wide fields of view into the forefront of display devices. However, attaining a satisfactory level of immersion and interactivity poses a primary challenge in the realm of VR, with latency potentially leading to user discomfort in the form of dizziness and nausea. Multiple studies have underscored the necessity of achieving a highly realistic VR experience while maintaining user comfort, entailing the elevation of the screen's image refresh rate to 1 800 Hz and keeping latency below 3～40 ms. Achieving real-time, photorealistic rendering at high resolution and low latency represents a formidable objective. Foveated rendering is an effective approach to address these issues by adjusting the rendering quality across the image based on gaze position, maintaining high quality in the fovea area while reducing quality in the periphery. This technique leads to substantial computational savings and improved rendering speed without a perceptible loss in visual quality. While previous reviews have examined technical approaches to foveated rendering, they focused more on categorizing the imple-

收稿日期:2023-10-17;修回日期:2023-11-29;预印本日期:2023-12-05

*通信作者:王璐　luwang_hcivr@sdu.edu.cn

基金项目:国家重点研发计划资助(2022YFB3303203);国家自然科学基金项目(62272275)

Supported by: National Key R&D Program of China (2022YFB3303203); National Natural Science Foundation of China (62272275)

mentation techniques. A comprehensive review within the domain of machine learning still needs to be explored. With the ongoing advancements in machine learning within the rendering field, combining machine learning and foveated rendering is considered a promising research area, especially in postprocessing, where machine learning methods have great potential. Nonmachine learning methods inevitably introduce artifacts. By contrast, machine learning methods have a wide range of applications in the postprocessing domain of rendering to optimize and improve foveated rendering results and enhance the realism and immersion of foveated images in a manner unattainable through nonmachine learning approaches. Therefore, this work presents a comprehensive overview of foveated rendering from a machine-learning perspective. In this paper, we first provide an overview of the background knowledge of human visual perception, including aspects of the human visual system, contrast sensitivity functions, visual acuity models, and visual crowding. Subsequently, this paper briefly describes the most representative nonmachine learning methods for point-of-attention rendering, including adaptive resolution, geometric simplification, shading simplification, and hardware implementation, and summarizes these methods' features, advantages, and disadvantages. Additionally, we describe the criteria employed for method evaluation in this review, including evaluation metrics for foveated images and gaze-point prediction. Next, we subdivide machine learning methods into super-resolution, denoise, image reconstruction, image synthesis, gaze prediction, and image application. We provide a detailed summary of them in terms of four aspects: results quality, network speed, user experience, and the ability to handle objects. Among them, super-resolution methods commonly use more neural blocks in the foveal region while fewer neural blocks in the periphery region, resulting in variable regional super-resolution quality. Similarly, foveated denoising usually performs fine denoising in the fovea and coarse denoising in the peripheral, but the denoising aspect has yet to receive extensive attention. The initial attempt to integrate image reconstruction with gaze utilized generative adversarial networks (GANs), yielding promising outcomes. Then, some researchers combined direct prediction and kernel prediction for image reconstruction, which is also the state of the art in this field. Gaze prediction is a key development direction for future VR rendering, which is mostly combined with saliency detection to predict the location of the viewpoint. Substantial work remains in the field, but unfortunately, only a tiny portion of the work can be achieved in real time. Finally, we present the current problems and challenges machine learning methods face. Our review of machine learning approaches in foveated rendering not only elucidates the research prospects and developmental direction but also provides insights for future researchers in choosing research direction and designing network architectures.

Key words: foveated rendering; deep learning; real-time rendering; eye fixations prediction; image reconstruction; super-resolution; ray tracing denoising

0 引　言

随着虚拟现实(virtual reality, VR)和增强现实(augmented reality, AR)技术在医疗、教育、军事、娱乐等领域的广泛应用,具有高分辨率、宽视野的头戴显示器(head-mounted display, HMD)成为一种主要的显示设备(Weier等, 2017; Mohanto等, 2022)。实现高度的沉浸感和交互性一直是虚拟现实的主要挑战之一,延迟可能导致用户出现晕眩和恶心的情况(Frank等, 1988)。当前,标准的虚拟现实帧率被确定为90 Hz (Mohanto等, 2022),而对于需要更高响应速度的交互式游戏显示器,帧率要求则需要达到120 Hz以上(Hsu等, 2017)。然而,一些研究者,例如Cuervo等人(2018)认为要实现逼真且不引起晕眩的虚拟现实沉浸感,需要将刷新率提高到1 800 Hz; Arabadzhiyska等人(2017)建议,处理360°视频流时,延迟应该在20 ms左右;同样,对于沉浸式应用程序,Koskela等人(2016)建议延迟应该保持在20 ms以下。显然,要在高分辨率和低延迟的环境下实现实时逼真渲染是一个充满挑战的目标。

解决这些挑战的一种方法是在渲染时结合注视点信息。如图1所示,由于人眼很难注意到注视区域之外的细节,因此可以根据用户的注视点来调整渲染过程,对注视区域进行高质量渲染,而在视觉的外围区域降低渲染质量。图中,圆圈内表示人眼注视点区域(fovea),圆圈外表示注视点的外围区域(periphery)。由于用户通常不会注意到外围区域的细节变化,因此该技术可以在用户感知质量没有明显损失的前提下大大提高渲染速度,降低渲染的计

算复杂性。这种将渲染过程与人眼特性相结合的技术在以往文献中有多种称呼，如凝视感应渲染（gaze-contingent rendering）或感知驱动渲染（perception-driven rendering）（Weier等，2017），本文称之为注视点渲染（foveated rendering）。

在现有的注视点渲染的研究综述中，Weier等人（2017）围绕人类视觉系统的局限性、针对局限性的建模方法以及注视点渲染方法3个问题进行探讨；Mohanto等人（2022）则详细讨论了各种注视点渲染方法，其将注视点渲染的实现技术分为自适应分辨率技术、着色简化技术、几何简化技术和时空退化技术4类，并对每一类技术中面向静态注视和动态注视的方法，以及面向光栅化和光线追踪的方法进行了概述；Wang等人（2023）则按照输入数据的类型（如图像、视频、体数据和点云等）、所采用的注视点渲染技术（如着色简化、几何简化等）和渲染方法（如光栅化、光线追踪和光子映射等）3个维度进行划分。这些综述为注视点渲染领域提供了深刻的见解和分类，有助于研究者更好地理解和应用这一关键技术。

近年来深度学习在计算机视觉、计算机图形学等领域不断取得成果，杨航等人（2023）对深度学习背景下的图像三维重建技术进行整合、论述，并总结当前图像三维重建领域存在的挑战及未来趋势。江俊君等人（2023）系统论述了基于深度学习的视频超分辨率技术进展。对于目标检测，赵永强等人（2020）对主流目标检测算法改进和优化方法进行分析，并对通用数据集进行介绍。潘晓英等人（2023）则从数据增强、多尺度特征融合、注意力机制等方面对小目标检测方法进行总结，并对其评价指标及数据集进行介绍。王自全等人（2022）从深度学习与显著性检测的角度对相关论述进行总结整理。因此，将深度学习与注视点渲染相结合会是一个很有前景的研究领域，特别是在渲染后处理方面，深度学习方法能够优化和改进注视点渲染的结果，从而显著提升注视点图像的真实性和沉浸性。然而，目前已有的综述文献都测重于对实现技术进行分类，对深度学习方法的探索尚不充分。因此，本文将重点对注视点渲染在深度学习方面的研究做分类概况：第1节介绍人类视觉系统的相关背景知识；第2节从自适应分辨率、着色简化、几何简化和硬件实现4个方面简述目前最先进的非深度学习的注视点渲染方法，总结其优缺点并与深度学习方法对比；第3节介绍本文综述的深度学习方法所使用的评估准则；第4节详细介绍使用深度学习进行注视点渲染的方法，具体细分为超分辨率、降噪、补全图像合成、视点预测和图像应用领域。最后，讨论并总结本研究的结果。

(a) 注视点图像　(b) 参考图像

图1　注视点渲染图像实例（Patney等，2016）

Fig. 1　Example of foveated rendering image (Patney et al., 2016) ((a) foveated image; (b) reference)

1　人类视觉感知背景知识

1.1　人类视觉系统

人类视觉系统（human visual system, HVS）是人类用于感知和理解视觉信息的复杂系统，其包括眼睛、大脑和其他相关的生物学组件。其中，眼睛是HVS的前沿部分，负责接收外部光线并将其转化为电信号。它包括角膜、晶状体、虹膜、视网膜和视神经等组件（Swafford等，2016）。

视网膜主要由视杆细胞和视锥细胞组成，这些细胞负责捕捉可见光谱中的不同信号，并将信号通过神经节细胞（ganglion cells）传递给大脑以形成视觉感知。视杆细胞（rods）主要分布在视网膜的外围

区域(periphery),它们对光线较暗的环境非常敏感,使人眼能在低光条件下观察物体,但它不能分辨颜色。相比之下,视锥细胞(cones)集中在视网膜的中央区域(fovea),它的跨度只有1.5°～2°(Swafford等,2016)。视锥细胞提供了更高的空间分辨率和色彩感知能力,使人能够看清细节和感知颜色。

图2显示了视杆细胞和视锥细胞在视网膜上的分布。横坐标中的"偏心率"(eccentricity)表示物体或视觉刺激相对于视网膜中央的位置,偏心率值越高,则该视觉刺激离中央区域越远。纵坐标中的"细胞密度"(density)则表示不同位置的细胞数量。可见,视网膜的中央区域包含了密集的视锥细胞,提供了比其他任何地方都高的空间和色彩分辨率,而随着偏心率的增加,细胞密度逐渐减小,意味着外围区域的视觉感知能力迅速下降。

1.2 对比敏感度函数

对比度是描述图像或视觉刺激中亮度差异或颜色差异的度量,它表示亮度或颜色之间的差异程度。而对比度的倒数通常称为对比度敏感性,它表示在特定的空间频率和时间频率下,人眼能够察觉到的最小对比度变化,较低的对比度敏感性值表明视觉系统对亮度或颜色差异更加敏感。不同的空间和时间频率会影响对比度感知的能力,在某些频率下,人眼可以察觉微小的亮度或颜色差异,而在其他频率下,需要更大的差异才能察觉到。

通过测量不同频率下的对比度敏感性,可以构建对比敏感度函数(contrast sensitivity function, CSF),这有助于了解在不同频率下人眼对图像感知的敏感性。如图3所示,Mantiuk等人(2022)将视觉刺激的所有主要方面(空间和时间频率、偏心率、亮度和面积)考虑在内,提出了一个统一的CSF函数,称为"stelaCSF"。图3从左到右分别展示了空间频率与时间频率、亮度、偏心率和尺寸的相互作用。

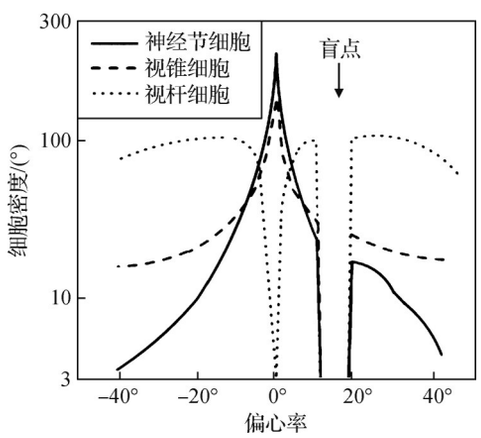

图2 视网膜中细胞的偏心率分布(Molenaar,2018)

Fig. 2 Eccentricity distribution of cells in the retina (Molenaar,2018)

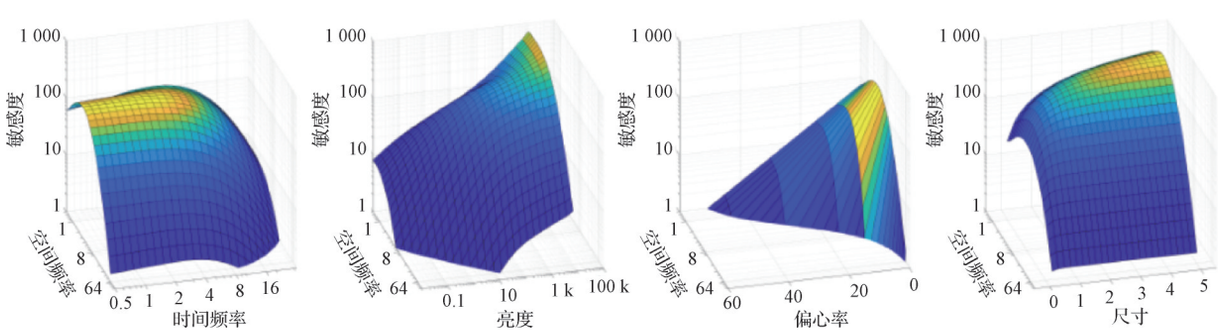

图3 对比敏感度函数的3D可视化(Mantiuk等,2022)

Fig. 3 3D visualisation of the contrast sensitivity function (Mantiuk et al., 2022)

1.3 视觉敏锐度模型

视觉敏锐度(acuity)简称为视敏度,它表示眼睛或视觉系统分辨两个紧密排列的对象或细节的能力,通常以角度或距离来衡量,即在给定条件下,人眼能够分辨的最小角度(minimum angle of resolution, MAR)或最小距离。在注视点渲染领域,通常用每度周期数(cycles per degree, CPD)作为视敏度单位。

偏心率(eccentricity)是指物体或视觉刺激相对于注视点中央区域的位置,通常以度为单位。如图4所示,偏心率越高,则距离视点中央区域越远,视敏度也越低,即分辨细节的能力越低。

Weymouth(1958)认为,在前20°～30°内,视敏度随着偏心率呈线性下降,随后偏心率越高,视觉性能下降得越快。Guenter等人(2012)将视敏度建模为偏心率的线性函数,具体为

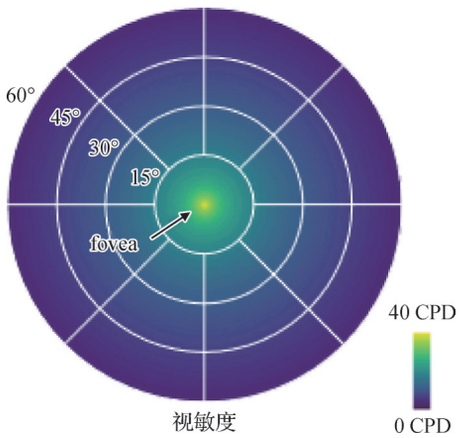

图4　视敏度与偏心率模型（Krajancich 等，2021）

Fig. 4　Model of visual acuity and eccentricity (Krajancich et al. , 2021)

$$\omega = m \times e + \omega_0 \tag{1}$$

式中，e 表示偏心率，m 表示斜率，ω_0 代表最小可分辨角度，ω 为视敏度。

Guenter 等人（2012）提出该线性模型后，许多注视点技术都基于此模型来讨论降低质量的方法。然而，除了偏心率，视敏度还受到场景内容、眼睛在明亮和黑暗区域的适应变化以及认知因素等的影响，因此，任何线性模型都只是近似。Tursun 等人（2019）指出，视敏度也受到显示内容的影响，如图 5 所示，在高对比度区域（右侧），降低图像质量会让用户很容易察觉到，但在低对比度区域（左侧），则不太容易引起注意。因此，Tursun 等人（2019）提出一种结合了局部亮度对比度等因素的新的计算模型。而 Krajancich 等人（2021）考虑了时间因素，通过实验测量和计算与偏心率相关的临界闪烁融合阈值，使模型能够预测在特定空间频率、偏心率和亮度水平范围内难以察觉的时间信息，这是独一无二的。

图5　根据底层纹理不同，注视点的不同可见性（Tursun 等，2019）

Fig. 5　Different visibility of gaze points depending on texture (Tursun et al. , 2019)

1.4　视觉拥挤

视觉拥挤（visual crowding）是指在人眼试图识别或辨认一个物体时，如果周围存在干扰性刺激，会导致人眼无法准确地感知或辨认物体的特征或细节。通常情况下，视觉拥挤更容易在注视点的外围区域中观察到，这是因为外围区域的感知容量较小，无法同时处理多个刺激。在拥挤的情况下，人眼往往无法分辨出被拥挤的物体的形状、大小、位置或其他细节信息。

如图 6 所示，以 Fridman 等人（2017）的研究为例，当视线聚焦在中间的加号时，人眼可以轻松识别左侧孤立的字母 A，但无法识别右侧两侧有其他字母的 A。观察者可能会以错误的顺序看到这些拥挤的字母，比如“BORAD”，也可能看不到字母 A，或者会以错误的顺序看到由多个字母的部分混合成的奇怪形状。

图6　视觉拥挤效应示例（Fridman 等，2017）

Fig. 6　Example of visual crowding (Fridman et al. , 2017)

视觉拥挤对视敏度的影响比偏心率或对比度还要显著，实际上，它是影响中央视觉和外围视觉之间的主要差异因素。

2　非深度学习方法概述

本节将简单归纳非深度学习领域中注视点渲染的主要技术和应用。

2.1　自适应分辨率

调整注视区域与外围区域的分辨率是注视点渲染常用的方法。Guenter 等人（2012）提出了一种代表性的方法，如图 7 所示，他们首先将视敏度建模为关于偏心率的线性函数，并根据该线性函数将图像划分为内层、中间层和外层。3 个层次均以注视点为中心，其中，内层以最高分辨率和细节层次渲染，中间层和外层则以逐渐降低的分辨率和更粗糙的细节渲染。随后这些层次进行插值合并，从而节省了一半的渲染成本。

在后续的研究中，许多方法采用了类似的多分辨率架构（Mohanto 等，2022）。Marianos（2018）根据

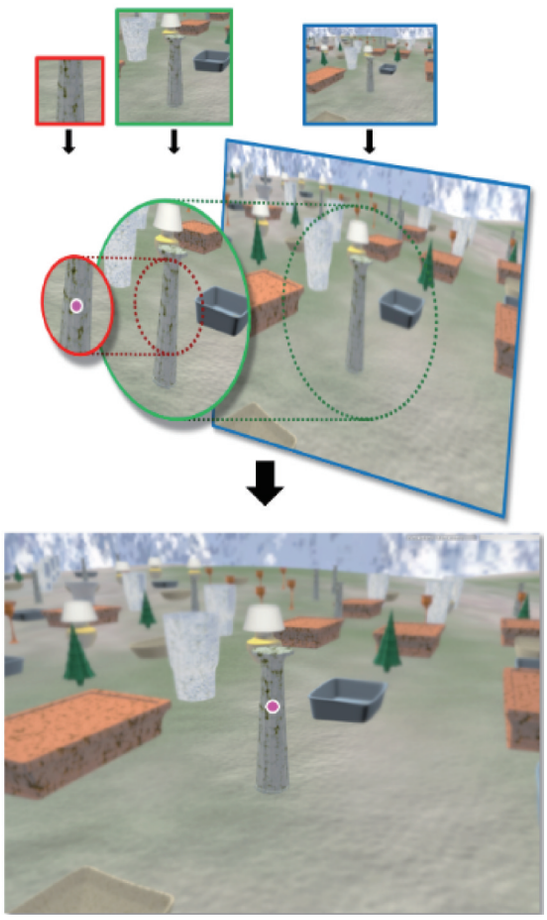

图7　多分辨率方法示意(Guenter等,2012)

Fig. 7　Illustration of multi-resolution method (Guenter et al., 2012)

视觉感兴趣区域(region of interest, ROI)与注视点的欧氏距离,将分辨率划分为3个层次,分别为100%、60%、40%。Swafford等人(2015)则采用两层分辨率,即在中央区域采用全分辨率,在外围区域采用25%的分辨率。此外,Peuhkurinen和Mikkonen(2021)还设计了一种用于混合现实环境的基于人眼分辨率的实时光线追踪方案。

2.2　着色简化

着色简化是注视点渲染的重要策略之一,该项技术通常会在注视区域使用更高精度的着色方法,而在外围区域则采用简化算法或者减少着色密度。例如,使用可变速率着色(variable-rate shading, VRS)(White等,2023;Reed,2015)和粗糙像素着色(coarse pixel shading, CPS)(Xiao等,2018;Vaidyanathan等,2014)等方法来动态地调整像素的着色密度,从而减少外围区域的着色工作。Franke等人(2021)提出了一种新方法,他们通过对之前帧进行重投影来重新利用外围区域的像素,并使用置信值检测来评估重投影可能引入的伪影,fovea区域以及低置信度的区域将被重新渲染。另外,研究人员还考虑了外围视觉的感知特点。Patney等人(2016)进行了一项用户研究,发现对外围区域进行滤波会降低对比度导致外围视野感到模糊,产生隧道视野感。但当应用后处理做对比度增强时,受试者可以容忍高达原来2倍的模糊半径,而不会察觉到与原始真实图像之间的差异。因此,Patney等人(2016)设计了一个注视点渲染系统,将着色数量减少了70%,并将粗糙着色的范围延伸至距离注视区域30°,比Guenter等人(2012)更接近注视区域。他们使用滤波来解决外围采样不足引起的锯齿问题,并使用对比度增强来帮助恢复由于滤波而降低的外围细节。如图8所示,Stengel等人(2016)设计了一个结合视敏度、眼睛运动、对比度和亮度适应的稀疏采样模型,他们使用该模型计算采样概率图,并生成采样点,以进行稀疏着色。最后,通过插值算法,他们完成了图像中缺失部分的渲染,从而将需要着色的片段数量减少了50%～80%。

2.3　几何简化

几何简化是最早与注视点渲染结合的技术(Mohanto等,2022),该技术通常通过降低外围区域的场景复杂性、调整对象细节或剔除不可见对象来实现。Horvitz和Lengyel(2013)提出了一种基于注意力模型的自适应细节级别(level of detail, LOD)生成和选择的方法,他们使用基于几何简化的成本函数和观看者注视点的概率分布,以在高分辨率网格的视觉质量和计算节省之间进行权衡。Ju等人(2019)通过在场景的低分辨率网格模型上覆盖新的高分辨率网格块,以最小的计算量实现从低分辨率到高分辨率的渐进更新。Zheng等人(2020)建立了一个感知模型来计算空间中每个三角形的视敏度,并剔除视敏度低于感知阈值的三角形。此外,该方法还通过视敏度自适应地调整曲面细分级别,使细分结果更好地满足视觉感知并提高实时渲染的计算性能。

2.4　硬件实现

注视点显示器(foveated displays)近年来已经成为一个备受关注的研究领域,它通常用于虚拟现实(VR)和增强现实(AR)系统,其主要工作原理是通过眼动追踪技术实时跟踪用户注视点,并相应地调整显示内容。Kim等人(2019)提出一种近眼增强现

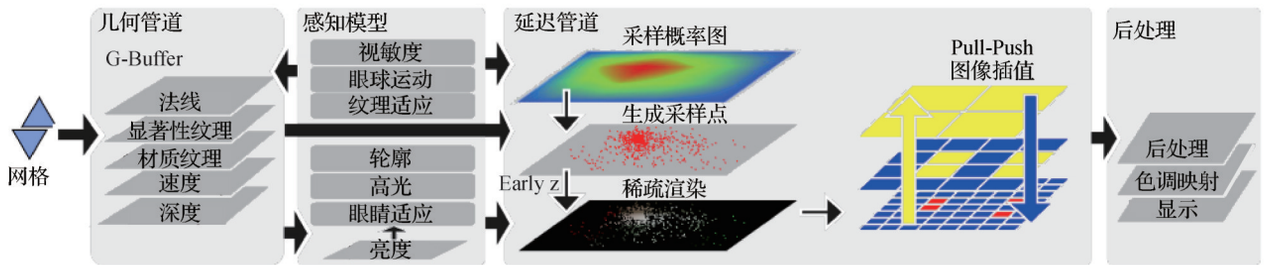

图 8　着色简化方法架构(Stengel 等,2016)

Fig. 8　Illustration of shading simplification method (Stengel et al. , 2016)

实(AR)显示技术,可以根据用户的注视点来动态调整分辨率和焦点深度。Yoo 等人(2020)使用偏振相关的双合 GP(geometric phase) 透镜和偏振复用的方法实现注视点图像的显示。Friess 等人(2021)的方法将硬件编码器与注视点编码相结合,提出了一种动态调整编码质量的方法,其通过跟踪用户的注视点,并局部调整编码器使用的每个宏块的质量参数,从而提高感兴趣区域的图像质量,同时保持总体所需带宽尽可能低。Illahi 等人(2017)同样根据当前注视点调整云游戏应用程序的视频编码器中每个宏块的质量参数,在玩家关注的宏块中增加质量,而在其他宏块中降低质量。Zare 等人(2016)将视频内容存储在两种不同的分辨率中,使用高效视频编码(high efficiency video coding, HEVC)标准将其分成多个子集,根据用户当前的视口,选择相应的分辨率。

非深度学习的方法主要关注注视点渲染的实现方式。如表 1 所述,尽管这些方法减少了渲染计算量,但不可避免地会导致图像质量下降,甚至在注视区域和外围区域之间引入视觉伪影。与此不同,深度学习方法在渲染的后处理领域有着广泛的应用,因此其可以优化和改进注视点渲染的结果,改善非深度学习方法造成的伪影,提高注视点图像的真实性和沉浸性,这是非深度学习方法所无法实现的。

3　评估准则

本节将描述对深度学习方法进行评估时所使用的评估准则。对于超分辨率、降噪、补全和图像合成方法,从结果质量、方法实时性、用户主观体验和方法的可处理对象能力 4 个方面进行对比分析。结果质量指应用深度学习方法所得到的图像与参考图像之间的差别,参考注视点图像评价指标进行评估;实时性参考网络推理所需平均时间进行评估;用户主观体验表示用户观看方法结果与观看参考图像之间的感受差异,根据方法所提供的用户研究数据与方法的结果质量进行综合评估;可处理对象能力表示该方法可以处理的数据类型,包括图像、视频、全景图像、光场等。

对于注视点预测方法,从准确性、实时性、可适用性和可处理对象能力 4 个方面进行对比分析。准确性指视点预测准确程度,参考注视点预测评价指标进行评估;实时性参考网络推理所需平均时间进行评估;可适用性参考深度学习方法的输入复杂程度进行评估,如除图像本身外,是否需要额外的眼睛、头部等辅助信息;可处理对象能力表示该方法可以处理的数据类型,包括图像、视频、全景图像、光场等。

3.1　注视点图像评价指标

1)FWQI (foveated wavelet quality index)。Wang 等人(2001)在小波变换领域提出了一种新的图像质量度量,称为注视点小波图像质量指标 (FWQI)。FWQI 综合考虑了人类视觉系统的多个因素,包括对比度敏感性函数的空间方差、局部视觉截止频率的空间方差、不同小波子带下的人眼视觉敏感度的方差,以及观看距离对显示分辨率的影响和 HVS 功能。FWQI 可用于焦点感兴趣区域(ROI)图像编码和质量增强。

2)FA-MSE(foveated mean squared error)。Rimac-Drlje 等人(2010)分析了人眼视觉系统对视频运动速度的依赖关系,即当视频内容在视网膜上的运动速度增加时,视觉敏感度会下降。他们提出了一种称为视点均方误差(FA-MSE)的评估指标,该指标更准确地考虑了注视点视觉特性和视频运动对质量感

表1　非深度学习方法总结

Table 1　Summary of non-deep learning methods

方法类别	代表算法	方法特点	优点	缺点
自适应分辨率	Guenter等人(2012)；Marianos (2018)；Swafford等人(2015)；Peuhkurinen和Mikkonen (2021)	主要在图像空间工作，根据用户的注视点调整渲染分辨率，并将不同层次的分辨率进行平滑混合。	节省了计算和传输资源。	无法避免出现伪影，需要在后处理中使用强大的抗锯齿算法。
着色简化	Patney等人(2016)；Stengel等人(2016)；Franke等人(2021)	主要通过减少每个像素的计算量来实现注视点渲染。	降低渲染的总体成本，显著减少计算时间。	在过于简化的区域容易产生抖动和伪影，且若着色质量过低，那么在物体快速运动时伪影将非常明显。
几何简化	Horvitz和Lengyel (2013)；Ju和Park (2019)；Zheng等人(2020)	主要在模型空间发挥作用，通过调整渲染的模型的几何复杂度来调整注视区域与外围区域的细节。	减少内存消耗，降低计算成本，提高渲染效率。	缺少简化范围和简化级别的统一标准。外围区域的几何模型细节水平低，很容易引入可闪烁伪影，因此几何简化技术很少单独使用，往往需要和外围模糊技术结合使用。
硬件实现	Kim等人(2019)；Yoo等人(2020)；Friess等人(2021)；Illahi等人(2017)；Zare等人(2016)	能够创建一个根据用户的注视点来调整分辨率的注视点显示器，也可以通过调整编码质量减少总体所需的带宽。	提高渲染效率，减少图像渲染和传输的延迟。	大多数显示器仅适用于简单场景，难以扩展到复杂场景。

知的影响，从而更好地预测和评估视频质量。

3）FA-SSIM（foveation-based content adaptive structural similarity index）。Rimac-Drlje等人（2011）还提出了一种新的视频质量评估指标，即基于注视点的内容自适应结构相似性指数（FA-SSIM）。该指标综合了结构相似性指数（structural similarity index，SSIM）、基于注视点的视敏度函数以及视频序列中运动引起的空间敏锐度降低的影响。

4）FovVideoVDP。Mantiuk等人（2021）提出了第1个用于评估注视点图像在空间、时间和偏心率方面的可见差异的评估指标，该指标旨在描述注视点渲染方法所产生的伪影对图像质量的影响。

3.2　注视点预测评价指标

1）AUC（area under the curve）。是通过在显著性图不同阈值下计算真阳性率与假阳性率之间的曲线下面积来评估显著性预测模型的性能。AUC值越高，表示模型的性能越好。

2）sAUC（shuffled-AUC）。是用于减轻AUC评分中的中心偏差效应的指标。它在计算AUC时对显著性图进行了打乱处理，以消除中心位置的影响，从而更公平地评估模型的性能。

3）NSS（normalized scanpath saliency）。NSS计算了在标准化显著性图中所有注视点位置上的平均值。该指标可以衡量模型对人类注视点的预测准确性。

4）Sim（similarity）。Sim计算了显著性分布和注视点分布的每个点上的最小值之和，用于评估模型的显著性分布与人类注视点分布之间的相似程度。

5）CC（linear correlation coefficient）。CC用于比较显著性图与人类注视点图之间的线性相关性。较高的CC值表示模型预测与人类注视点吻合程度较高。

6）KLD（Kullback-Leibler divergence）。KLD用于计算显著性图与人类注视点图之间的Kullback-Leibler散度，该指标表示两者之间的差异程度。

7）EMD（earth mover's distance）。EMD将真值显著性图和预测的显著性图视为两个概率分布，并衡量将一个分布转换成另一个分布的代价。

4　深度学习方法

深度学习方法在注视点渲染方面发挥了越来越重要的作用，它可以帮助提高注视点渲染的效果。本节将根据模型的作用，从超分辨率、降噪、补全、图像合成、注视点预测以及图像应用几个方面对该领域进行概述。深度学习方法脉络图如图9所示。

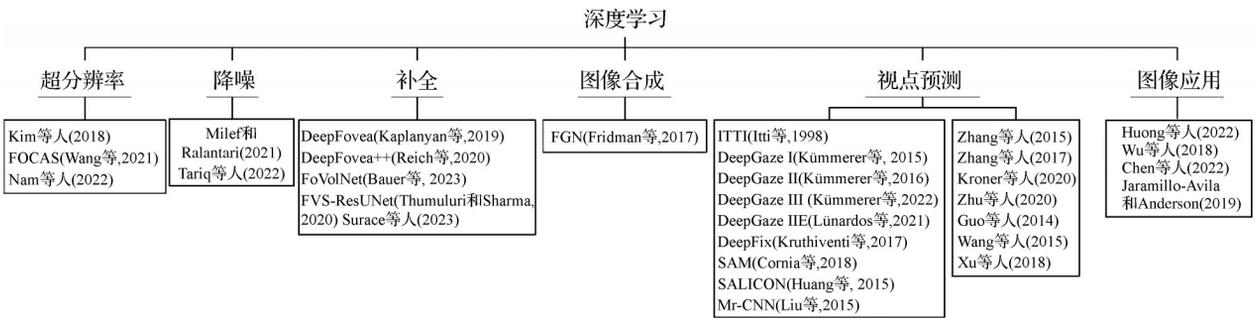

图 9　深度学习方法脉络

Fig. 9　Illustration of ddep learning methods

4.1　超分辨率

超分辨率(super-resolution，SR)是一种用于将低分辨率图像重建到高分辨率的技术。注视点超分辨率的核心思想是在进行超分辨率处理时，通过深度学习方法对用户的注视区域进行精细处理，而在外围区域，由于用户的视觉关注度较低，通常采用插值方法。

Kim 等人(2018)率先将注视点与超分辨率网络结合，提出了一个适用于虚拟现实头戴设备的注视点超分辨率方法。该方法利用眼动追踪设备和目标跟踪算法获取用户的注视区域，然后将超分辨率卷积神经网络应用于注视区域，同时在外围区域采用普通插值方法以降低计算成本。如图 10 所示，输入是一个低分辨率的图像序列(图 10(a))，利用眼动追踪设备和目标追踪算法来追踪注视区域(图 10(b))。在这个过程中，眼动追踪设备用于重新定位主要对象并获取感兴趣区域的位置信息，但文章使用的追踪设备以 60 帧/s 的速度进行定位，因此在眼动追踪设备不可用的时间段内，使用目标追踪算法进行追踪，从而提高图像处理的效率和准确性。随后，根据注视点信息对图像进行超分辨率处理。通过该算法，将渲染高分辨率图像所需的计算量降低了 90. 405 9%。

Wang 等人(2021)提出了一种基于深度学习的实时注视点超分辨率方法 FOCAS (foveated cascaded video super resolution)。该方法在注视区域分配更多的神经网络块以提供更高的图像质量，而在外围区域使用较少块以提供较低但是足够的质量。

如图 11 所示，FOCAS 的网络模型采用了循环结构。在每个循环迭代 t 中，将当前帧和前一帧的低分辨率图像 $\boldsymbol{I}_t$ 和 $\boldsymbol{I}_{t-1}$、上一帧的特征值 $\boldsymbol{H}_{t-1}$ 以及上一帧的超分辨率结果 $\boldsymbol{O}_{t-1}$ 作为输入。输入数据在经过 Unshuffle 操作后被送入一系列堆叠的残差块(res-block)。最后，模型输出当前帧特征值 $\boldsymbol{H}_t$ 和超分辨率结果 $\boldsymbol{O}_t$。每个残差块都会向特征图引入更精细的图像细节，从而提高超分辨率输出的视觉质量。因此，FOCAS 只允许注视区域的特征图能够使用更多残差块以获得更高质量，外围区域则仅使用少量残

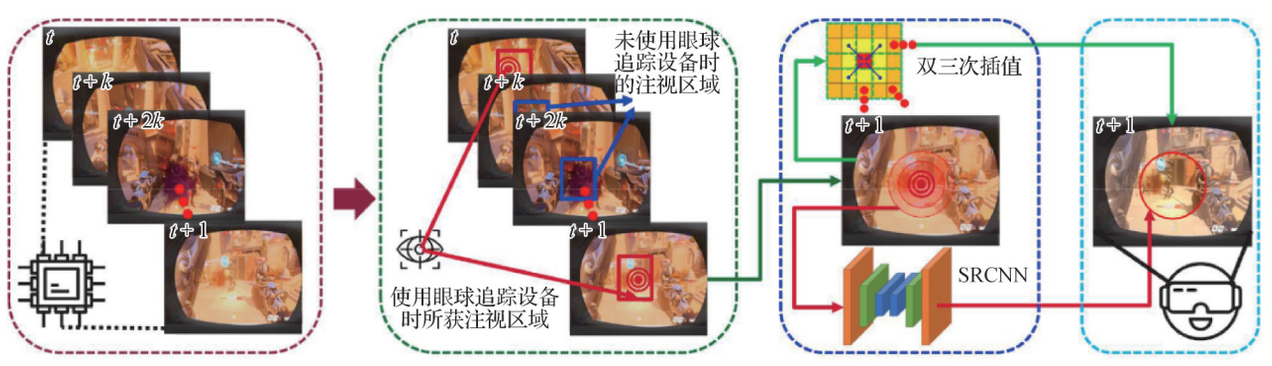

图 10　基于追踪设备的注视点超分辨率方法(Kim 等，2018)

Fig. 10　Illustration of the tracking device-based foveated super-resolution method (Kim et al. , 2018)((a) low resolution image sequence; (b) object tracking-based gazed region detection; (c) resolution conversion; (d) high resolution image sequenc)

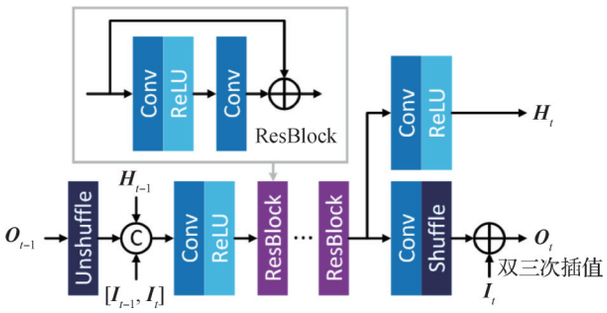

图11　FOCAS网络架构(Wang等,2021)

Fig. 11　Illustration of FOCAS network architecture (Wang et al., 2021)

差块保持适度的质量。

FOCAS方法架构图(Wang等,2021)如图12所示。FOCAS方法包括3个阶段:训练模型、优化求解和推理。在推理阶段,FOCAS实现注视点超分辨率。具体来说,FOCAS采用了低、中、高3种质量分布来分别对应外围区域、中间混合区域和注视区域,如图中阶段3)(推理阶段)所示,输入数据首先经过d_3个残差块处理得到低质量的特征,随后,从低质量特征中裁剪出大小为r_2的混合区域子特征,并将该子特征接着传递给(d_2-d_3)个残差块以获得中等质量的特征,高质量特征同理获得。经过上述步骤,FOCAS将用于外围区域的低质量特征、用于混合区域的中等质量特征和用于注视区域的高质量特征混合在一起再次卷积,输出当前帧的超分辨率结果$\boldsymbol{O}_t$和特征$\boldsymbol{H}_t$。

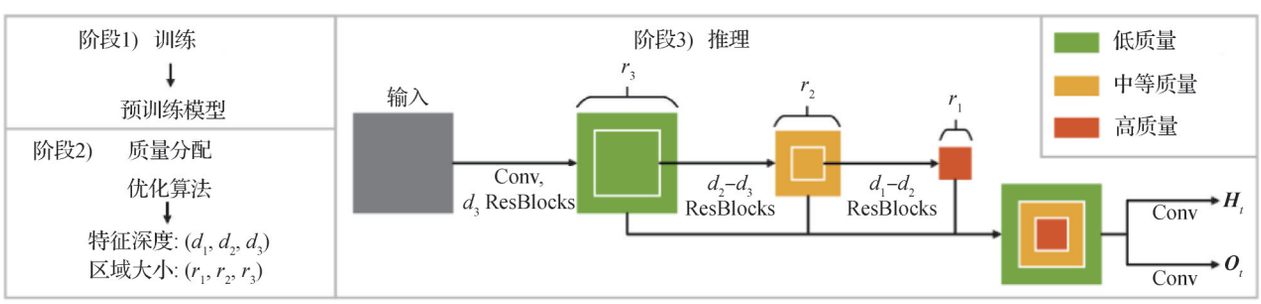

图12　FOCAS方法架构(Wang等,2021)

Fig. 12　Illustration of FOCAS method (Wang et al., 2021)

区域大小(r_1,r_2,r_3)以及残差块数量(d_1,d_2,d_3)的选择在阶段2)进行。FOCAS构建了一个优化问题并给出求解算法。经过实验证明,FOCAS可以减少50%～70%的推理时间,从而导致帧率提高2～3倍,同时获得与传统SR相当的视频质量。

与FOCAS类似,Nam等人(2022)同样在注视区域提供了更多的神经网络块以提供更高的质量。如图13所示,Nam等人(2022)首先从低分辨率输入中裁剪出注视区域,并将其传递给卷积层进行二倍的超分辨率,获得二倍超分辨率图像。随后,再次从该二倍图像中裁剪出需要进一步精细超分的区域,传入卷积层获得四倍超分图像,最后对低分辨率输入、二倍超分图像和四倍超分图像进行插值和合并。

4.2　降噪

与注视点超分辨率类似,注视点降噪通常也是通过对注视区域精细降噪、外围区域粗糙降噪来实现,但该方面的工作并不多。

Milef和Kalantari(2021)提出了一种时域稳定的支持注视点降噪的多尺度深度学习方法。该方法将输入图像和辅助特征逐步进行两倍的降采样以在不同空间尺度上分别降噪,并将不同尺度的降噪结果根据权重进行混合以获得最终结果。此外,将降噪后的历史帧重新投影后也输入网络,以提高时间稳定性。根据Guenter等人(2012)提出的偏心率模型,该降噪器在不同尺度上建模偏心率,具体为

$$e_i = \frac{s_{i+1} \times \omega^* - \omega_0}{m} \tag{2}$$

式中,s_i是尺度i的降采样因子,而其他参数采用默认设置(Guenter等,2012),即,$\omega^* = 0.0516$、$\omega_0 = 1/48$、$m = 0.022$。根据距屏幕的距离和显示器的分辨率,将以角度为单位的偏心率e_i转换为以像素为单位的偏心率r_i,这使得只有注视区域的一小部分使用原始尺度上的去噪结果,而其他区域则使用低尺度上的去噪结果。

如图14所示,Tariq等人(2022)提出了一种基于注视点渲染的噪声增强方法。该方法通过噪声增强

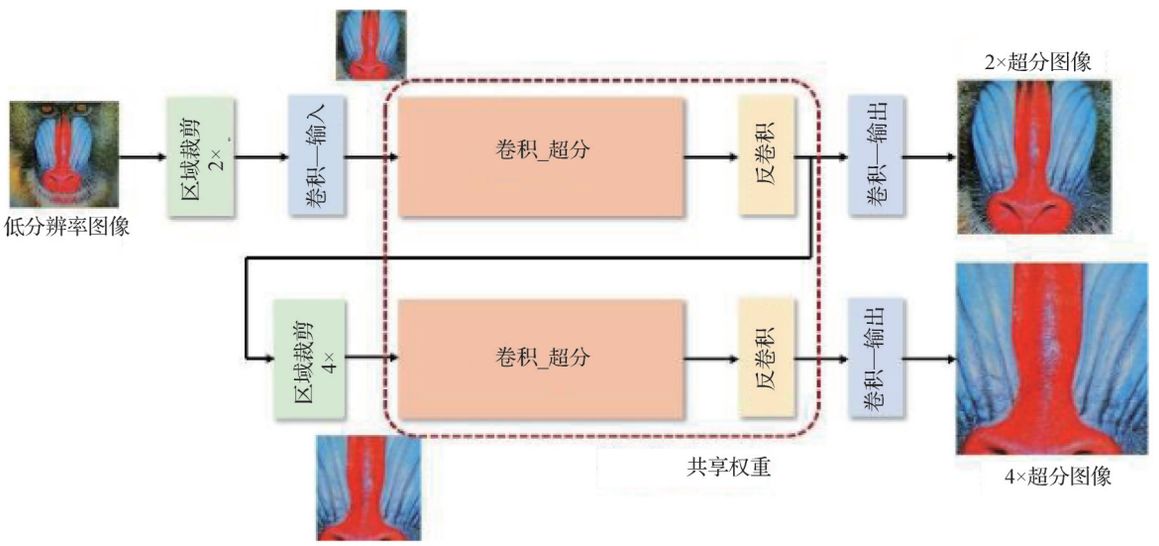

图 13　Nam 等人(2022)的网络架构

Fig. 13　Network architecture of Nam et al.（2022）

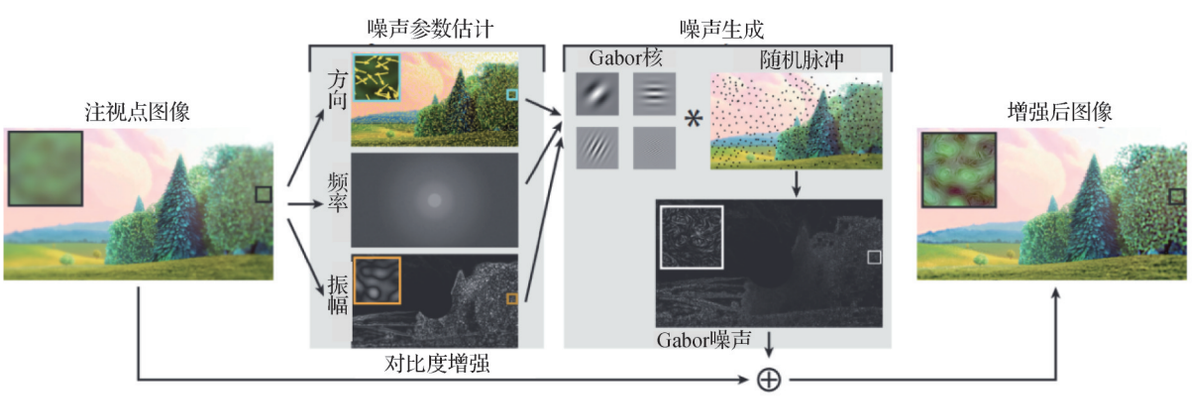

图 14　噪声增强方法的网络架构(Tariq 等，2022)

Fig. 14　Network architecture of noise enhancement method (Tariq et al. , 2022)

替代了对空间细节的准确渲染，即仅对注视点区域进行全分辨率渲染，而对外围区域进行噪声增强，以此在增强标准的注视点渲染输出的同时降低渲染成本。具体来说，该方法将注视点渲染所得的图像作为输入进行处理，通过感知启发方法估计 Gabor 噪声所对应的 3 个参数：空间频率、振幅和方向，随后，所得 Gabor 核与随机脉冲卷积合成程序噪声。同时，对输入的注视点图像进行对比度增强，以有效地提高那些已经因注视点渲染而减弱但尚未被完全移除的空间频率。然后，将 Gabor 噪声添加到经过对比度增强的注视点图像上，形成最终的增强图像输出。由于它只是一系列简单的图像处理步骤，因此适用于直接集成到注视点渲染系统中。

表 2 从结果质量、方法实时性、用户主观体验和方法的可处理对象能力 4 个方面对上述基于深度学习的注视点超分和注视点降噪方法进行了对比分析，对比分析参考了上述论文中汇报的数据。表 2 中的结果质量指应用注视点超分和注视点降噪后所得到的图像与对应目标分辨率的参考图像之间的差别，参考 SSIM、PSNR、FWQI、FovVideoVDP、FA-SSIM 等注视点图像评估指标，点数越多表示质量越高；实时性为网络推理所需平均时间，点数越高表示所需时间越少，实时性越好；用户主观体验表示用户观看方法结果与观看参考图像之间的感受差异，点数越高表示用户体验感越好；可处理对象能力表示该方法可以处理的数据类型，包括图像、视频、全景图像、光场等，点数越高表示该方法的可处理能力越强。

在注视点超分辨率的方法中，由于 Wang 等人(2021)的方法对注视区域和外围区域均使用卷积层进行超分辨率，因此所获得的结果质量最好，用户对

表2　超分方法和降噪方法对比分析

Table 2　Comparative analysis of super-resolution methods and denoising methods

类别	方法	结果质量	实时性	用户主观体验	可处理对象能力
注视点超分	Kim等人(2018)	●●●●○	●●●○○	●●●●○	●●●○○
	Wang等人(2021)	●●●●○	●●●●●	●●●●●	●●●●○
	Nam等人(2022)	●●●○○	●●●●○	●●●○○	●●●○○
注视点降噪	Milef和Kalantari (2021)	●●●●○	●●●●●	●●●○○	●●●●○
	Tariq等人(2022)	●●●●●	●●●●●	●●●●●	●●●●○

注:实点数越多质量越高。

注视区域和外围区域之间的质量差异感最小。该方法从低分辨率的每帧图像4倍超分至1 080 P仅需20 ms左右,实现了实时的性能,是目前注视点超分辨率领域最好的方法。与之不同的是,另外两种方法在外围区域采用简单的插值技术,因此结果质量略差于Wang等人(2021)的方法。而Kim等人(2018)的方法网络推理时间为0.134 8 s,并不具备很好的实时性。

在注视点降噪的方法中,Milef和Kalantari(2021)以及Tariq等人(2022)的方法都能够实现实时性能,但相对于全分辨率降噪而言,Tariq等人(2022)的结果在图像质量上存在一定损失。两种方法均能对图像和视频进行注视点降噪,具有良好的可处理对象能力。目前,针对深度学习的注视点降噪领域的研究工作并不多,因此该领域仍然具有巨大的发展潜力。

4.3　补全

鉴于人眼只能注意到注视区域的细节特征而忽视外围区域的特性,一种注视点渲染的思路是仅对注视区域进行全分辨率渲染,而外围区域进行稀疏渲染,以降低渲染计算量。然而,如图15最左侧,这种渲染方法无疑会引入大量视觉伪影,因此需要对外围的稀疏图像进行补全。

Kaplanyan等人(2019)探索了一种新颖的注视点补全方法DeepFovea,该方法采用生成对抗网络框架,从每帧提供的一小部分像素中补全出一个合理的外围区域。如图16,该生成器网络G使用了具有跳跃连接的循环编码器—解码器架构和ELU激活函数等技术来增强网络性能,为了让网络能够利用帧间相关性,该方法还将循环层添加到网络的解码器部分。通过生成器网络,输入的稀疏像素可以被重建为完整的图像。

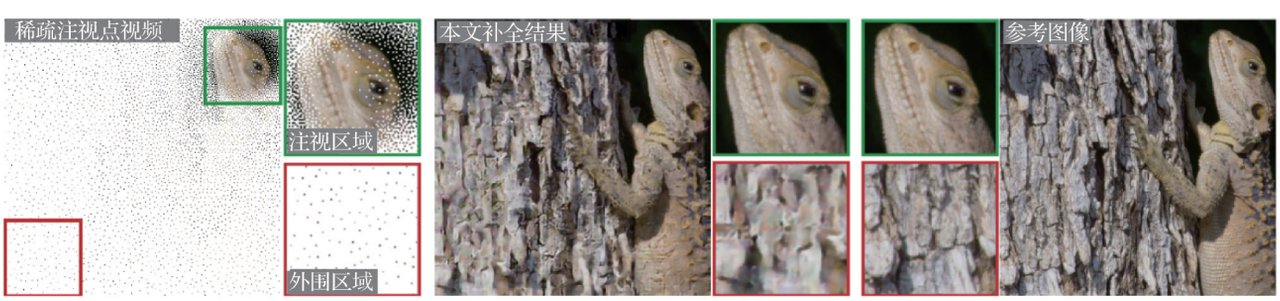

图15　DeepFovea结果图(Kaplanyan等,2019)

Fig. 15　Results of DeepFovea(Kaplanyan et al., 2019)

如图17所示,DeepFovea包含了两个鉴别器网络,分别为D1和D2。鉴别器网络D1具有3D漏斗结构,由多个残差块组成。此外,考虑到频谱信息可以用于分析视频序列中的频域特征,而自然图像的频谱通常在高频处逐渐消失,Kaplanyan等人(2019)还设计了鉴别器网络D2,即傅里叶鉴别器,该鉴别器首先对视频序列应用快速傅里叶变换(fast Fourier transform, FFT)以获得频谱信息,然后再应用鉴别器D1的架构。该方法使用对抗性损失、感知损失和光流损失来优化生成器网络G,以实时的速度生成时域稳定且没有明显感知质量下降的补全图像。

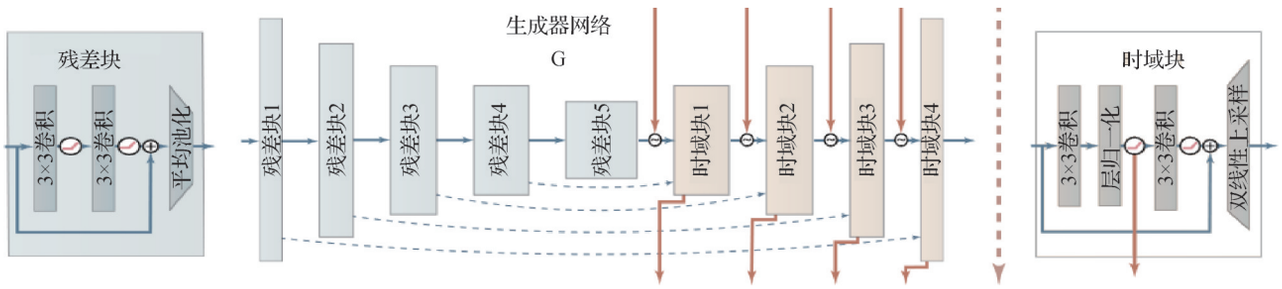

图 16　DeepFovea 生成器网络架构（Kaplanyan 等，2019）

Fig. 16　Illustration of DeepFovea generator network architecture (Kaplanyan et al. , 2019)

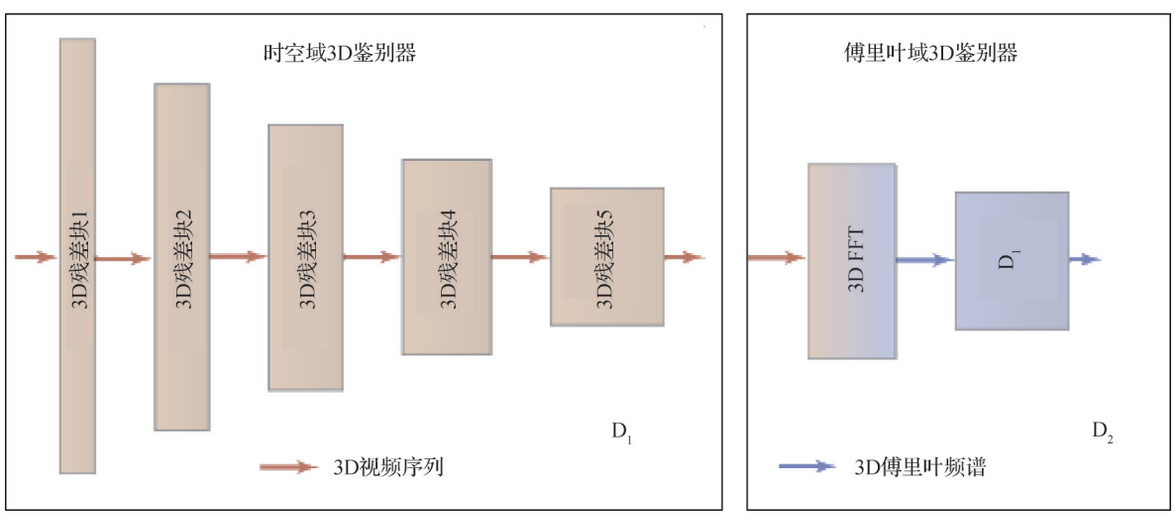

图 17　DeepFovea 鉴别器网络架构（Kaplanyan 等，2019）

Fig. 17　Illustration of DeepFovea discriminator network architecture (Kaplanyan et al. , 2019)

在 DeepFovea（Kaplanyan 等，2019）的基础上，Reich 等人（2020）设计了一个补全超分模型 DeepFovea++。如图 18 所示，该模型首先是一个与 DeepFovea 类似的循环 U-Net 网络，紧接着使用两个基于可变形卷积的超分辨率块，从而在补全图像后对图像进行超分。该方法同样使用两个鉴别器网络，第 1 个鉴别器网络接收整个 3D 视频序列，第 2 个鉴别器接收视频序列的 3D FFT 频谱。

Bauer 等人（2023）开发了一个完整的基于注视点的体渲染管道 FoVolNet，如图 19 所示，整个渲染过程由两个关键步骤组成。首先，基于注视点信息对整帧进行稀疏采样。然后，使用神经网络补全出完整帧，不同于 DeepFovea（Kaplanyan 等，2019）只使用直接预测的方法，该神经网络采用了直接预测和核预测结合的方法来实时地生成稳定且感知上可以接受的输出。

具体来说，该管道首先基于 STBN（spatio-temporal blue noise）噪声生成了用于稀疏采样的二进制遮罩，随后，将得到的稀疏图像传入补全网络以生成完整图像。如图 20 所示，该补全网络由两个 U-Net 架构

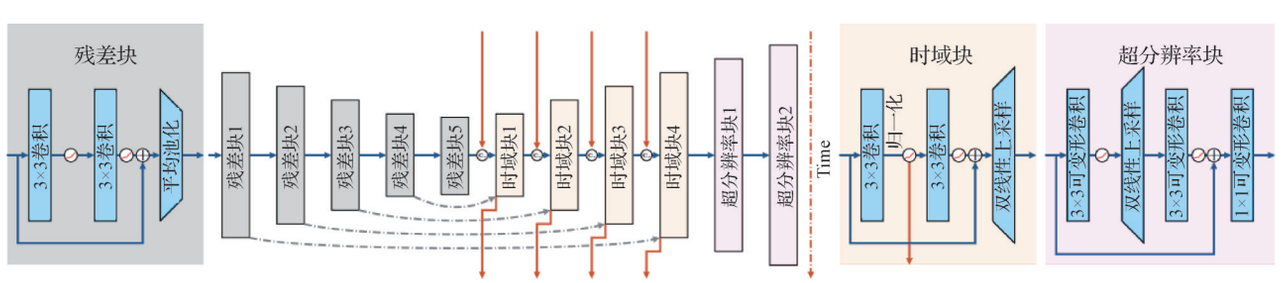

图 18　DeepFovea++网络架构（Reich 等，2020 ）

Fig. 18　Illustration of DeepFovea++ network architecture (Reich et al. , 2020)

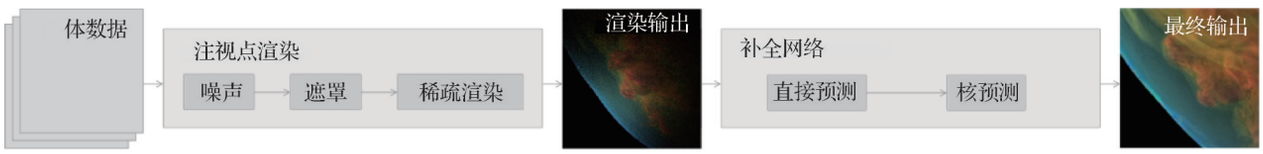

图19　FoVolNet方法架构（Bauer等，2023）

Fig. 19　Illustration of FoVolNet method (Bauer et al., 2023)

的直接预测网络和核预测网络组成，通过直接预测网络执行粗略的补全，再将粗糙补全结果传入核预测网络以获得高质量的补全结果。FoVolNet（Bauer等，2023）引入了循环连接，允许将当前时间步的输出隐藏状态（hidden state）传递回到下一个训练步骤的输入中，以此来积累关于图像序列的信息，以便更好地重建具有时间稳定性的图像序列。与DeepFovea（Kaplanyan等，2019）相比，FoVolNet提供了更高更稳定的视觉质量，且能以实时速率运行。与传统渲染相比，FoVolNet在显著节省计算时间的同时保证了感知质量。

Thumuluri和Sharma（2020）设计了一种全新的端到端卷积神经网络FVS-ResUNet（foveated view synthesis-resuNet），这是首次尝试从稀疏的RGB图像和深度信息合成整个光场。他们首先提出了一种用于稀疏渲染的对数极坐标采样方案，通过该方法，稀疏渲染只需渲染原来总像素的1.2%。随后，将稀疏图像进行插值，根据式（3）使用最近的4个像素做反距离插值，从而初步恢复出完整图像。具体为

$$p=\frac{\sum_{i=1}^{4}\frac{p_i}{d_i}}{\sum_{i=1}^{4}\frac{1}{d_i}} \tag{3}$$

式中，p是目标像素颜色值，p_i是邻域像素颜色值，d_i是目标像素到邻域像素的距离。

最后，将该图像传入FVS-ResUNet中进行进一步的补全。

如图21所示，FVS-ResUNet网络由两个主要组件构成：一个由残差块组成的高效卷积神经网络和一个编码器—解码器架构的U-Net网络。Thumuluri和Sharma（2020）指出，残差网络能够保留输入图像的高频信息，而U-Net架构能够纠正插值步骤中由于感受野较小而引入的块状结构。因此，他们提出了将残差网络和U-Net网络并行工作最后将两个网络的输出相加的方法，以平滑外围区域并减少块状效果。该方法在注视点区域获得了较高的真实性，在外围区域没有任何可感知的伪影。

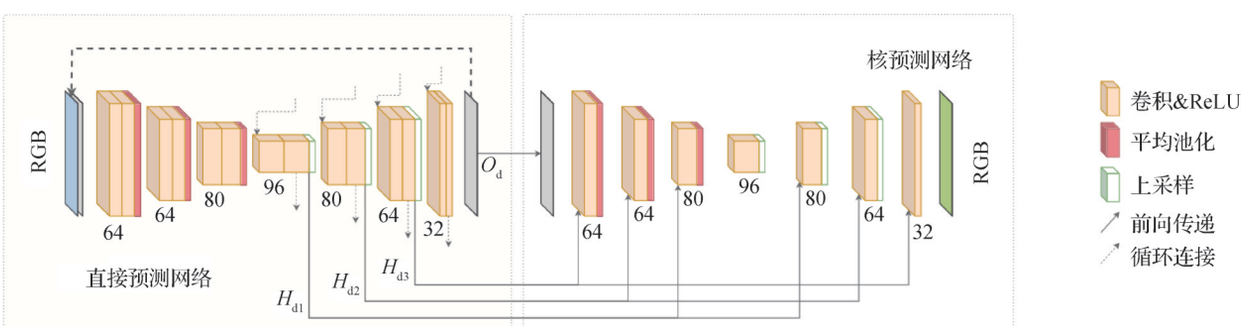

图20　FoVolNet网络架构（Bauer等，2023）

Fig. 20　Illustration of FoVolNet network architecture (Bauer et al., 2023)

Surace等人（2023）针对如何有效地指导注视点补全技术的训练进行研究。他们采用了一种创新的方法，使用包含了无法被观察者察觉的图像扭曲的数据集来为鉴别器的训练提供图像，而不是在训练时向鉴别器提供原始图像。这些数据集是通过纹理合成等方法生成的，这样做的目的是让鉴别器能够学习到与人眼视觉系统（HVS）的限制相关的数据表示。

如图22所示，该文提出的补全模型首先对稀疏图像进行插值，以初步恢复出完整图像，然后将该图像传入生成器网络进一步补全。该生成器网络的架构与Kaplanyan等人（2019）使用的DeepFovea模型

类似，鉴别器网络 D 则基于 PatchGAN 架构。

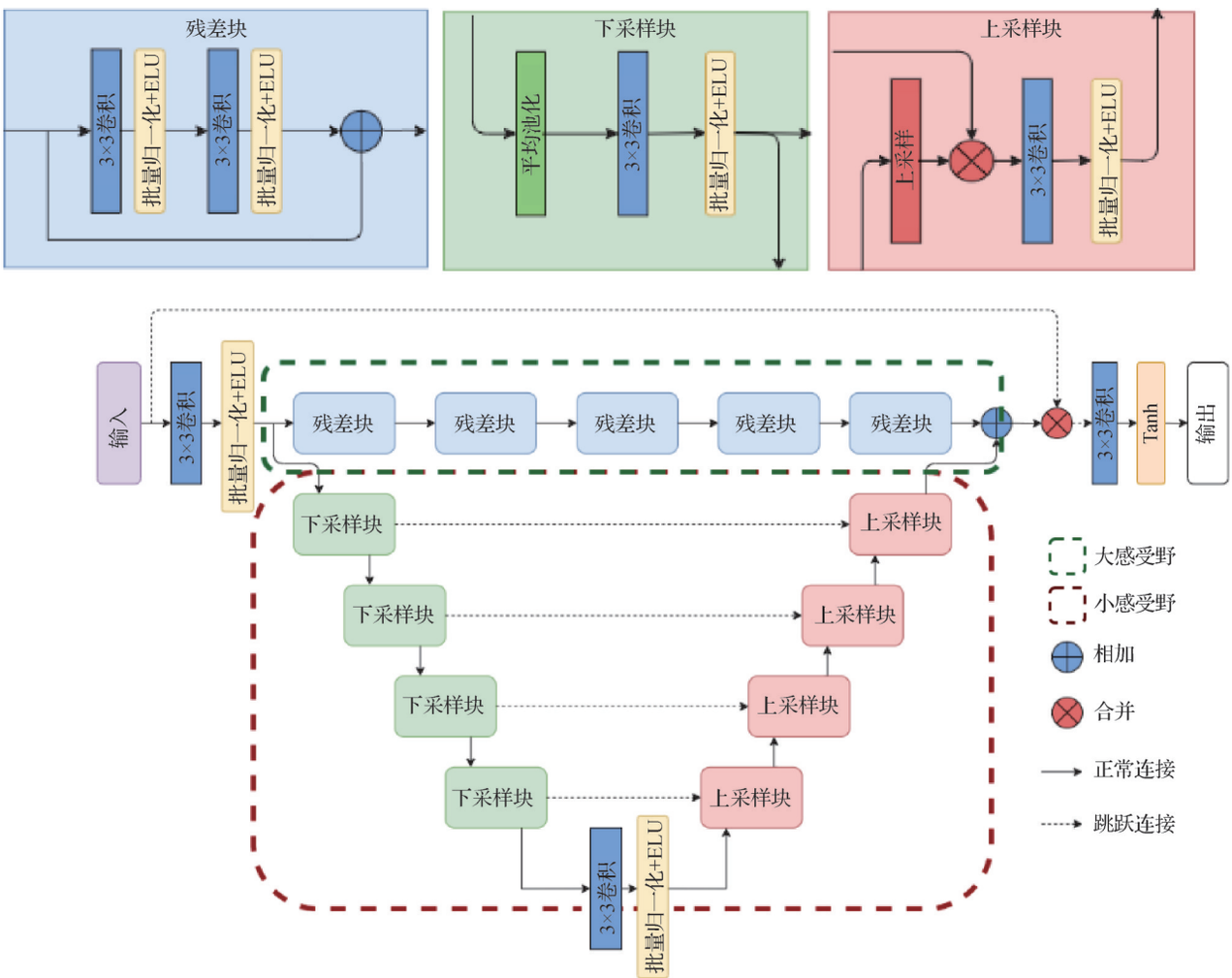

图 21　FVS-ResUNet 网络架构（Thumuluri 和 Sharma，2020）

Fig. 21　Illustration of FVS-ResUNet network architecture (Thumuluri and Sharma, 2020)

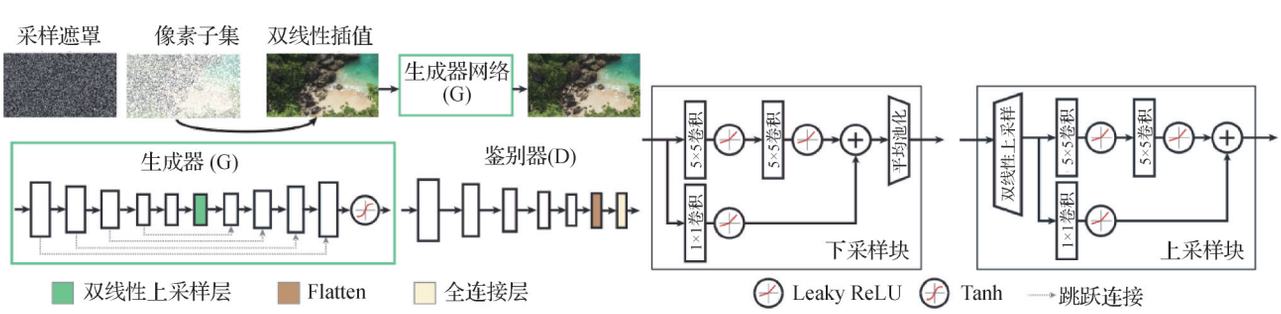

图 22　Surace 等人（2023）的网络架构

Fig. 22　Network architecture of Surace et al. (2023)

为了将感知信息传递到鉴别器网络中，Surace 等人（2023）向鉴别器提供了一组包含人眼难以察觉的结构扭曲的图像，而不是原始图像。其中一种扭曲图像如图 23 所示，通过调整高斯滤波器的 sigma 值，可以控制图像的扭曲程度，增加 sigma 值将导致图形的扭曲程度增加。Surace 等人（2023）表示，这个方法的目的并不是让生成器生成扭曲图像，而是让生成器对人类无法检测到的扭曲不敏感，从而使其专注于感知上重要的伪影。为了实现这一点，他们依赖纹理合成技术来生成扭曲图像，并使用如式

(4)的对抗损失来训练鉴别器,具体为

$$L_{\text{adv}}^{*} = D(x^{*}) - D(G(z)) \tag{4}$$

式中,z表示生成器网络G的输入,$D(x^{*})$是鉴别器网络对扭曲图像的输出,$D(G(z))$是鉴别器对G生成的图像的输出。

4.4 图像合成

与从稀疏图像补全出完整图像不同,图像合成的目标是模拟在注视条件下所看到的整体图像。基于视觉拥挤带来的信息丢失效应,Fridman等人(2017)提出了一种外围视觉模拟方法,他们开发了一个端到端的注视点生成网络FGN(foveated generative network),并设计了一个用于实时的外围视觉模拟在线工具SideEye。

如图24所示,该方法首先根据每个像素到注视点的距离生成注视遮罩,然后将该遮罩与输入图像逐元素相乘。在前向传播过程中,遮罩将被传递到每个隐藏层,并与每层的偏置进行组合,从而考虑了与注视点的关系。FGN由4个卷积层组成。TTM(texture tiling model)(Rosenholtz等,2012)在模拟注视点图像时具有出色的效果,但推理时间较长,因此该网络使用TTM的输出作为训练的参考图像。该方法是第1个将周边视觉模拟的真实性和速度结合起来的方法,在一定程度上提供了一种全新的视觉设计方法,即外围视觉可视化。

表3从结果质量、方法实时性、用户主观体验和方法的可处理对象能力4个方面对上述基于深度学习的注视点图像补全方法和图像合成方法进行了对

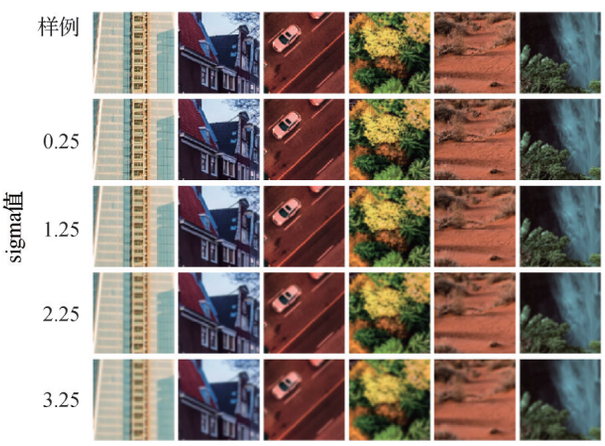

图23 Surace等人(2023)的扭曲图像示例

Fig. 23 Example of distorted image from Surace et al. (2023)

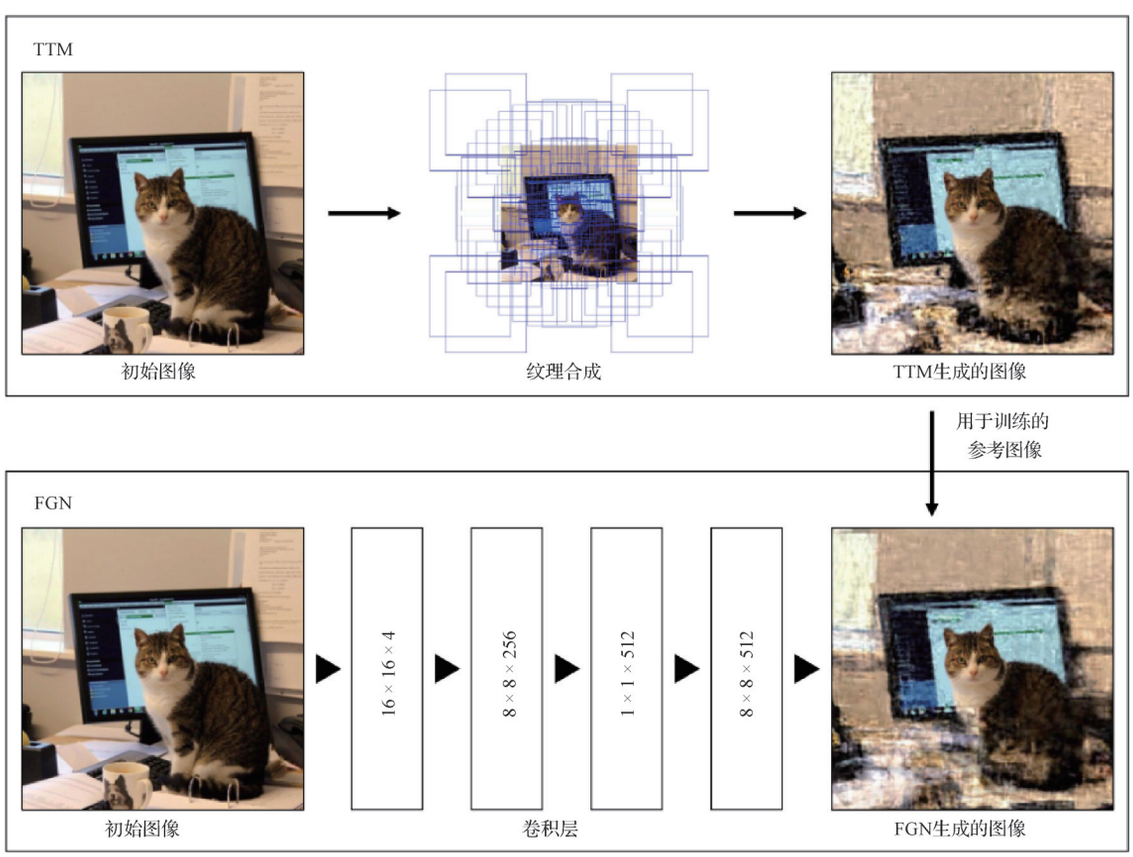

图24 FGN方法示意(Fridman等,2017)

Fig. 24 Illustration of FGN method (Fridman et al., 2017)

表 3　图像补全方法和图像合成方法对比分析

Table 3　Comparative analysis of image reconstruction methods and image synthesis methods

类别	方法	结果质量	实时性	用户主观体验	可处理对象能力
图像补全	DeepFovea（Kaplanyan 等，2019）	●●●●○	●●●●●	●●●●○	●●●●○
	DeepFovea++（Reich 等，2020）	●●●○○	●●●●●	●●●○○	●●●●○
	FoVolNet（Bauer 等，2023）	●●●●●	●●●●●	●●●●●	●●●●○
	FVS-ResUNet（Thumuluri 和 Sharma，2020）	●●●●○	●●●○○	●●●○○	●●●●○
	Surace 等人（2023）	●●●●○	●●●●○	●●●●○	●●●●○
图像合成	FGN（Fridman 等，2017）	●●●○○	●●●○○	●●●○○	●●●○○

注：实点数越多质量越高。

比分析，对比分析参考了上述论文中汇报的数据。表 3 中的结果质量指应用注视点图像补全或图像合成后所得到的图像与参考图像之间的差别，参考 SSIM、PSNR、FWQI、FovVideoVDP、FA-SSIM 等注视点图像评估指标，点数越高表示质量越高；实时性为网络推理所需平均时间，点数越高表示所需时间越少，实时性越好；用户主观体验表示用户观看方法结果与观看参考图像之间的感受差异，点数越高表示用户体验感越好；可处理对象能力表示该方法可以处理的数据类型，包括图像、视频、全景图像、光场等，点数越高表示该方法的可处理能力越强。

在注视点图像补全的方法中，DeepFovea（Kaplanyan 等，2019）和 DeepFovea++（Reich 等，2020）的网络推理时间均为几毫秒，但 DeepFovea 的结果存在一定的模糊现象，而 DeepFovea++的结果存在无法忽视的横条状视觉伪影。相比之下，FoVolNet 方法（Reich 等，2020）同样能够在实时范围内运行，并且其结果质量目前在该领域处于领先地位。FVS-ResUNet 方法（Thumuluri 和 Sharma，2020）和 Surace 等人（2023）的方法都是先对图像做插值以获得粗略的补全结果，再使用网络进行更精细的补全，从结果上来看，FVS-ResUNet 方法会导致用户感知到较为明显的视觉差异，而 Surace 等人（2023）由于改进了训练策略，因此能够获得更出色、更符合用户感知的结果。上述方法均能对图像和视频进行补全，具有足够的可处理对象能力。

在注视点图像合成的方法中，Fridman 等人（2017）的方法虽然网络架构简单，但网络推理时间约为每帧 0.7 s，并不能达到实时，且只能进行图像上的合成。

4.5　视点预测

将视点预测与注视点渲染结合是未来 VR 渲染的重要发展方向。目前的高精度视点预测方法是利用眼动仪等硬件设备来跟踪眼球运动轨迹，实时计算视线焦点。然而，这种方法存在硬件依赖性强、实施成本高、适用范围有限等问题。另一个解决方案是通过深度学习来预测视点。在基于深度学习的注视估计中，通常会根据场景中不同区域的视觉吸引力或重要性来进行视点预测，这种方法可以进一步分为自下而上和自上而下两种模型。

在自下而上模型中，通常根据场景中的低级特征，如颜色、亮度、对比度、物体运动等因素进行注视点预测。自下而上的预测是潜意识的，并不依赖于主动的注意力（Weier 等，2017），因此，此类模型常通过检测显著性来预测人类的注视点。Itti 等人（1998）的经典研究使用了多尺度的低级特征，如颜色和边缘方向，并通过神经网络将这些特征组合在一起，以预测显著性图。Kümmerer 等人（2015，2016）提出了两个深度显著性预测网络 DeepGaze I 和 DeepGaze II，分别基于 AlexNet 网络和 VGG-19（Visual Geometry Group）网络进行构建，利用预先训练的分类模型，从少数编码层中提取显著图像位置。Kruthiventi 等人（2017）提出了 DeepFix 模型，该模型由 20 个卷积层组成，通过采用 Inception 风格的卷积块来高效地捕捉可能在多个尺度上出现的对象级语义，并使用了具有非常大感受野的卷积层来捕捉对于显著性预测至关重要的全局上下文信息。

如图 25 所示，Cornia 等人（2018）提出，可以通过结合神经注意机制来预测注视点。其核心是一个结合了注意力机制的卷积长短时记忆网络 Attentive

ConvLSTM（attentive convolutional long short-term memory network），该网络专注于输入图像的最显著区域，并通过迭代细化显着性图的预测。此外，卷积层中的池化操作通常会引起特征尺度的缩小，从而降低显著性预测的性能。为应对这一问题，Cornia等人（2018）提出了针对VGG-16和ResNet-50的扩展，从而减小尺度变化的影响，同时保持图像的空间分辨率，保留详细的视觉信息。Kroner等人（2020）设计了一个编码器—解码器架构的网络，该网络包含一个卷积层模块，其具有不同的膨胀率且以并行方式捕获多尺度特征。此外，该网络还结合了多个空间尺度的语义表示，以在预测过程中融合上下文信息。

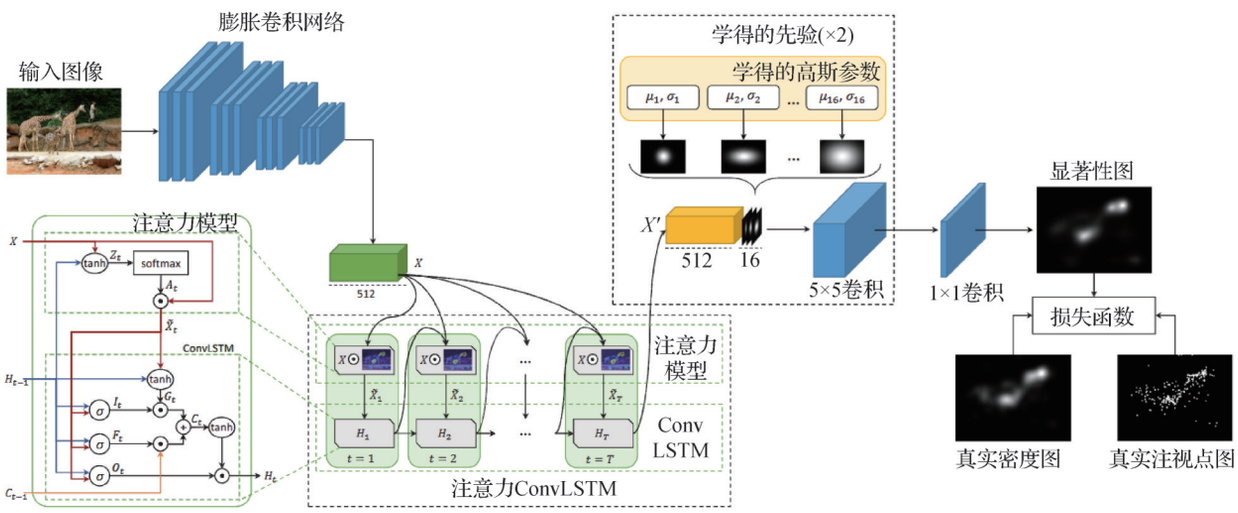

图25　Attentive ConvLSTM网络架构（Cornia等，2018）

Fig. 25　Illustration of Attentive ConvLSTM network architecture (Cornia et al., 2018)

自上而下模型是人有意识地、带有主观意图去观察图像某些部分。这些模型可能结合了先验知识、场景语义、对象识别和上下文信息等，以确定哪些区域在特定任务中最具有关注价值。因此，此类模型通常会利用眼睛的外观信息（如眼球位置和运动方向、瞳孔大小和形状等）来推断用户正在注视的区域或位置。Huang等人（2015）提出的SALICON利用预训练好的对象识别模型中高级语义的表示能力来研究视点预测，并使用显著性评估指标作为微调DNN（deep neural network）的目标，在多个尺度上合并显著性信息。Zhang等人（2015）为基于外观的注视点预测模型提供了有意义的数据集MPIIGaze，该数据集是从15位参与者在3个多月的日常笔记本电脑自然使用过程中收集的213 659幅图像，涵盖了眼球和环境的各种实际变化。

如图26所示，Zhang等人（2017）描述了一种新颖的基于外观的注视估计方法。在这个方法中，CNN（convolutional neural network）利用完整的面部图像作为输入，并在特征图上具有空间权重，以抑制或增强不同面部区域的信息。这种方法在不同的照明和极端的头部姿势下实现了高精度和稳健的性能。

目前，大多数视点预测模型会综合自上而下和自下而上两种因素来进行预测。Liu等人（2015）组合多种分辨率的信息来推断自下而上的视觉显著性，并结合眼睛注视信息以预测眼睛注视点。Nakashima等人（2015）提出了一种基于头部方向的注视预测方法，通过观察者在观看自然场景时眼睛和头部方向的经验数据，将眼睛位置和头部方向之间的关系与视觉显著性相结合来预测注视位置。Zhu等人（2020）同样将场景的显著性信息和头部运动信息结合起来，建立了一个基于头部运动和场景信息的视点预测系统，并在此基础上实现了注视点渲染，该系统能够根据预测结果为视野内的物体分配不同的渲染优先级，优先级较高的区域直接渲染，而优先级低的区域则丢弃一些像素，不进行渲染，仅通过计算相邻像素的平均值来重建。

大多数视点预测模型目前只能用于常规的图像或视频，还有研究人员研究了适用于立体图像（Guo等，2014）和视频（Wang等，2015）的视点预测方法。此外，Xu等人（2018）的工作是首次专门解决动态360°视频中的注视点预测的研究，该方法利用了现

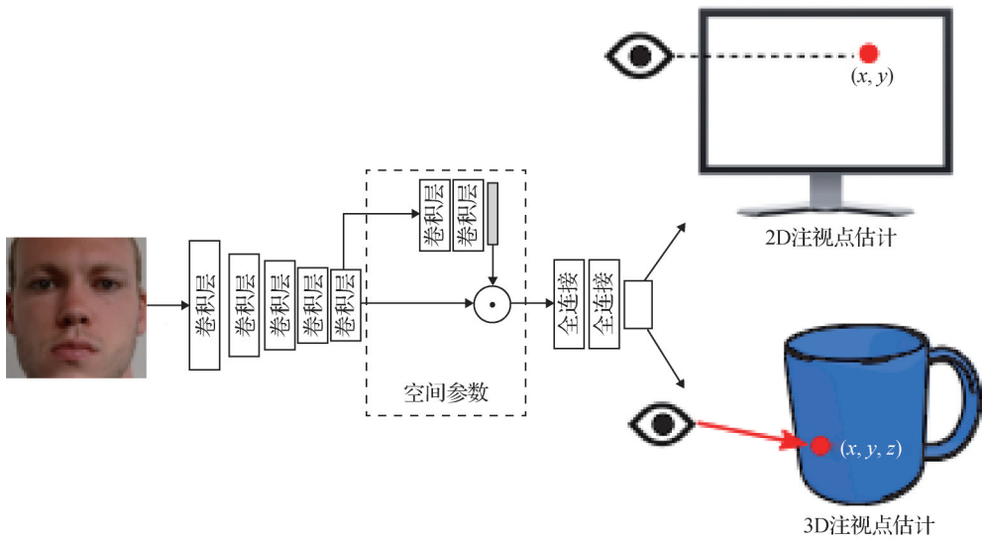

图 26　Zhang 等人（2017）的网络架构

Fig. 26　Network architecture of Zhang et al.（2017）

有的显著性模型，在不同的空间尺度上生成显著性图，并将这些显著性图与相应的图像一同输入卷积神经网络（CNN）以进行特征提取。同时，Xu 等人还使用长短期记忆（long short-term memory，LSTM）来编码用户的历史注视路径。最后，将 CNN 特征和 LSTM 特征组合起来，用于预测当前时间的注视点以及未来时间内的注视点位移。

表 4 从准确性、实时性、可适用性和可处理对象能力 4 个方面对上述基于深度学习的视点预测方法进行对比分析，对比分析参考了上述论文中汇报的数据。表 4 中的准确性指视点预测准确程度，参考 AUC、sAUC、NSS、Sim、CC、KLD、EMD 等注视点预测评估指标，点数越高表示准确性越高；实时性为深度学习网络进行视点预测实际所需平均时间，点数越高表示所需时间越少，实时性越好；可适用性表示深度学习输入复杂度，如是否需要输入头部信息、眼部信息等，点数越高表示网络所需输入复杂度越低，可适用性越高；可处理对象能力表示该方法可以处理的数据类型，包括图像、视频、全景图像和光场等，点数越高表示该方法的可处理能力越强。

在视点预测方法中，准确性方面普遍未能达到接近人类感知视觉的实际水准，存在一定的进步空间。但是相比较而言，DeepGaze 系列（Kümmerer 等，2015，2016，2022）、DeepFix（Kruthiventi 等，2017）、SAM-VGG（Cornia 等，2018）、SAM-ResNet（Cornia 等，2018）等均有不错表现。实时性方面，基于深度学习的视点预测普遍存在所需时间长、实时性能较差的问题，但其中仍有部分模型表现良好。根据论文描述及汇报数据，Kroner 等人（2020）提出的方法可以达到实时，Mr-CNN（Liu 等，2015）所需平均时间达到 14 s，难以应用到实时渲染中。SALICON（Huang 等，2015）与 Mr-CNN 相比，在网络更大的情况下，通过全卷积架构，提高推理时间至 0.27 s。Zhu 等人（2020）提出方法可以达到 1.2 ms，具有比较优越的实时性能。可适用性方面，上述方法普遍表现优异，绝大多数方法仅需输入图像即可进行处理，部分方法如 DeepGazeIII（Kümmerer 等，2022）还需输入先前注视信息。Zhu 等人（2020）提出方法需要通过设备传感器提前收集运动信息，包括速度、加速度等，并从场景截图中计算显著性信息，输入复杂性相对较高。可处理对象能力方面，绝大多数基于深度学习的视点预测方法停留在处理图像阶段，对实际应用领域未进行进一步扩展。相比之下。Guo 等人（2014）提出方法可处理立体图像，Wang 等人（2015）提出方法可处理视频，Xu 等人（2018）提出方法可对 360°沉浸式视频进行处理，均具有较高的泛化能力。

综合来看，结合各方面影响因素和评价指标，DeepGaze IIE（Linardos 等，2021）和 Kroner 等人（2020）提出模型具有较优性能，在准确性、实时性、可适用性和可处理对象能力方面均有突出表现。

4.6　图像应用

除了前述的研究领域，还有一些研究将深度学习用于注视点图像的应用领域。例如，Huong 等人（2022）构建的基于图卷积网络（graph convolutional

表4　视点预测方法对比分析

Table 4　Comparative analysis of gaze prediction methods

类别	方法	准确性	实时性	可适用性	可处理对象能力
视点预测	ITTI(Itti等,1998)	●●○○○	●●○○○	●●●●●	●●●○○
	DeepGazeI(Kümmerer等,2015)	●●●○○	●●●●●	●●●●●	●●●○○
	DeepGazeII (Kümmerer等,2016)	●●●○○	●●●●●	●●●●●	●●●○○
	DeepGazeIII (Kümmerer等,2022)	●●●●○	●●●●●	●●●●○	●●●○○
	DeepGazeIIE (Linardos等,2021)	●●●●○	●●●●●	●●●●●	●●●○○
	DeepFix(Kruthiventi等,2017)	●●●●○	●●●●○	●●●●●	●●●○○
	SAM-VGG(Cornia等,2018)	●●●●○	●●●○○	●●●●●	●●●○○
	SAM-ResNet(Cornia等,2018)	●●●●○	●●●○○	●●●●●	●●●○○
	Kroner等人(2020)	●●●●○	●●●●●	●●●●●	●●●○○
	SALICON(Huang等,2015)	●●●●○	●●●●○	●●●●●	●●●○○
	Zhang等人(2015)	●●●○○	●●○○○	●●●●○	●●●○○
	Zhang等人(2017)	●●●●○	●●○○○	●●●●●	●●●○○
	Mr-CNN(Liu等,2015)	●●●○○	●●○○○	●●●●●	●●●○○
	Nakashima等人(2015)	●●●○○	●●○○○	●●●●●	●●●●○
	Zhu等人(2020)	●●●●○	●●●●○	●●●○○	●●●○○
	Guo等人(2014)	●●●○○	●●○○○	●●●●●	●●●●○
	Wang等人(2015)	●●●○○	●●●●○	●●●●●	●●●●○
	Xu等人(2018)	●●●○○	●●●●○	●●●●○	●●●●●

注:实点数越多质量越高。

network,GCN)的质量模型是首个使用深度学习算法来评估360°注视点图像质量的模型,该模型能够自动有效地学习每个像素对360°注视点图像整体感知质量的贡献。Wu等人(2018)提出了一种卷积神经网络,首次将视线信息应用到动态视频摘要领域。Chen等人(2022)创建了一个基于注视点的端到端深度学习视频压缩框架,引入了FGU(foveation generator unit)来生成注视点遮罩,以指导压缩过程,在保持视觉质量的前提下提高了压缩效率。Jaramillo-Avila和Anderson(2019)还提出,在进行目标检测和识别时,使用注视点采样来减小图像的尺寸,从而减少卷积数量,提高处理速度。

5　结　语

注视点渲染是一项根据用户的视点来分配计算资源的技术,提高注视点附近图像区域的渲染质量,同时降低外围区域的渲染质量以节省计算资源。这在虚拟现实(VR)、增强现实(AR)、实时渲染和人机交互等领域具有重要意义。本文综述了注视点渲染中使用深度学习的主要方法,着重介绍了超分辨率、降噪、补全、图像合成、视点预测和注视点图像应用的相关研究。

注视点超分辨率方法的核心是对注视区域使用网络模型进行高精度的超分辨率处理,外围使用简单的插值或更少的网络块进行粗略的超分辨率处理。注视点降噪方法同样将降噪分为精细降噪和粗糙降噪以降低工作的计算复杂度。而图像补全方法则致力于通过深度学习的技术对基于注视点稀疏渲染所得到的稀疏图像进行补全,从而生成完整图像。图像合成的目的是模拟注视条件下观察到的整体图像,这项工作有助于设计出更具可识别性的广告标志等。注视点预测是通过显著性检测或结合外观信息如头部、眼部姿态预测注视点,它也是实现注视点渲染的基础。

对于未来的工作,本文总结了一些基于深度学

习的注视点渲染方法的主要挑战和研究方向：

1)实时性。虚拟现实和增强现实等应用通常要求非常高的实时性，用户的头部和眼部运动会导致注视点的迅速变化。因此，深度学习在注视点渲染中的实时性是一个重要挑战，尤其是在输入为高分辨率的情况下。目前，特别是在视点预测领域，大部分工作尚不能实现实时性。

2)时域稳定性。深度学习方法在处理视频输入时可能导致时域不稳定的结果，这可能表现为视频中的伪影或其他问题。未来的研究可以探索通过改进深度学习方法以生成更稳定的视频渲染结果，特别是在头部和眼部运动频繁的情况下，例如增加辅助信息作为网络输入以补充更多场景细节。

3)自适应采样。在没有眼球追踪设备的情况下，大多数基于深度学习的图像重建工作(如超分辨率、降噪、补全)都是通过随机生成注视点的方式来模拟视点，但这可能不足以满足实际应用需求。未来的研究可以考虑结合视点预测模型，使用显著性、纹理、物体轮廓等因素引导注视点的采样，以提高模型准确性。

4)尽管深度学习在注视点渲染中取得了不错的进展，但在降噪领域仍然存在较大的发展空间。未来的工作可以更多考虑注视点降噪方向，以提高注视点渲染结果的质量。

5)多项技术结合。目前，大多数深度学习方法仍是单一任务目标，但实际的图像优化需要采用多种不同方法。因此，可以考虑将注视点渲染与多种技术相结合，例如注视点渲染与降噪、超分辨率一体化网络，从而在提高结果质量的同时显著缩短处理时间。

总的来说，深度学习在注视点渲染中的应用为虚拟环境的感知质量和计算效率提供了显著的提升，这对于虚拟现实和增强现实等领域的发展具有重要作用，但是这部分工作与非深度学习方法相比仍有很大发展创新空间。未来的研究工作需要平衡图像质量和网络实时性，以能够在实时的速率下生成没有明显感知损失的图像，这将是实时渲染和人机交互领域的新方向和新挑战。

参考文献(References)

Arabadzhiyska E, Tursun O T, Myszkowski K, Seidel H P and Didyk P. 2017. Saccade landing position prediction for gaze-contingent rendering. ACM Transactions on Graphics, 36(4): #50 [DOI: 10.1145/3072959.3073642]

Bauer D, Wu Q and Ma K L. 2023. FoVolNet: fast volume rendering using foveated deep neural networks. IEEE Transactions on Visualization and Computer Graphics, 29(1): 515-525 [DOI: 10.1109/TVCG.2022.3209498]

Chen M X, Webb R and Bovik A C. 2022. Foveated MOVI-Codec: foveation-based deep video compression without motion//2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP). Nafplio, Greece: IEEE: 1-5 [DOI: 10.1109/IVMSP54334.2022.9816221]

Cornia M, Baraldi L, Serra G and Cucchiara R. 2018. Predicting human eye fixations via an LSTM-based saliency attentive model. IEEE Transactions on Image Processing, 27(10): 5142-5154 [DOI: 10.1109/TIP.2018.2851672]

Cuervo E, Chintalapudi K and Kotaru M. 2018. Creating the perfect illusion: what will it take to create life-like virtual reality headsets?// Proceedings of the 19th International Workshop on Mobile Computing Systems and Applications. Tempe, USA: ACM: 7-12 [DOI: 10.1145/3177102.3177115]

Frank L H, Casali J G and Wierwille W W. 1988. Effects of visual display and motion system delays on operator performance and uneasiness in a driving simulator. Human Factors, 30(2): 201-217 [DOI: 10.1177/001872088803000207]

Franke L, Fink L, Martschinke J, Selgrad K and Stamminger M. 2021. Time-warped foveated rendering for virtual reality headsets. Computer Graphics Forum, 40(1): 110-123 [DOI: 10.1111/cgf.14176]

Fridman L, Jenik B, Keshvari S, Reimer B, Zetzsche C and Rosenholtz R. 2017. SideEye: a generative neural network based simulator of human peripheral vision [EB/OL]. [2023-09-25]. https://arxiv.org/pdf/1706.04568.pdf

Friess F, Braun M, Bruder V, Frey S, Reina G and Ertl T. 2021. Foveated encoding for large high-resolution displays. IEEE Transactions on Visualization and Computer Graphics, 27(2): 1850-1859 [DOI: 10.1109/TVCG.2020.3030445]

Guenter B, Finch M, Drucker S, Tan D and Snyder J. 2012. Foveated 3D graphics. ACM Transactions on Graphics, 31(6): #164 [DOI: 10.1145/2366145.2366183]

Guo F, Shen J B and Li X L. 2014. Learning to detect stereo saliency// Proceedings of 2014 IEEE International Conference on Multimedia and Expo (ICME). Chengdu, China: IEEE: 1-6 [DOI: 10.1109/ICME.2014.6890321]

Horvitz E J and Lengyel J. 2013. Perception, attention, and resources: a decision-theoretic approach to graphics rendering [EB/OL]. [2023-09-25]. https://arxiv.org/pdf/1302.1547.pdf

Hsu C F, Chen A, Hsu C H, Huang C Y, Lei C L and Chen K T. 2017. Is foveated rendering perceivable in virtual reality?: Exploring the efficiency and consistency of quality assessment methods//Proceed-

ings of the 25th ACM International Conference on Multimedia. Mountain View, USA: ACM: 55-63 [DOI: 10.1145/3123266.3123434]

Huang X, Shen C Y, Boix X and Zhao Q. 2015. SALICON: reducing the semantic gap in saliency prediction by adapting deep neural networks//Proceedings of 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE: 262-270 [DOI: 10.1109/ICCV.2015.38]

Huong T T, Ha D T, Tran H T T, Viet N D, Tien B D, Thanh N H, Thang T C and Nam P N. 2022. An effective foveated 360° image assessment based on graph convolution network. IEEE Access, 10: 98165-98178 [DOI: 10.1109/access.2022.3204766]

Illahi G, Siekkinen M and Masala E. 2017. Foveated video streaming for cloud gaming//Proceedings of the 19th IEEE International Workshop on Multimedia Signal Processing (MMSP). Luton, UK: IEEE: 1-6 [DOI: 10.1109/MMSP.2017.8122235]

Itti L, Koch C and Niebur E. 1998. A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(11): 1254-1259 [DOI: 10.1109/34.730558]

Jaramillo-Avila U and Anderson S R. 2019. Foveated image processing for faster object detection and recognition in embedded systems using deep convolutional neural networks//Martinez-Hernandez U, Vouloutsi V, Mura A, Mangan M, Asada M, Prescott T J and Verschure P F M J, eds. 8th International Conference on Biomimetic and Biohybrid Systems. Nara, Japan: Springer: 193-204 [DOI: 10.1007/978-3-030-24741-6_17]

Jiang J J, Cheng H, Li Z Y, Liu X M and Wang Z Y. 2023. Deep learning based video-related super-resolution technique: a survey. Journal of Image and Graphics, 28(7): 1927-1964 (江俊君, 程豪, 李震宇, 刘贤明, 王中元. 2023. 深度学习视频超分辨率技术综述. 中国图象图形学报, 28(7): 1927-1964) [DOI: 10.11834/jig.220130]

Ju Y G and Park J H. 2019. Foveated computer-generated hologram and its progressive update using triangular mesh scene model for near-eye displays. Optics Express, 27(17): 23725-23738 [DOI: 10.1364/OE.27.023725]

Kaplanyan A S, Sochenov A, Leimkühler T, Okunev M, Goodall T and Rufo G. 2019. DeepFovea: neural reconstruction for foveated rendering and video compression using learned statistics of natural videos. ACM Transactions on Graphics, 38(6): #212 [DOI: 10.1145/3355089.3356557]

Kim J, Jeong Y, Stengel M, Akşit K, Albert R, Boudaoud B, Greer T, Kim J, Lopes W, Majercik Z, Shirley P, Spjut J, McGuire M and Luebke D. 2019. Foveated AR: dynamically-foveated augmented reality display. ACM Transactions on Graphics, 38(4): #99 [DOI: 10.1145/3306346.3322987]

Kim S, Seo M W, Lee S J and Kang S J. 2018. Object tracking-based foveated super-resolution convolutional neural network for head mounted display//SIGGRAPH Asia 2018 Posters. Tokyo, Japan: ACM: #60 [DOI: 10.1145/3283289.3283325]

Koskela M, Viitanen T, Jääskeläinen P and Takala J. 2016. Foveated path tracing: a literature review and a performance gain analysis//Bebis G, Boyle R, Parvin B, Koracin D, Porikli F, Skaff S, Entezari A, Min J Y, Iwai D, Sadagic A, Scheidegger C and Isenberg T, eds. 12th International Symposium on Advances in Visual Computing. Las Vegas, USA: Springer International Publishing: 723-732 [DOI: 10.1007/978-3-319-50835-1_65]

Krajancich B, Kellnhofer P and Wetzstein G. 2021. A perceptual model for eccentricity-dependent spatio-temporal flicker fusion and its applications to foveated graphics. ACM Transactions on Graphics, 40(4): #47 [DOI: 10.1145/3450626.3459784]

Kroner A, Senden M, Driessens K and Goebel R. 2020. Contextual encoder-decoder network for visual saliency prediction. Neural Networks, 129: 261-270 [DOI: 10.1016/j.neunet.2020.05.004]

Kruthiventi S S S, Ayush K and Babu R V. 2017. DeepFix: a fully convolutional neural network for predicting human eye fixations. IEEE Transactions on Image Processing, 26(9): 4446-4456 [DOI: 10.1109/TIP.2017.2710620]

Kümmerer M, Bethge M and Wallis T S A. 2022. DeepGaze III: modeling free-viewing human scanpaths with deep learning. Journal of Vision, 22(5): #7 [DOI: 10.1167/jov.22.5.7]

Kümmerer M, Theis L and Bethge M. 2015. Deep gaze I: boosting saliency prediction with feature maps trained on ImageNet [EB/OL]. [2023-09-25]. https://arxiv.org/pdf/1411.1045.pdf

Kümmerer M, Wallis T S A and Bethge M. 2016. DeepGaze II: reading fixations from deep features trained on object recognition [EB/OL]. [2023-09-25]. https://arxiv.org/pdf/1610.01563.pdf

Linardos A, Kummerer M, Press O and Bethge M. 2021. DeepGaze IIE: calibrated prediction in and out-of-domain for state-of-the-art saliency modeling//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE: 12899-12908 [DOI: 10.1109/ICCV48922.2021.01268]

Liu N, Han J W, Zhang D W, Wen S F and Liu T M. 2015. Predicting eye fixations using convolutional neural networks//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE: 362-370 [DOI: 10.1109/CVPR.2015.7298633]

Mantiuk R K, Ashraf M and Chapiro A. 2022. stelaCSF: a unified model of contrast sensitivity as the function of spatio-temporal frequency, eccentricity, luminance and area. ACM Transactions on Graphics, 41(4): #145 [DOI: 10.1145/3528223.3530115]

Mantiuk R K, Denes G, Chapiro A, Kaplanyan A, Rufo G, Bachy R, Lian T and Patney A. 2021. FovVideoVDP: a visible difference predictor for wide field-of-view video. ACM Transactions on Graphics, 40(4): #49 [DOI: 10.1145/3450626.3459831]

Marianos N X. 2018. Foveated Rendering Algorithms Using Eye-Tracking Technology in Virtual Reality. Chania, Greece: Techni-

cal University of Crete: 1-113

Milef N and Kalantari N. 2021. Foveated monte-carlo denoising//Proceedings of 2021 Special Interest Group on Computer Graphics and Interactive Techniques Conference Posters. Virtual Event, USA: ACM: #33 [DOI: 10.1145/3450618.3469140]

Mohanto B, Islam A T, Gobbetti E and Staadt O. 2022. An integrative view of foveated rendering. Computers and Graphics, 102: 474-501 [DO18. Towa I: 10.1016/j.cag.2021.10.010]

Molenaar E N. 20rds real-time ray tracing through foveated rendering [EB/OL]. [2023-09-25]. https://studenttheses. uu. nl/bitstream/handle/20.500.12932/28691/Thesis%20Erik%20Molenaar.pdf?sequence=2andisAllowed=y

Nakashima R, Fang Y, Hatori Y, Hiratani A, Matsumiya K, Kuriki I and Shioiri S. 2015. Saliency-based gaze prediction based on head direction. Vision Research, 117: 59-66 [DOI: 10.1016/j. visres. 2015.10.001

Nam H, Kang H and Cho H. 2022. 65-4: foveated super resolution network for virtual reality head-mounted displays. SID Symposium Digest of Technical Papers, 53(1): 869-872 [DOI: 10.1002/sdtp. 15631]

Pan X Y, Jia N X, Mu Y Z and Gao X R. 2023. Survey of small object detection. Journal of Image and Graphics, 28(9): 2587-2615 (潘晓英, 贾凝心, 穆元震, 高炫蓉. 2023. 小目标检测研究综述. 中国图象图形学报, 28(9): 2587-2615) [DOI: 10.11834/jig. 220455]

Patney A, Salvi M, Kim J, Kaplanyan A, Wyman C, Benty N, Luebke D and Lefohn A. 2016. Towards foveated rendering for gaze-tracked virtual reality. ACM Transactions on Graphics, 35(6): #179 [DOI: 10.1145/2980179.2980246]

Peuhkurinen A and Mikkonen T. 2021. Real-time human eye resolution ray tracing in mixed reality//Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications. Virtual Event: SciTePress: 169-176 [DOI: 10.5220/0010205701690176]

Reed N. 2015. VR direct: how NVIDIA technology is improving the VR experience [EB/OL]. [2023-09-25]. https://developer. nvidia. com/sites/default/files/akamai/gameworks/vr/GameWorks_VR_2015_Final_handouts.pdf

Reich C, Memmel M and Grebe J H. 2020. DeepFovea++: reconstruction and super-resolution for natural foveated rendered videos [EB/OL]. [2023-09-25]. https://github.com/ChristophReich1996/DeepFoveaPP_for_Video_Reconstruction_and_Super_Resolution

Rimac-Drlje S, Martinović G and Zovko-Cihlar B. 2011. Foveation-based content Adaptive Structural Similarity index//Proceedings of the 18th International Conference on Systems, Signals and Image Processing. Sarajevo, Bosnia and Herzegovina: IEEE: 1-4

Rimac-Drlje S, Vranješ M and Žagar D. 2010. Foveated mean squared error—a novel video quality metric. Multimedia Tools and Applications, 49(3): 425-445 [DOI: 10.1007/s11042-009-0442-1]

Rosenholtz R, Huang J and Ehinger K A. 2012. Rethinking the role of top-down attention in vision: effects attributable to a lossy representation in peripheral vision. Frontiers in Psychology, 3: #13 [DOI: 10.3389/fpsyg.2012.00013]

Stengel M, Grogorick S, Eisemann M and Magnor M. 2016. Adaptive image-space sampling for gaze-contingent real-time rendering. Computer Graphics Forum, 35(4): 129-139 [DOI: 10.1111/cgf.12956]

Surace L, Wernikowski M, Tursun C, Myszkowski K, Mantiuk R and Didyk P. 2023. Learning GAN-based foveated reconstruction to recover perceptually important image features. ACM Transactions on Applied Perception, 20(2): #7 [DOI: 10.1145/3583072]

Swafford N T, Cosker D and Mitchell K. 2015. Latency aware foveated rendering in unreal engine 4//Proceedings of the 12th European Conference on Visual Media Production. London, United Kingdom: ACM: #17 [DOI: 10.1145/2824840.2824863]

Swafford N T, Iglesias-Guitian J A, Koniaris C, Moon B, Cosker D and Mitchell K. 2016. User, metric, and computational evaluation of foveated rendering methods//Proceedings of the ACM Symposium on Applied Perception. Anaheim, California, USA: ACM: 7-14 [DOI: 10.1145/2931002.2931011]

Tariq T, Tursun C and Didyk P. 2022. Noise-based enhancement for foveated rendering. ACM Transactions on Graphics, 41(4): #143 [DOI: 10.1145/3528223.3530101]

Thumuluri V and Sharma M. 2020. A unified deep learning approach for foveated rendering and novel view synthesis from sparse RGB-D light fields//Proceedings of 2020 International Conference on 3D Immersion (IC3D). Brussels, Belgium: IEEE: 1-8 [DOI: 10.1109/IC3D51119.2020.9376340]

Tursun O T, Arabadzhiyska-Koleva E, Wernikowski M, Mantiuk R, Seidel H P, Myszkowski K and Didyk P. 2019. Luminance-contrast-aware foveated rendering. ACM Transactions on Graphics, 38(4): #98 [DOI: 10.1145/3306346.3322985]

Vaidyanathan K, Salvi M, Toth R, Foley T, Akenine-Möller T, Nilsson J, Munkberg J, Hasselgren J, Sugihara M, Clarberg P, Janczak T and Lefohn A. 2014. Coarse pixel shading [EB/OL]. [2023-09-25]. https://www.intel.cn/content/www/cn/zh/developer/articles/technical/coarse-pixel-shading.html

Wang L D, Hajiesmaili M and Sitaraman R K. 2021. FOCAS: practical video super resolution using foveated rendering//Proceedings of the 29th ACM International Conference on Multimedia. Virtual Event, China: ACM: 5454-5462 [DOI: 10.1145/3474085.3475673]

Wang L L, Shi X H and Liu Y. 2023. Foveated rendering: a state-of-the-art survey. Computational Visual Media, 9(2): 195-228 [DOI: 10.1007/s41095-022-0306-4]

Wang W G, Shen J B and Shao L. 2015. Consistent video saliency using local gradient flow optimization and global refinement. IEEE Transactions on Image Processing, 24(11): 4185-4196 [DOI: 10.1109/

TIP.2015.2460013]

Wang Z, Bovik A C, Lu L G and Kouloheris J L. 2001. Foveated wavelet image quality index//Proceedings of Volume 4472, Applications of Digital Image Processing XXIV. San Diego, United States: SPIE: 42-52 [DOI: 10.1117/12.449797]

Wang Z Q, Zhang Y S, Yu Y, Min J and Tian H. 2022. Review of deep learning based salient object detection. Journal of Image and Graphics, 27(7): 2112-2128 (王自全, 张永生, 于英, 闵杰, 田浩. 2022. 深度学习背景下视觉显著性物体检测综述. 中国图象图形学报, 27(7): 2112-2128) [DOI: 10.11834/jig.200649]

Weier M, Stengel M, Roth T, Didyk P, Eisemann E, Eisemann M, Grogorick S, Hinkenjann A, Kruijff E, Magnor M, Myszkowski K and Slusallek P. 2017. Perception-driven accelerated rendering. Computer Graphics Forum, 36(2): 611-643 [DOI: 10.1111/cgf.13150]

Weymouth F W. 1958. Visual sensory units and the minimal angle of resolution. American Journal of Ophthalmology, 46(1): 102-113 [DOI: 10.1016/0002-9394(58)90042-4]

White S, ClAndrew, Lucas E D, Mabee D, Jenks A, Sherer T and gongwaner. 2023. Variable-rate shading (VRS) [EB/OL]. [2023-09-25]. https://learn.microsoft.com/zh-cn/windows/win32/direct3d12/vrs

Wu J X, Zhong S H, Ma Z, Heinen S J and Jiang J M. 2018. Foveated convolutional neural networks for video summarization. Multimedia Tools and Applications, 77(22): 29245-29267 [DOI: 10.1007/s11042-018-5953-1]

Xiao K, Liktor G and Vaidyanathan K. 2018. Coarse pixel shading with temporal supersampling//Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games. Montreal, Canada: ACM: #1 [DOI: 10.1145/3190834.3190850]

Xu Y Y, Dong Y B, Wu J R, Sun Z Z, Shi Z R, Yu J Y and Gao S H. 2018. Gaze prediction in dynamic 360° immersive videos//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 5333-5342 [DOI: 10.1109/CVPR.2018.00559]

Yang H, Chen R, An S P, Wei H and Zhang H. 2023. The growth of image-related three dimensional reconstruction techniques in deep learning-driven era: a critical summary. Journal of Image and Graphics, 28(8): 2396-2409 (杨航, 陈瑞, 安仕鹏, 魏豪, 张衡. 2023. 深度学习背景下的图像三维重建技术进展综述. 中国图象图形学报, 28(8): 2396-2409) [DOI: 10.11834/jig.220376]

Yoo C, Xiong J H, Moon S, Yoo D, Lee C K, Wu S T and Lee B. 2020. Foveated display system based on a doublet geometric phase lens. Optics Express, 28(16): 23690-23702 [DOI: 10.1364/OE.399808]

Zare A, Aminlou A, Hannuksela M M and Gabbouj M. 2016. HEVC-compliant tile-based streaming of panoramic video for virtual reality applications//Proceedings of the 24th ACM International Conference on Multimedia. Amsterdam, the Netherlands: ACM: 601-605 [DOI: 10.1145/2964284.2967292]

Zhang X C, Sugano Y, Fritz M and Bulling A. 2015. Appearance-based gaze estimation in the wild//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE: 4511-4520 [DOI: 10.1109/CVPR.2015.7299081]

Zhang X C, Sugano Y, Fritz M and Bulling A. 2017. It's written all over your face: Full-face appearance-based gaze estimation//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Honolulu, USA: IEEE: 2299-2308 [DOI: 10.1109/CVPRW.2017.284]

Zhao Y Q, Rao Y, Dong S P and Zhang J Y. 2020. Survey on deep learning object detection. Journal of Image and Graphics, 25(4): 629-654 (赵永强, 饶元, 董世鹏, 张君毅. 2020. 深度学习目标检测方法综述. 中国图象图形学报, 25(4): 629-654) [DOI: 10.11834/jig.190307]

Zheng Z P, Cao W J, Zhou W J, Yang Z and Zhan Y W. 2020. Perceptual model based surface tessellation for VR foveated rendering//Proceedings of the 6th International Conference on Robotics and Artificial Intelligence. Singapore, Singapore: ACM: 64-68 [DOI: 10.1145/3449301.3449313]

Zhu F, Lu P, Li P, Sheng B and Mao L J. 2020. Gaze-contingent rendering in virtual reality//Magnenat-Thalmann N, Stephanidis C, Wu E H, Thalmann D, Sheng B, Kim J, Papagiannakis G and Gavrilova M, eds. 37th Computer Graphics International Conference on Advances in Computer Graphics. Geneva, Switzerland: Springer International Publishing: 16-23 [DOI: 10.1007/978-3-030-61864-3_2]

作者简介

李英群,女,硕士研究生,主要研究方向为注视点渲染。
E-mail: 291921787@qq.com
王璐,通信作者,女,博士,教授,博士生导师,主要研究方向为真实感渲染、实时渲染、真实感材质建模和高性能渲染。
E-mail:luwang_hcivr@sdu.edu.cn
胡啸,女,硕士研究生,主要研究方向为注视点渲染。
E-mail: huxiao@mail.sdu.edu.cn
徐翔,男,博士,讲师,主要研究方向为并行渲染和真实感渲染。E-mail: xuxiang@sdufe.edu.cn.
徐延宁,男,博士,副教授,主要研究方向为真实感渲染、实时渲染和真实感材质建模。E-mail: xyn@sdu.edu.cn