E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(2025)06-1828-44

论文引用格式:Zhang Y N, Wang H Y, Yan Q S, Yang J Q, Liu T, Fu M Q, Wu P and Zhang L. 2025. Research progress of unmanned mobile vision technology for complex dynamic scenes. Journal of Image and Graphics, 30(6):1828-1871(张艳宁, 王昊宇, 闫庆森, 杨佳琪, 刘婷, 符梦芹, 吴鹏, 张磊. 2025. 面向复杂动态场景的无人移动视觉技术研究进展. 中国图象图形学报, 30(6):1828-1871)[DOI:10.11834/jig.240458]

面向复杂动态场景的无人移动视觉技术研究进展

张艳宁,王昊宇,闫庆森,杨佳琪,刘婷,符梦芹,吴鹏,张磊*

1. 西北工业大学计算机学院,西安　710129; 2. 西北工业大学空天地海一体化大数据应用技术国家工程实验室,西安　710129

摘　要:随着人类活动范围的不断扩大和国家利益的持续发展,新域新质无人系统已成为世界各大国科技战略竞争的制高点和制胜未来的关键力量。无人移动视觉技术是无人系统辅助人类透彻感知理解物理世界的核心关键之一,旨在基于无人移动平台捕获的视觉数据,精准感知理解复杂动态场景与目标特性。深度神经网络凭借其超强的非线性拟合能力和区分能力,已经成为无人移动视觉技术的基准模型。然而,实际应用中无人系统通常面临成像环境复杂动态、成像目标高速机动—伪装对抗、成像任务需求多样,导致基于深度神经网络的无人移动视觉模型成像质量大幅退化,场景重建解译与目标识别分析精度显著下降,从而严重制约无人系统在复杂动态场景下对物理世界的感知解译能力与应用前景。针对这一挑战,本文深入探讨面向复杂动态场景的无人移动视觉技术发展现状,从图像增强处理、三维重建、场景分割、目标检测识别以及异常检测与行为分析等5个关键技术入手,介绍每项技术的基本研究思路与发展现状,分析每项技术中典型算法的优缺点,探究该技术目前依然面临的问题与挑战,并展望未来研究方向,为面向复杂动态场景的无人移动视觉技术长远发展与落地奠定基础。

关键词:无人移动视觉;复杂动态场景;图像增强;三维重建;场景分割;目标检测;异常检测

Research progress of unmanned mobile vision technology for complex dynamic scenes

Zhang Yanning, Wang Haoyu, Yan Qingsen, Yang Jiaqi, Liu Ting, Fu Mengqin, Wu Peng, Zhang Lei*

1. School of Computer Science, Northwestern Polytechnical University, Xi'an　710129, China; 2. National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, Northwestern Polytechnical University, Xi'an　710129, China

Abstract: In today's era of raid automation and technological advancement, unmanned systems are increasingly becoming a key area of strategic competition among major global powers. These new domains and capabilities of unmanned systems are not only key to supporting national security and strategic interests but also serve as the core force driving future technological innovation and application development. Unmanned systems are reshaping the boundaries of national security and

收稿日期:2024-08-05;**修回日期**:2024-10-28;**预印本日期**:2024-11-04

*通信作者:张磊　nwpuzhanglei@nwpu.edu.cn

基金项目:国家自然科学基金项目(62101454, 62106201, 62301432, 62372377, 62372379);陕西省自然科学基础研究计划项目(2023-JC-QN-0685);中央高校基本科研业务费专项资金资助(D5000220444);西安市重点产业链技术攻关项目(人工智能关键技术攻关)(23ZDCYJSGG0003-2023)

Supported by: National Natural Science Foundation of China(62101454, 62106201, 62301432, 62372377, 62372379); Natural Science Basic Research Program of Shaanxi Province, China(2023-JC-QN-0685); Fundamental Research Funds for the Central Universities(D5000220444); Xi'an Key Industrial Chain Core Technology Breakthrough Project: AI Core Technology Breakthrough(23ZDCYJSGG0003-2023)

redefining the connotations of strategic advantages. As a key component of unmanned systems, unmanned mobile visual technology is demonstrating its immense potential in assisting humans to gain a deeper understanding of the physical world. The advancement of this technology not only equips unmanned systems with richer and more precise perceptual capabilities but also offers humans with new perspectives to observe, analyze, and ultimately master the complex and dynamic physical environment. In the early stages of unmanned mobile visual technology development, researchers mainly relied on traditional learning methods for image processing. These methods focused on manual feature extraction, which depended heavily on the experience and knowledge of domain experts. For instance, feature descriptors such as scale-invariant feature transform(SIFT) and histogram of oriented gradients(HOG) played crucial roles in tasks such as image matching and target detection. Although traditional visual analysis methods still hold value in specific situations, their dependence on manual feature extraction and professional knowledge limits efficiency and accuracy. With the advent of deep neural network technology, unmanned mobile visual technology has ushered in revolutionary progress. Deep neural networks, through automatic feature extraction and hierarchical structures, can learn feature representations ranging from simple to complex, allowing them to capture local image features but also understand and interpret higher-level semantic information. Thus, these networks notably enhance the fitting and discriminative capabilities of models, offering advantages that traditional methods cannot match. Consequently, deep neural networks have become the benchmark for unmanned mobile visual technology. However, in practical applications, unmanned systems often encounter complex, diverse, and dynamically changing application scenarios, which present considerable challenges for the deployment and effectiveness of deep learning models. First, the complexity and dynamics of the imaging environment present notable problems in unmanned systems. Drastic changes in lighting, unpredictable weather conditions, and interference from other moving objects can degrade image quality, thereby affecting subsequent processing and analysis. Second, the high-speed maneuverability and camouflage strategies of imaging targets add another layer of difficulty for unmanned mobile visual systems. The rapid movement of targets complicates stable tracking, while camouflage and concealment make detection notably more difficult. These factors collectively reduce the accuracy of scene reconstruction, interpretation, and target identification in deep neural network-based unmanned mobile visual models. Furthermore, the diversity of imaging tasks introduces additional challenges. Different tasks often require tailored visual processing strategies, and the system must possess sufficient flexibility and adaptability to effectively handle different tasks. However, current deep neural network models are often tailored for specific tasks, limiting their adaptability across diverse applications. The uncertainty and unpredictability of environmental factors impose demanding requirements on unmanned mobile visual systems. These systems need to offer precise perception and in-depth analysis to provide decision support, enabling automated systems to respond quickly and accurately to environmental changes, thus improving overall efficiency and reliability. In response to the visual challenges of unmanned systems in complex, dynamic environments, this article delves into the current state of development of unmanned mobile visual technology in addressing these challenges, focusing on five key technical areas: image enhancement, 3D reconstruction, scene segmentation, object detection, and anomaly detection. Image enhancement, being the first step, is crucial for improving the quality of visual data. This process improves the contrast, clarity, and color of images, providing highly reliable input for subsequent analysis and processing, which enhances the performance of unmanned systems under various environmental conditions. 3D reconstruction technology facilitates the recovery of three-dimensional structures from two-dimensional images, enabling unmanned systems to gain a more comprehensive understanding of the environment, making it better suited for tasks in complex settings. Scene segmentation involves partitioning an image into semantically meaningful regions or objects, providing a basis for precise environmental perception and target recognition. Object detection is central to unmanned mobile visual technology, enabling the system to locate and identify specific targets within images or video streams. In contrast, anomaly detection focuses on identifying anomalies or events in the scene, allowing unmanned systems to timely identify and respond to potential threats. This article will provide an in-depth exploration of the research ideas, current status, and the advantages and disadvantages of typical algorithms for these key technologies, while also analyzing their performance in practical applications. The integration and collaboration of these technologies have substantially enhanced the visual perception capabilities of unmanned systems in dynamic and complex scenes, enabling them to perform tasks more intelligently and autonomously. Although some progress has been made in unmanned mobile visual technol-

ogy, numerous problems in its practical application within complex dynamic scenes are still encountered. This review aims to provide a comprehensive perspective, systematically examining and analyzing the latest research advancements in unmanned mobile visual technology for such scenes. This paper explores the advantages and limitations of the above key tasks in practical applications. In addition, this paper will discuss the gaps and challenges in current research and propose future possible research directions. Through in-depth exploration of these research directions, unmanned mobile visual technology will continue to advance, offering more robust and flexible solutions to address the challenges posed by complex dynamic scenes. This progress will lay a solid foundation for the long-term development and practical application of unmanned systems in the fields of automation and intelligence.

Key words: unmanned mobile vision; complex dynamic scenes; image enhancement; 3D reconstruction; scene segmentation; object detection; anomaly detection

0 引 言

在自动化和智能化不断推进的今天，无人系统正迅速崛起，成为全球各大国科技战略竞争的新焦点。新域新质的无人系统不仅是国家安全和战略利益的关键支撑，也是推动未来技术革新和应用发展的核心力量。无人系统正在重新定义国家安全的边界和战略优势的内涵。无人移动视觉技术，作为无人系统的关键组成部分，正在展现出其在辅助人类深入理解物理世界方面的巨大潜力。这一技术的进步不仅为无人系统提供了更为丰富和精准的感知能力，也为人类提供了全新的视角，以观察、分析并最终掌握复杂多变的物理环境。

在无人移动视觉技术的发展初期，主要依赖传统学习方法进行处理，这些方法侧重于手工特征提取，依赖领域专家的经验和知识。例如，尺度不变特征变换（scale-invariant feature transform，SIFT）（Lowe，2004）和方向梯度直方图（histogram of oriented gradients，HOG）（Dalal 和 Triggs，2005）等特征描述符，在图像匹配和目标检测任务中曾发挥重要作用。虽然传统视觉分析方法在特定情况下仍具有其价值，但其依赖于人工特征提取和专业知识，限制了分析的效率和准确性。随着深度神经网络技术的兴起，无人移动视觉技术迎来了革命性的进步。深度神经网络通过自动提取特征，利用其层级结构，能够自动学习从简单到复杂的特征表示。这使得它们在捕捉图像局部特征的同时，也能理解和解释更高层次的语义信息。这显著提升了模型的拟合能力和判别能力，展现了传统方法无法比拟的优势，使深度神经网络成为无人移动视觉技术的基准模型。

然而在实际应用中，无人系统通常面临复杂多样且动态变化的应用场景，对深度学习的应用带来了极大的挑战。首先，成像环境的复杂动态性是无人系统必须面对的问题。环境光照的剧烈变化、天气条件的不确定性，以及场景中其他移动物体的干扰，都可能导致图像质量下降，进而影响后续的处理和分析；其次，成像目标的高速机动性和伪装对抗行为，对无人移动视觉系统提出更高的要求。目标的快速移动使得系统难以稳定跟踪，而伪装和隐蔽行为则使得目标检测变得更加困难。这些因素共同作用，使得基于深度神经网络的无人移动视觉模型在场景重建解译和目标识别分析的精度显著下降；此外，成像任务需求的多样性也给无人移动视觉技术带来了挑战。不同的任务可能需要不同的视觉处理策略和分析方法，而系统需要具备足够的灵活性和适应性，以满足不同任务的需求。然而，当前的深度神经网络模型往往在设计时针对特定的任务进行优化，对于多样化任务的适应能力有限。环境因素的不确定性和不可预测性对无人移动视觉技术的应用提出极为苛刻的要求，这要求无人移动视觉技术能够提供精确的感知与深入的分析，从而为自动化系统提供决策支持，使其能够快速且准确地响应环境变化，提高系统的效率和可靠性。

针对无人系统在复杂动态场景下的视觉挑战，本文深入探讨了无人移动视觉技术在应对这些挑战时的发展现状，特别关注了图像增强处理、三维重建、场景分割、目标检测识别以及异常检测与行为分析这5个关键技术领域。图像增强处理是提升视觉数据质量的首要步骤，它通过改善图像的对比度、清晰度和色彩等，为后续的分析和处理提供了更可靠的输入，从而增强了无人系统在各种环境条件下的性能；三维重建技术通过从二维图像中恢复三维结构，使无人系统能够理解场景的深度和空间布局，以

增强系统对复杂环境的理解和适应性；场景分割将图像分割成多个语义上有意义的区域或对象，为精确环境感知和目标识别提供了基础；目标检测识别是无人移动视觉技术中的一个核心任务，使系统能够在图像或视频中定位和识别特定的目标；异常检测与行为分析技术关注于识别场景中的异常行为或事件，为无人系统提供了及时识别和响应潜在威胁的能力。针对这些关键技术，本文将深入探讨它们的研究思路、发展现状以及典型算法的优缺点，分析这些技术在实际应用中的表现。这些技术的集成和协同工作，使得无人系统在复杂动态场景下的视觉感知能力得到了显著提升，使其能够更加智能和自主地执行任务。

尽管已有研究取得了一定的进展，但面向复杂动态场景的无人移动视觉技术在实际应用中仍面临诸多问题。本综述旨在提供一个全面的视角，对面向复杂动态场景的无人移动视觉技术的最新研究进展进行系统梳理和分析。探讨上述关键任务在实际应用中的优势和局限性。此外，本文还将讨论当前研究中存在的空白和挑战，并提出未来可能的研究方向。通过这些研究方向的深入探索，无人移动视觉技术将不断进步，为解决复杂动态场景下的挑战提供更加强大和灵活的解决方案，为无人系统在自动化和智能化领域的长远发展与实际应用奠定坚实基础。

1　图像增强处理

随着无人移动视觉技术在复杂动态场景中的广泛应用，图像增强处理的重要性愈加凸显。复杂动态场景往往伴随各种环境干扰和图像质量问题，如噪声、模糊和低分辨率等。这些问题严重影响了无人移动设备的感知能力和决策准确性。为了解决这些挑战，研究人员提出多种图像增强技术，包括图像去噪、图像去模糊、图像超分辨率、单帧图像增强和多帧图像增强。

1.1　图像去噪

无人设备拍摄过程中，图像的质量往往受到机器的震动和外界环境因素的干扰，导致图像中出现各种噪点，例如由于传感器本身固有的噪声、光线条件的变化、摄像头镜头的污垢或损伤等引起的视觉干扰。这些噪点不仅影响了图像的清晰度和质量，还可能干扰对图像内容的准确理解和后续处理的可靠性。因此，为了还原图像，基于无人移动视觉的图像去噪算法应运而生。如图 1 所示，以保障高质量的可用图像，这具有重要的理论和实际价值。根据图像去噪算法处理图像的域不同，可以将其划分为空间域去噪（Chang 等，2016）、变换域去噪（Satapathy 等，2019）和基于深度学习的去噪算法（Dabov 等，2007）。

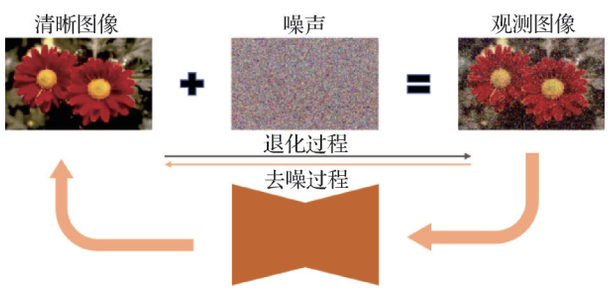


图 1　图像退化过程和去噪过程

Fig. 1　Image degradation and denoising processes

空间域去噪是指直接在图像的空间域中对像素点进行处理来消除噪声。常见的空间域图像去噪算法有：均值滤波、中值滤波以及非局部均值滤波等。例如，Erkan 等人（2019）提出一种递归均值滤波算法，对含噪声的图像进行多次均值滤波处理，可以较好地去除图像中相应的噪声。虽然空间域去噪方法有效利用局部信息或自相似性去除噪声，但会损失图像细节信息。

变换域去噪是指将图像变换到另一个域（如频域、小波域等）进行去噪处理。经典的变换域去噪算法有：傅里叶变换去噪、小波变换去噪等。例如，Kumari 和 Mustafi（2022）采用一种基于分数阶傅里叶变换（fractional Fourier transform，FrFT）的算法来对噪声图像进行去噪处理，将输入图像转换到 FrFT 域，再通过在 FrFT 域中进行滤波操作来去除图像中的噪声。与空间域去噪算法相比，变换域去噪算法计算量较大，但去噪效果相对较好。

基于深度学习的图像去噪算法一般分为基于多层感知机（multilayer perceptron，MLP）的图像去噪算法和基于卷积神经网络（convolutional neural network，CNN）模型的图像去噪算法。例如，Zhang 等人（2017）设计了一种前馈去噪 CNN，采用了批标准化和残差学习相结合的方法，加快了网络的训练速度，并提升了图像去噪的性能。基于深度学习的去噪方

法利用深度神经网络来学习图像的噪声模型和去噪策略，去噪效果较好，但需要大量的标记数据进行训练。

尽管图像去噪算法在复杂场景下展现出强大性能，但仍面临挑战。在不同噪声类型和强度混合的场景下，准确判定噪声类型和强度，并设计出高鲁棒性和泛化性的算法仍是一个挑战。此外，去噪算法的平滑处理可能减少图像高频成分，破坏原始纹理细节，因此在噪声识别和细节保护方面仍需改进。

各种去模糊经典算法实验如表1与图2所示，实验的模型有DnCNN（Murali 和 Sudeep，2020）、Restormer（Zamir等，2022）、MaskDenoising（Chen等，2023a）、HAT（hybrid attention Transformer）（Chen等，2023c）和DIL（distortion invariant representation learning）（Li等，2023b），实验的数据集有两种，分别为CBSD68（color BSD68）（Martin等，2001）和McMaster（Zhang 等，2011）。MaskDenoising在CBSD68和McMaster数据集上处理Spatial Gauss和Salt&Pepper两种噪声的表现非常好，DIL在Poisson噪声的性能更好。HAT模型在处理3种噪声类型时，可看出指标变化不大，表现稳定。由此可得，在具体的应用场景和噪声类型上选择不同算法可以达到最佳的去噪效果。

表1　经典去噪模型在CBSD68和McMaster数据集上的评估结果

Table 1　The evaluation results of the classic deblurring model on the CBSD68 and McMaster datasets

模型	Spatial Gauss σ = 55		Poisson α = 3.5		Salt&Pepper d = 0.02	
	CBSD68	McMaster	CBSD68	McMaster	CBSD68	McMaster
	PSNR(/dB)/SSIM	PSNR(/dB)/SSIM	PSNR(/dB)/SSIM	PSNR(/dB)/SSIM	PSNR(/dB)/SSIM	PSNR(/dB)/SSIM
DnCNN	25.91/0.699	26.18/0.649	24.37/0.627	25.50/0.651	26.53/0.746	25.72/0.691
Restormer	23.51/0.595	24.01/0.539	22.20/0.559	21.93/0.579	23.59/0.679	23.05/0.640
MaskDenoising	**26.72/0.762**	**26.89/0.709**	24.24/0.638	25.17/0.590	**29.74/0.843**	**29.28/0.773**
HAT	26.39/0.713	26.62/0.665	26.61/0.733	27.54/0.723	27.55/0.782	26.62/0.727
DIL	24.61/0.630	24.82/0.574	**27.64/0.819**	**28.91/0.825**	29.45/0.822	**29.28/0.773**

注：加粗字体表示各列最优结果。

1.2　图像去模糊

无人移动平台下的视觉图像成像涉及大量视觉信息的获取、传输和处理，但相机与场景的相对运动导致图像模糊，影响数据的精确度和可用性。因此，研究和应用去模糊技术对各领域都有重大意义，受到学术界关注。图像退化过程如图3所示，图像复原是从退化图像中恢复清晰场景。去模糊方法可分为基于迭代优化的传统方法和基于深度学习的方法。

基于迭代优化的传统方法主要包含图像先验设计和模糊核估计两部分。对于图像先验设计的常用的方法有基于概率的统计先验，主要是依靠贝叶斯理论和最大后验概率准则，将图像去模糊任务构建为包含模糊核和清晰图像的复原模型。经典的图像先验设计方法有基于L2范数正则化的图像梯度先验（Tikhonov，1963），以及过完备字典（Chen等，2021a）、判别学习框架（Cho等，2020）和去噪先验（Zuo等，2015）。这些方法通过拟合大量图像数据特性，实现清晰图像的先验建模。为了准确估计模糊核，通常采用的方法是利用图像中的锐利结构，尤其是强边缘，来设计模糊核的先验。经典的模糊核估计方法有通过双边滤波（Zhong等，2013）和自动梯度选择算法等技术。这些方法在去噪的同时，能够保留图像中的关键信息，为图像去模糊提供了坚实的基础。

随着图像处理技术的发展，深度学习算法在去除动态场景非均匀模糊方面表现出色。Nah等人（2017）提出多尺度卷积网络架构，通过逐步恢复不同分辨率的清晰图像来提高效果。Tao等人（2018）通过跨尺度权值共享机制降低了训练难度。Gao等人（2019）引入参数选择性共享和嵌套跳层连接模块，提升了训练速度和图像恢复效果。这些研究展示了卷积神经网络在图像去模糊领域的潜力，有效地提升了图像恢复的性能，为未来图像处理技术的发展提供了新的思路和方法。

随着注意力机制的提出，越来越多的学者尝试

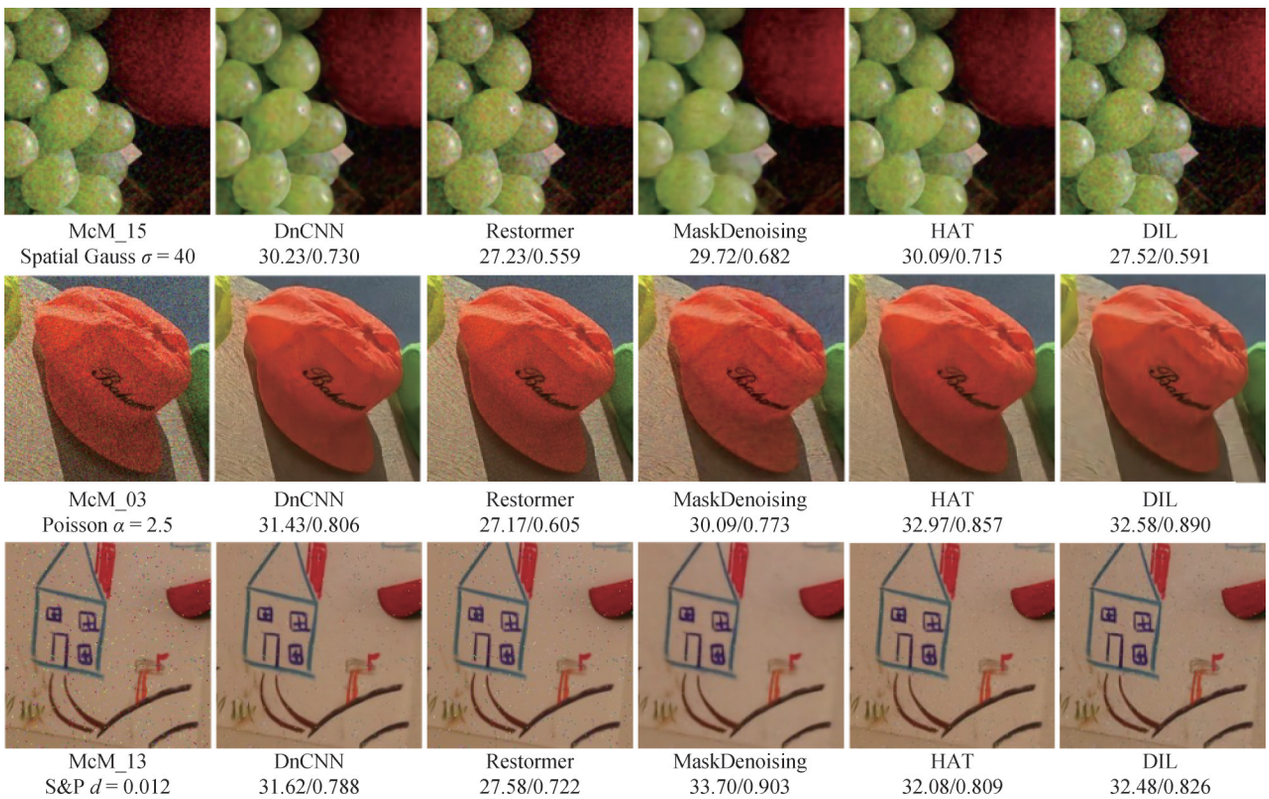


图 2　经典去噪模型定性实验对比(PSNR(/dB)/SSIM)

Fig. 2　Qualitative experimental comparison of classic denoising models(PSNR(/dB)/SSIM)

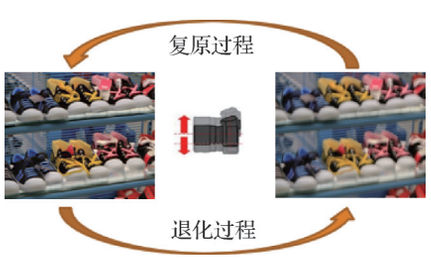


图 3　图像去模糊退化和复原过程

Fig. 3　The process of image deblurring degradation and restoration

用 Transformer 架构进行去模糊。Restormer(Zamir 等,2022)通过多头转置注意力模块聚合局部和非局部像素信息,Uformer(Wang 等,2022c)采用非重叠窗口的自注意力模型和深度卷积,提升了处理效率和性能。这些方法为图像处理领域提供了新的解决方案,推动了技术的发展。

尽管图像去模糊技术已有进展,但仍面临算法复杂度高、模型准确性和泛化能力不足,以及在复杂场景中对噪声敏感的问题。未来的发展方向包括优化算法以实现实时去模糊,开发自适应不同模糊类型和程度的模型,结合深度信息或运动数据提高去模糊准确性。需要不断创新和跨任务合作,以解决现有问题并开拓新应用领域。

各种去模糊经典算法实验如表 2 与图 4 所示,实验的模型有 IFAN(iterative filter adaptive network)(Lee 等,2021)、DDDNet(dilated deformable deep net)(Pan 等,2021)、DeepRFT(deep residual fourier transformation)(Mao 等,2023)、Restormer(Zamir 等,2022)和 K3DN(disparity-aware kernel estimation for dual-pixel defocus deblurring)(Yang 等,2023)。其中,K3DN 和 Restormer(Zamir 等,2022)模型在图像重建质量上表现最为出色。在峰值信噪比(peak signal-to-noise ratio,PSNR)、结构相似性(structural similarity index measure,SSIM)、平均绝对误差(mean absolute error,MAE)和 MAE_rel 这 4 个关键指标上都显示了较高的性能。特别是 K3DN 模型,在图像去模糊的准确性和视觉质量上都有优势。

1.3　图像超分辨率

随着无人移动视觉技术的发展,对高质量图像的需求不断增加,如无人机航拍、自动驾驶和视频监控。然而,动态复杂环境、成像设备和目标的移动性以及不同光照和噪声等因素,使得采集高分辨率图像变得困难。这些外界环境导致观测图像存在几何

表2　在由37个室内场景和39个室外场景组成的DPD-blur（Abuolaim和Brown，2020）数据集上进行的定量实验

Table 2　Quantitative experiments conducted on the DPD-blur (Abuolaim and Brown, 2020) dataset, which consists of 37 indoor scenes and 39 outdoor scenes

方法	PSNR/dB	SSIM	MAE	MAE_rel
IFAN	25.99	0.804	0.037	0.050
DDDNet	25.36	0.768	0.041	0.054
DeepRFT	25.71	0.801	0.037	0.051
Restormer	26.66	0.833	0.035	0.046
K3DN	**27.06**	**0.835**	**0.034**	**0.044**

注：加粗字体表示各列最优结果。

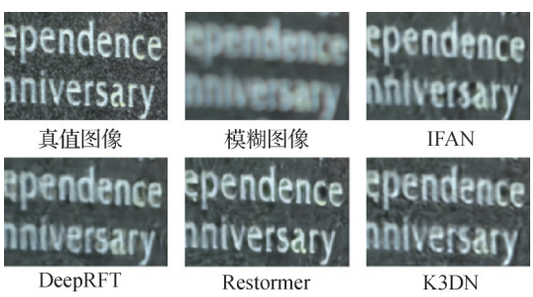


图4　图像去模糊定性实验结果

Fig. 4　Qualitative experimental results of image deblurring

失真、散焦模糊、运动模糊、噪声干扰和明暗变化。为了满足应用需求，图像超分辨重建技术应运而生，旨在通过信号处理去除退化因素干扰，提高图像分辨率。目前主流的面向复杂场景的图像超分辨重建方法主要有基于退化过程显式建模的方法（Cornillère等，2019）、基于退化过程隐式建模的方法（Wei等，2020）、基于域转换的方法（Zhao等，2018b）和基于自监督学习的方法（Li等，2022d）。

基于退化过程显式建模的图像超分辨重建方法的关键在于提出退化核估计方法，以模拟低分辨图像的退化过程。这些方法使用估计的退化模型来生成逼近真实世界的低分辨率图像，或将其作为先验知识嵌入超分辨模型中。然而，这些方法的一个显著缺点是对估计的退化核高度依赖。如果估计的退化核与实际退化过程不符，性能将急剧下降。

基于退化过程隐式建模的方法则是利用大量高—低分辨率图像对，设计专门的网络模型来学习低、高分辨率图像之间的特征映射关系。该类方法的有效性主要依赖人工采集的大量对齐的高、低分辨率图像对。然而真实的成像环境是复杂多变的，有时很难采集高、低分辨率对应的图像对用以模型的训练，这限制了上述方法的进一步应用。

基于域转换技术的超分辨重建方法放松了对训练数据集的要求，认为真实世界的低分辨率图像、理想的低分辨率图像（仿真低分辨率图像）和高分辨率图像处于不同的域中。其目标是将真实世界低分辨率图像域转换到高分辨率图像域。如图5所示，主要有两种方法实现这一转换：1）基于两阶段的域转换方法；2）基于单阶段的域转换方法。两者的主要区别在于是否引入理想的低分辨率图像辅助域转换。尽管这些方法在一定程度上提高了超分辨技术在真实世界中的泛化能力，但其训练过程仍依赖大量的高分辨率图像，而在实时成像环境中获取大量满足条件的高分辨率图像仍然是一个重大挑战。

基于自监督技术的超分辨方法正是解决上述问题的关键。其模型的构建仅依赖输入图像自身，而

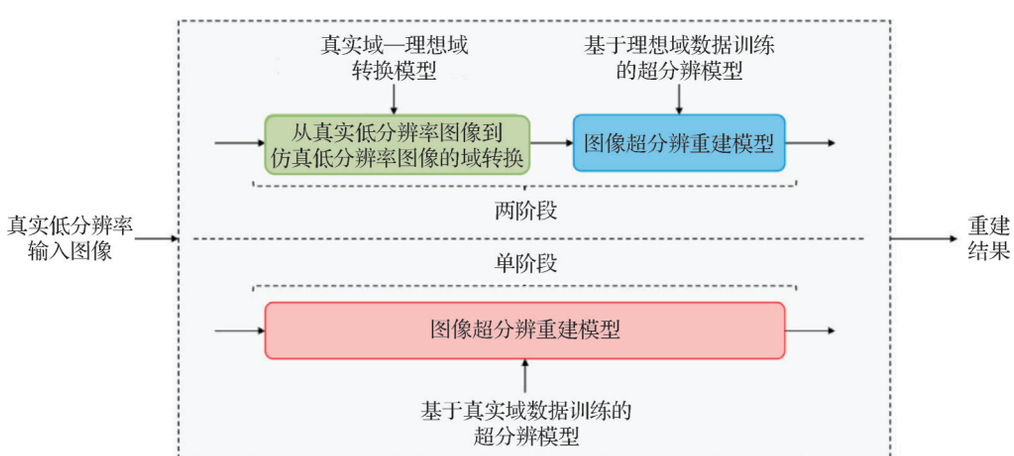


图5　单阶段和两阶段图像超分辨率重建模型

Fig. 5　Single-stage and two-stage image super-resolution reconstruction models

无需外部数据，突破了训练集对模型的限制，使得超分辨技术的实际应用变得更加广泛。

如表 3 所示，RCAN（residual channel attention networks）（Zhang 等，2018b）、DAN（deep alternating network）（Huang 等，2024）和 IKC（iterative kernel correction）（Gu 等，2019）模型在这些指标上的表现相对较为均衡，其中 IKC 模型计算复杂度最小，BSRGAN（blind super-resolution generative adversarial networks）（Zhang 等，2021a）模型在 PSNR、LPIPS（learned perceptual image patch similarity）和 SSIM 这 3 个衡量图像质量的指标上都表现出色，可以认为在图像超分辨率任务中的 LPIPS 整体性能最佳。Real-ESRGAN（real-enhanced super-resolution generative adversarial networks）（Wang 等，2021b）模型在 LPIPS 表现最佳，但在 PSNR 和 SSIM 上略逊一筹，这意味着它在某些视觉感知方面做得很好，但在像素级别的精确重建上不如 BSRGAN。ZSSR（zero-shot super-resolution）（Shocher 等，2018）的 3 个指标均表现较差，有待提升。

表 3　图像超分辨重建经典模型在 NTIRE-RWSR（Lugmayr 等，2020）数据集上实现 ×4 超分辨

Table 3　Classical models of image super-resolution reconstruction achieve ×4 super-resolution (SR) on the NTIRE-RWSR (Lugmayr et al. ,2020) dataset

方法	MACs/G	LPIPS	PSNR/dB	SSIM
Bicubic	–	0.537	22.35	0.617
RCAN	260	0.472	22.32	0.604
ZSSR	–	0.639	22.21	0.603
IKC	80	0.479	22.25	0.600
DAN	315	0.471	22.41	0.609
BSRGAN	291	0.299	**22.47**	**0.623**
Real-ESRGAN	291	**0.238**	22.08	0.622

注：加粗字体表示各列最优结果，“–”表示无数据。

1.4　单帧图像增强

单帧图像增强在提升低照度条件下的图像质量和可见性方面至关重要，尤其对于无人驾驶汽车、无人机侦察和夜间监控等无人移动视觉系统。它支持更准确的图像分析和决策制定。主要挑战包括处理噪声放大、细节丢失和颜色失真。解决这些问题的方法主要分为 3 类：传统图像处理方法、基于 Retinex（retina-exponential）理论的低照度增强算法和基于深度学习的方法，如图 6 所示。

图 6　单帧图像增强技术

Fig. 6　Single frame image enhancement technology

传统的低照度增强方法主要基于图像处理技术，如直方图均衡化、局部对比度增强和伽马校正。这些方法无需训练数据，计算效率高。直方图均衡化调整图像的直方图以扩展整体对比度，伽马校正通过非线性变换调整亮度，局部对比度增强则改善局部细节的清晰度。然而，这些方法难以适应不同场景变化，且容易引入伪影和过度增强的问题。

基于 Retinex 理论的低照度增强算法能够解决直方图均衡等传统算法常见的过度增强或低对比度问题，有效地保持图像的细节信息，从而提升图像质量和视觉体验。Li 等人（2018）提出一种基于鲁棒 Retinex 模型的低光照图像增强方法。该方法通过引入图像对比度的局部自适应性来改进传统的 Retinex 模型，从而增强低光照图像的结构信息，显著提高了图像的清晰度和视觉质量。虽然基于 Retinex 理论的低照度增强算法可以增强低照度图像的整体信息，但算法结构不灵活，局部信息增强效果不好。

在单帧图像增强领域，基于深度学习的方法已成为解决低光照图像增强问题的重要手段。Xu 等人（2022）基于不同光照区域具有不同的信噪比特点，结合 Transformer 理论与卷积神经网络，通过减少对低信噪比区域的信息聚合来提高图像处理效果。这些基于深度学习的低光照图像增强方法通过复杂的神经网络架构和大量的训练数据，能够有效地处理图像的噪声放大、细节丢失和颜色失真等问题，显著提升低光照图像的质量和可见性。

单帧图像低照度增强技术是无人移动视觉系统中的一个重要研究领域，涉及多种技术和方法。传统方法虽然简单高效，但可能在复杂环境下表现不佳。而基于深度学习的方法虽然计算成本较高，但能够提供更加精细和适应性强的增强效果，是当前

和未来研究的重点方向。

对单帧图像低光照增强进行定量实验，实验模型有PairLIE（paired low-light instances）（Fu等，2023）、Zero-DCE（zero-reference deep curve estimation）（Guo等，2020）、RUAS（retinex-inspired unrolling with cooperative prior architecture search）（Liu等，2021）、SCI（self-calibrated illumination）（Ma等，2022）和CLIP-LIT（contrastive language-image pre-training for pixel—level image enhancement）（Liang等，2023），结果如表4所示，RUAS和CLIP-LIT模型所有指标上的表现并不突出。Zero-DCE和SCI模型的PSNR和SSIM指标较高，但在保持图像结构和亮度信息方面有待提高。PairLIE模型在所有评估指标上均表现最佳，在无监督的弱光单帧图像增强任务中具有较高的性能和较好的图像质量保持能力。图7为单帧图像低照度增强定性实验结果。

表4　单帧图像低光照增强定量实验结果

Table 4　Quantitative experimental results of low-light enhancement for single-frame images

方法	PSNR/dB	SSIM	LPIPS	LOE
PairLIE	**19.70**	**0.774**	**0.235**	**0.278**
Zero-DCE	17.64	0.572	0.316	0.296
RUAS	15.47	0.490	0.305	0.330
SCI	16.97	0.532	0.312	0.289
CLIP-LIT	14.82	0.524	0.371	0.320

注：加粗字体表示各列最优结果。

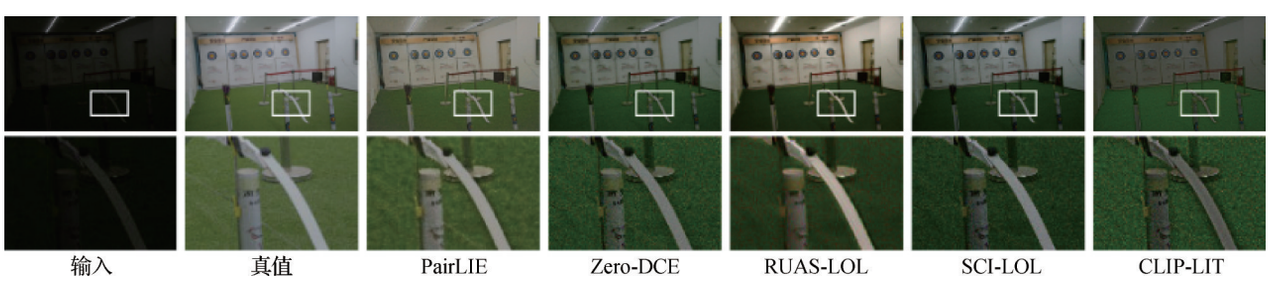


图7　单帧图像低照度增强定性实验结果

Fig. 7　Qualitative experimental results of low-light enhancement for single-frame images

1.5　多帧图像增强

在复杂动态场景中，无人移动视觉技术面临光照变化、遮挡和动态物体等挑战。多帧图像增强技术通过融合多帧图像的信息提高图像质量和细节展现，提升视觉系统的可靠性和精度。主要方法包括多模态图像增强和多曝光图像增强。多模态图像增强融合不同传感器数据，广泛应用于无人机、自动驾驶和机器人，以准确识别目标物体和检测障碍物。多曝光图像增强融合不同曝光时间的图像生成高动态范围（high-dynamic range，HDR）图像，提升在高对比度环境下的图像质量和场景识别能力。

多模图像融合技术通过整合多源图像中的有用信息生成单幅融合图像，以实现对场景的有效全面表征。典型方法包括基于自编码器（Li等，2020）、卷积神经网络（Han等，2022）和生成对抗网络（Liu等，2022a）的图像融合框架。基于自编码器的方法在大型数据集预训练自编码器进行特征提取和图像重建，并通过手工设计的融合策略整合深度特征；基于CNN的方法通过设计网络结构和损失函数，实现端到端的特征提取、融合和图像重建，避免了手动设计融合规则的烦琐（Ma等，2021）；基于生成对抗网络（generative adversarial network，GAN）的融合框架则将图像融合建模为生成器与判别器之间的对抗博弈，通过判别器迫使生成器生成的融合结果在概率分布上与目标分布一致，从而隐式实现特征提取、融合和图像重建。总之，这3种方法各有优缺点，自编码器方法依赖于手工设计的融合策略，适用性较差；CNN方法通过端到端训练，实现了特征提取和融合过程的一体化，减少了手工设计的复杂性；GAN方法通过对抗博弈机制，提高了融合图像的质量，但训练过程较为复杂。

多曝光图像融合技术通过融合同一场景下不同曝光度拍摄的多幅图像，生成HDR图像。该技术分为基于运动检测（Hu等，2019b；Ye等，2022）、基于图像配准（Trinidad等，2019）和基于深度学习（Deng等，2020；唐凌峰 等，2022）3类方法。基于运动检测的方法通过去除或削弱运动像素的权重生成融合图像，适用于少量运动像素的场景，但在大量内容运动

时会丢失信息，导致图像质量较低；基于图像配准的方法通过对齐参考图像和源图像序列的信息生成融合图像，利用输入图像的信息，效果较好；基于深度学习的方法在多曝光融合中表现优越。Yan 等人（2019）提出的基于注意力机制的 HDR 重建网络，通过引入注意力模块，改善了不同曝光下细节丢失的问题，同时消除多幅图像间的运动差异，提升合成 HDR 图像的质量和逼真度。3 种多曝光图像融合方法各有优缺点：基于运动检测的方法简单直接，但在大量运动场景中效果不佳；基于图像配准的方法通过对齐技术有效融合多幅图像，生成高质量的 HDR 图像；基于深度学习的方法通过学习最佳融合策略，在处理复杂和动态场景时表现出色。

多帧图像增强技术通过综合利用多帧图像的信息，显著提升了图像的质量和分辨率。随着计算机技术和深度学习技术的发展，多帧图像增强技术在无人驾驶、安防监控、视频增强以及虚拟现实等领域实现高质量的图像和视频处理成为可能，不仅提高了图像质量和分辨率，还增强了系统的鲁棒性和稳定性。

多曝光图像融合经典模型有 Sen（Sen 等，2012）、Hu（Hu 等，2013）、Kalantari（Kalantari 和 Ramamoorthi，2017）、DeepHDR（Wu 等，2018）、AHDRNet（attention-guided end-to-end deep neural network）（Yan 等，2019）、NHDRR（Yan 等，2020）、HDRGAN（Niu 等，2021）和 APNT（attention-guided progressive neural texture fusion）（Chen 等，2022a），表 5 为多曝光图像融合经典模型在 Kalantari（Kalantari 和 Ramamoorthi，2017）数据集的定量实验结果。

从表 5 中的 PSNR-μ 和 PSNR-L 两个指标来看，APNT 模型在平均和亮部的峰值信噪比上都取得最高的分数，分别为 43.92 dB 和 41.57 dB，表明 APNT 在整体和亮部区域的细节恢复上，能够更好地保持图像的质量和细节信息。在 SSIM-μ 和 SSIM-L 两个指标上，HDRGAN 和 APNT 模型同样表现出色，HDRGAN 的结果为 0.990 5 和 0.986 5。APNT 的结果为 0.989 8 和 0.987 9。SSIM 值越接近 1，表示图像的结构信息保持得越好，表明这两个模型在图像增强过程中能够较好地保留原始图像的结构特征。HDR-VDP-2 指标用于评估 HDR 图像的视觉质量，得分越高表示视觉质量越好。在这个指标上，HDRGAN 以 65.45 高分领先。

表 5　多曝光图像融合经典模型在 Kalantari（Kalantari 和 Ramamoorthi，2017）数据集定量实验结果

Table 5　Quantitative experimental results of classical multi-exposure image fusion models on the Kalantari（Kalantari and Ramamoorthi，2017）dataset

方法	PSNR-μ /dB	PSNR-L /dB	SSIM-μ	SSIM-L	HDR-VDP-2
Sen	40.95	38.31	0.980 5	0.972 6	55.72
Hu	32.19	30.84	0.971 6	0.950 6	55.25
Kalantari	42.74	41.22	0.987 7	0.984 8	60.51
DeepHDR	41.64	40.91	0.986 9	0.985 8	60.50
AHDRNet	43.62	41.03	0.990 0	0.986 2	62.30
NHDRR	42.41	41.08	0.988 7	0.986 1	61.21
HDRGAN	43.92	41.57	**0.990 5**	0.986 5	**65.45**
APNT	**43.94**	**41.61**	0.989 8	**0.987 9**	64.05

注：加粗字体表示各列最优结果。

2　三维重建

2.1　多视图几何点云三维重建

多视图几何点云三维重建通过拍摄目标物体场景的一系列图像或视频作为输入，首先利用运动结构恢复算法（structure from motion，SfM）或利用相机传感器自身标定来计算相机的位置姿态信息，然后采用多视图立体重建算法（multi-view stereo，MVS）来从一系列已知相机参数的图像中恢复场景的稠密点云表征。依据输出结果表征，MVS 方法可以大致分为基于体积标量场的方法（Hernández 等，2007）、基于网格的方法（Esteban 和 Schmitt，2004）、基于点云的方法（Furukawa 和 Ponce，2009）和基于深度图的方法（Campbell 等，2008）。依据方法类别，MVS 方法又大致可分为基于传统几何的 MVS 方法以及基于深度学习的 MVS 方法。然而，目前常用的基于传统几何的 MVS 方法和基于深度学习的 MVS 方法，其输出表征均为深度图。因此，本节将依据方法类别进行介绍。

2.1.1　基于传统几何的多视图立体重建方法

表 6 为传统多视图立体方法对比。基于传统几何的多视图立体重建算法，大多为块匹配算法的变种，块匹配算法最初由 Barnes 等人（2009）提出，其主要目标是快速识别两幅图像中的近似最近邻补丁。

Shen(2013)首先将块匹配的思想扩展到MVS。为了提高效率,Galliani等人(2015)提出Gipuma,其利用红黑棋盘图案执行类似扩散的传播方案,更好地利用了图形处理器(graphics processing unit,GPU)的并行性以提高效率。然而,Gipuma的主要缺点是每个像素的采样点是提前预设的,进而导致了对图像区域信息的忽略。因此,Xu和Tao(2019)设计了一种自适应棋盘采样和多假设联合视图选择(adaptive checkerboard sampling and multi-hypotesis joint view selection,ACMH)的方法来解决这一缺陷。针对低纹理区域的深度估计,他们进一步将ACMH与多尺度几何一致性引导相结合,提出多尺度几何一致性引导的MVS算法(combine ACMH with multi-scale geometric consistency guidance,ACMM)。尽管前述方法显著增强了基于传统几何的MVS方法的性能,但恢复低纹理区域的深度信息始终是一个具有挑战性的问题。块匹配多视图立体重建算法流程如图8所示。

表6　传统多视图立体方法对比

Table 6　Comparison of traditional multi-view stereo methods

方法	优点
Gipuma	时间开销低
COLMAP	精度高
ACMM	完整度高
APD-MVS	能够处理无纹理区域
ACMP	时间开销低,精度与完整达到平衡
HPM-MVS	在大规模场景表现优异,时间开销与性能达到平衡

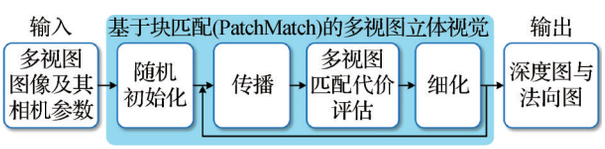


图8　块匹配多视图立体重建算法流程

Fig. 8　Block matching multi-view stereo reconstruction

为了解决这个长期存在的问题,许多方法考虑结合先验信息来重建。Romanoni和Matteucci(2019)提出TAPA-MVS(textureless-aware patch-match multi-view sereo),将图像分为两个不同尺度的超像素,并将其用做区域结构,随后利用可信的三维点和随机抽样一致性(random sample consensus,RANSAC)方法在超像素内拟合平面,然后利用先验假设优化低纹理区域。与TAPA-MVS类似,Kuhn等人(2019)采用一种基于扩展超像素的平面补全方法,以恢复低纹理区域深度图的完整性。Xu和Tao(2020)基于德劳内三角化构建了平面先验假设,随后通过概率图模型将先验假设融入MVS过程中。Xu等人(2023b)设计了一个多尺度几何一致性引导和平面先验辅助的多视图立体方法,通过结合ACMM和ACMP,充分利用各种尺度的深度信息,并结合先验辅助有效地处理低纹理区域。Ren等人(2023)提出将非局部的思想融合到MVS中,其中引入了层次先验挖掘的框架,通过构建由粗到精的平面先验模型,逐步重建目标的三维几何。

2.1.2　基于深度学习的多视图立体重建方法

表7为深度学习多视图立体方法对比。深度学习已广泛用于三维视觉问题,并在多视图立体重建中发挥了关键作用。MVSNet(Yao等,2018)引入一个端到端的深度学习网络,从二维图像特征构建三维成本体积。然而,MVSNet对整个三维体积进行正则化,导致了过多的内存需求。MVSNet网络结构如图9所示。R-MVSNet(Yao等,2019)通过卷积GRU依次对成本体积沿深度方向进行正则化,以实现大幅降低内存的目标。为了进一步减少内存占用并提高性能,CIDER(MVS network with correlation cost volume and inverse depth regression)(Xu等,2020)使用平均分组相关相似度度量构建轻量级的成本体积。

表7　深度学习多视图立体方法对比

Table 7　Comparison of deep learning-based multi-view stereo methods

方法	优点
MVSNet	首个开创性工作
CasMVSNet	显存较低
PatchMatchNet	时间开销低,训练资源占用低
TransMVSNet	精度高
WT-MVSNet	完整度高,能够处理大规模场景

近年来,基于深度学习的MVS的效率引起了更多关注。CVP-MVSNet(Yang等,2020)、CasMVSNet(Gu等,2020)和UCS-Net(uncertainty-aware cascaded

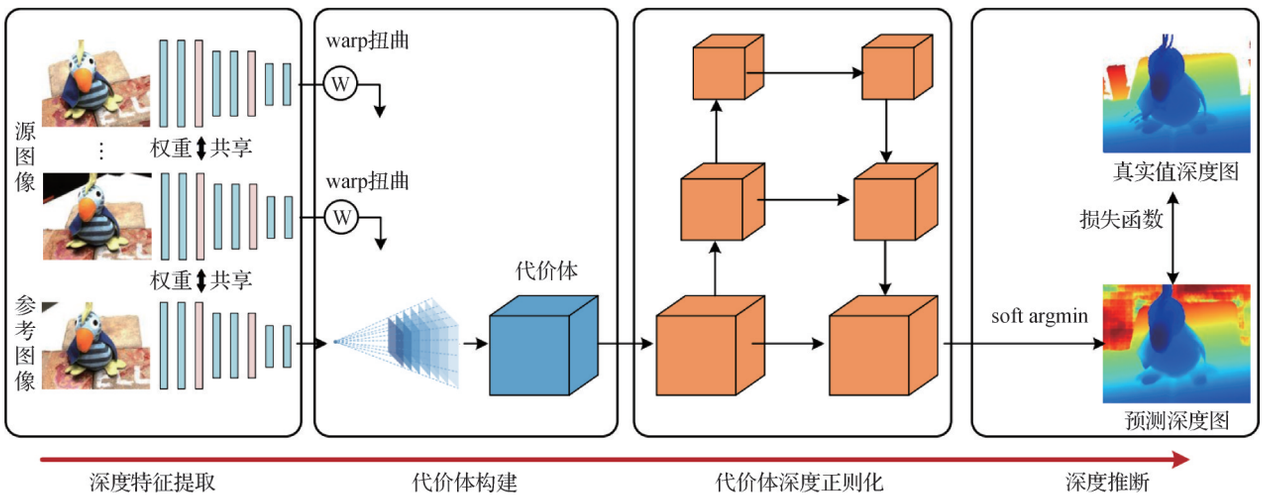


图 9　MVSNet 网络结构

Fig. 9　MVSNet architecture

stereo network)(Cheng 等,2020b)通过集成粗到精的策略构建级联成本体积。这些多阶段方法首先预测粗略的深度图,然后基于初始预测,在高分辨率尺度上进行深度恢复。因此,它们可以在占用较小内存空间的同时快速预测图像的深度。

此外,PatchMatchNet(Wang 等,2021a)将块匹配算法嵌入深度学习框架中,以实现高质量的重建,同时减少内存消耗。最近,Transformers 通过其捕获远距离关系的能力显著推动了 MVS 的研究。AttMVS(Luo 等,2020a)引入了增强的注意力匹配置信度体积,以融入局部上下文信息,增强匹配的鲁棒性。TransMVSNet(Ding 等,2022)借助 Transformer 实现了跨图像内部和跨图像的稳健远程全局上下文聚合,以提取密集特征。WT-MVSNet(Liao 等,2022)提出基于窗口的 Transformer,用于多视图立体匹配中的局部特征匹配和全局特征聚合。

2.2　基于神经隐式表征的三维重建

传统多视图几何重建方法虽然提供了更高精度的重建结果,但是复杂动态场景中视图间相差大使得特征点匹配不足等问题往往会导致重建失败。而基于神经网络的方法可以有效地从图像中学习三维结构,不仅可以取得更好的静态三维重建结果,还可以实现动态三维场景重建。

目前,面向复杂动态场景的无人移动视觉三维重建问题,已经有大量的研究采用神经网络的方法。本文根据三维结构的表达方法从基于体积表示的三维重建方法、基于面表示的三维重建方法和基于隐式表示的三维重建方法进行总结。

2.2.1　基于体积表示的三维重建方法

体素(volume pixel)作为常见的三维体积表示方法,是二维像素在三维空间的扩展,表示为三维空间中的立方体,具有位置、颜色等属性。体素表示方法是将三维几何结构离散化为三维体素网格形式存储,通过每个体素的属性来描述三维几何结构。其中基于二值占用体素网格的方法是通过神经网络学习三维体素网格上二值张量的概率分布,被占用的体素属性为 1,否则是属性为 0 的空位置(Wu 等,2015;Girdhar 等,2016;Xie 等,2019);概率占用体素网格中每个体素的属性值表示该体素被占用的概率(Kar 等,2017;Tulsiani 等,2018)。体素网格采取固定分辨率会导致计算复杂和内存压力,一些方法采取八叉树及其他改进的存储结构来支持更高分辨率(Tatarchenko 等,2017;Wang 等,2017)。

符号距离函数(sign distance function,SDF)表示方法将三维空间中的每个体素的属性设置为其到最近表面的有符号距离:正值表示在物体内部,负值表示在几何外部。为了减少计算和存储的开销,截断符号距离函数(truncated signed distance function,TSDF)在符号距离函数基础上以较小的正值和负值对距离值进行截断,仅在存储靠近物体表面的体素内存储有符号距离值(Curless 和 Levoy,1996;Newcombe 等,2011)。基于体素实现的符号距离函数广泛用于三维重建(Zeng 等,2017),但是体素的存储和计算成本仍然将重建结果限制在较低分辨率,Canelhas 等人(2016)采取降维的方法压缩截断符号距离体。近年来更多的方法借助隐式表示实现符号

距离函数，如Park等人（2019）提出的DeepSDF方法实现了更复杂的拓扑结构和更高的分辨率。

由于体素的位置固定约束了其对复杂动态场景的描述，需要依赖额外的变形场作为辅助。重建的时候通常将当前帧通过变形场变形到固定参考帧上，并采用类似静态重建的方式在参考帧上进行几何融合，实现动态三维重建（Huang等，2024）。Dou等人（2016）提出Fusion4D方法通过嵌入式变形场（embedded deformation，ED）实现了实时非刚性动态场景重建，Dou等人（2017）提出Motion2Fusion方法，通过同时包含几何和光度信息的非刚性对齐策略，相比Fusion4D方法可以更有效应对复杂拓扑结构变化。

2.2.2　基于面表示的三维重建方法

点云作为一种数据结构，由三维空间中的离散点构成。通过将这些离散点分布于三维几何结构的表面，能够精确表征该几何结构的形状。Qi等人（2017）提出PointNet方法使用最大池化操作提取输入点云的全局形状特征，后续大量点云生成网络的编码器借鉴该思路（Fan等，2017；Achlioptas等，2018）。针对动态场景重建往往采取时空点云的方式，为点云中每个点加入轨迹信息（Wittner等，2015）。

Kerbl等人（2023）提出高斯泼溅的方法将点云表示为三维空间中各向异性的高斯分布，然后通过抛雪球算法进行渲染来自监督学习，在训练时间和外观质量上得到了惊人的效果。由于高斯泼溅具有高分辨率和高效的推理速度，所以在复杂动态场景具有更多发展潜力。Yang等人（2024b）首次将变形场与高斯泼溅相结合，实现了高质量的单目动态重建结果。

由于点云中的点之间缺少连接关系，采用点云表示的表面会存在缺陷、不平滑的现象。而网格的表示方法通过顶点、边构成面片来描述三维结构，相邻点之间具有连接关系，结构细节更丰富，最近很多方法直接优化纹理网格的几何形状和外观。当然直接重建显式的纹理网格是非常困难的，特别是针对具有复杂拓扑结构的场景。基于网格表示的研究大多数方法，如Wang等人（2018）提出的Pixel2Mesh方法和Chen等人（2021b）提出的Dib-r++方法，都是基于一个假设的固定拓扑结构的网格模板。近期的方法已经开始解决拓扑优化的问题。Munkberg等人（2022）提出的NVdiffrec方法将可微的行进四面体与可微渲染相结合，通过图像的色度损失直接优化网格。此方法还具备分解材质和光照的能力，并且在Hasselgren等人（2022）提出的NVdiffrecMC方法中使用蒙特卡洛渲染得到进一步改进。针对动态场景表示，网格表示方法通过每个面片的顶点的运动轨迹进行描述。

2.2.3　基于隐式表示的三维重建方法

基于隐式表示的三维重建方法通过神经网络参数化连续的隐式函数来表达物体的几何和纹理信息，可以通过输入三维空间采样点和观察的角度信息进行获取。隐式函数的呈现形式包括占用网络（occupancy networks）、符号距离函数（signed distance function，SDF）和神经辐射场（neural radiance fields，NeRF）。

Mescheder等人（2019）使用深度神经网络分类器实现占用网络，允许从单视图、点云和体素输入中重建三维结构。Park等人（2019）提出DeepSDF方法，使用多层感知机来回归SDF分布表示三维表面，需要稀疏的点云进行监督。受DeepSDF等方法启发，Mildenhall等人（2020）提出NeRF方法，核心思想是将场景的三维结构表示为一个函数。该方法中函数通过多层感知机来实现，输入场景采样点的三维坐标和采样角度信息，输出颜色和密度信息，根据输出的颜色和密度信息体积渲染得到的预测图像与输入图像计算损失进行自监督学习。

由于NeRF带来的巨大研究价值，出现了大量在动态场景中采取神经辐射场进行重建的工作。Pumarola等人（2021）提出的D-NeRF方法和Park等人（2021）提出Nerfies方法都通过优化额外的连续体积变形场来增强NeRF，但此类方法很难恢复复杂的运动，训练和渲染速度非常慢。Song等人（2023）提出NeRFPlayer方法，根据动态场景的时间特征使用分解场将每个采样点划分为静态的内容、变形的内容和新出现的内容3种类别，每种类别采用单独的神经场表示，实现了10 s/帧的交互式渲染速度。为了追求更高的训练和渲染效率以及内存利用率，一些方法采用与显式表示相结合的方式进行实现，Fridovich-Keil等人（2023）提出K-Planes方法将动态神经辐射场的四维体分解为6个特征平面进行表示，前3个平面对应空间维度，后3个描述时空变化，实现了高质量重建结果。

2.3 三维点云配准重建

2.3.1 基于传统的点云配准方法

表 8 为传统点云配准方法对比。传统的点云配准方法主要分为粗配准和精配准两个阶段。粗配准的目的是为精配准提供良好的初始变换值；而精配准则是在给定初始变换的基础上进行优化，以获得更精确的变换。

表 8 传统点云配准方法对比

Table 8 Comparison of traditional point cloud registration methods

类型	方法	优点
粗配准	RANSAC	鲁棒性强，对局部噪声和离群点有良好抵抗能力
	4PCS	简单易用，适合粗配准场景
	Super4PCS	精度高，处理速度快
	SAC-CoT	能够处理复杂场景，精度高
	Ltgv	极高的配准精度，适合细致配准
精配准	3D-NDT	在大规模场景中表现优异，鲁棒性好
	ICP	迭代过程收敛性好，支持多种变换模型
	Go-ICP	结合全局和局部优化，精度更高

在粗配准方法的研究领域内，RANSAC 策略（Schnabel 等，2007）备受瞩目。其中，4PCS 技术（Aiger 等，2008）尤为突出，它依赖于对比四点集交叉对角线的比例来识别几乎共面的对应四点集。为了提升效率和性能，Super4PCS（Mellado 等，2014）对 4PCS 方法进行了优化，显著降低了等四点集采样的计算复杂性，从二次时间复杂度降低到线性复杂度，极大地推动了其在实际场景中的应用。此外，SAC-CoT（sample consensus by sampling compatibility triangles）（Yang 等，2021）是一种基于图模型和兼容性三角形采样的 RANSAC 变体算法，用于解决三维点云配准中的噪声和离群点问题。它通过建立点云之间的图模型关系，并采样兼容性三角形，从而识别出可靠的点对应关系。与传统 RANSAC 相比，SAC-CoT 在存在大量噪声和离群点的情况下表现更为鲁棒，能够提高配准的准确性。另外，Ltgv（loose-tight geometric voting）（Yang 等，2022b）方法利用松紧结合的几何投票机制，从匹配的特征点中挑选出高质量对应关系，从而实现快速、鲁棒的点云配准。这种高效的粗配准方法为后续的精细配准奠定了良好的基础，在各类三维感知任务中都有重要应用价值。

另外，基于特征的配准方法也在粗配准中占据一席之地。该方法的核心在于在三维点云上生成特征描述符，并基于这些描述符实现配准。当特征描述符在刚性变换下保持稳定时，具有相同描述符的点对可以被视为对应点对，从而用于计算刚性变换。

在基于特征的配准方法中，Martin 等人（2001）提出的 3D-NDT（three-dimensional normal-distributions transform）方法是一个重要里程碑。该方法将点云划分为多个等大小的网格，并用正态分布来表示这些网格，通过匹配每个网格的正态分布来实现点对的匹配，无需逐点寻找对应点。此外，Wu 等人（2022）提出的 SingleMatch 算法也为精确配准提供了新的思路。该算法首先识别出具有稳定方向的特征点，然后使用卷积神经网络为这些点生成描述符。然后，通过描述符的匹配，算法能够获取候选的对应点对，从而实现精确的配准。此外，Zhang 等人（2023d）提出基于极大团（maximal cliques，MAC）的三维配准方法。该方法不仅利用了图的特性来准确描述匹配之间的亲和关系，还通过挖掘更多的局部信息来生成高质量的假设。MAC 流程图如图 10 所示。

在精配准技术领域，迭代最近点（iterative clos-

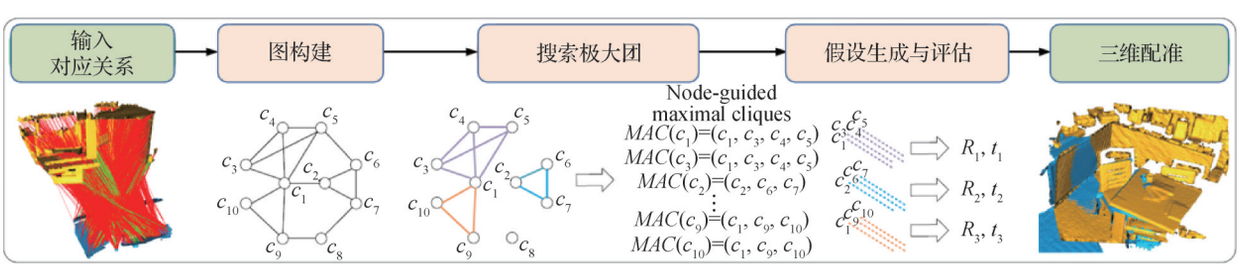


图 10 MAC 流程图

Fig. 10 MAC pipeline

est point, ICP)算法是众所周知的经典方法。它通过迭代地将一个点云中的每个点关联到另一个点云中的最近点,并利用最小二乘法计算点对距离来优化目标函数。然而,ICP的局限性在于它仅依赖空间坐标进行点对搜索,对初始设定敏感,容易在噪声或离群点的影响下收敛到错误的极小值。因此,ICP在初始刚性变换准确、噪声低且无离群点的情况下表现最佳,而在噪声扫描配准和全局配准等复杂场景中则可能不适用。

为了克服ICP的局限性,研究者提出多种改进算法。Sharp等人(2002)引入点云旋转不变性,开发了ICPIF(iterative closest point invariant features)算法,通过构建旋转不变的特征点提高了匹配的鲁棒性。Bouaziz等人(2013)提出稀疏ICP算法,通过引入p范数惩罚项优化配准模型,提升了匹配点的精度。Yang等人(2016)的Go-ICP算法则基于全局特征匹配,采用分支界限法求解最优旋转平移矩阵,有效避免了局部极小值问题。Zhou等人(2016)的FGR(fast global registration)方法利用Geman-McClure估计量处理噪声,通过迭代优化提升了算法速度。杨沐杰等人(2022)则将灰狼优化算法与差分进化算法结合,提出一种改进的配准方法,通过优化旋转矩阵参数提高了ICP的配准精度。王太学等人(2022)则融合主成分分析(principal components analysis, PCA)对ICP进行改进,先通过PCA测量点云主位置获得初始变换,再调整主轴基向量进行配准,进一步提升了算法的准确性和效率。此外,Mutual Voting(Yang等,2023)方法利用相互投票的机制从匹配的特征点中挑选出高质量的对应关系,从而实现快速、鲁棒的点云精细配准。在这种方法中,每个3D点都根据其附近的特征描述符进行评分和投票,最终得票最高的点对成为最佳匹配。这种方法能够有效应对噪声、部分遮挡等常见问题,提高3D点云匹配的准确性和鲁棒性。

传统点云配准方法面临的主要挑战在于粗配准和精配准两个阶段各自存在的问题。粗配准虽然能大致对齐点云,但配准精度相对较低,且网络复杂度较高,效率不够理想;而精配准方法,虽然能进一步提高配准精度,但对初始化和噪声非常敏感,容易陷入局部最优解,从而限制了其在实际应用中的效果。因此,传统点云配准通常需要采用先粗配准后精配准的策略,这种串行处理的方式不仅增加了配准流程的复杂性,也降低了整体的配准效率。这些都是传统点云配准方法需要克服的重要问题。

2.3.2 基于深度学习的点云配准方法

表9为深度学习点云配准方法对比。随着深度学习技术的飞速发展,研究者开始将其应用于点云配准领域(武越 等,2023b)。然而,点云作为一种非结构化的三维数据,其无序性和非结构化特性给深度学习带来了挑战。传统的点云配准方法通常需要将点云转化为规则的体积表示,不仅增加了内存开销,还可能损失点云的细节特征。

表9 深度学习点云配准方法对比

Table 9 Comparison of deep learning-based point cloud registration methods

方法	优点
PointNet	直接处理无序点云,解决了置换不变性问题
PointNetLK	提升了配准精度,结合了优化算法
PCRNet	速度快,鲁棒性强,占用内存少
DCP	结合局部与全局信息,精度高

为解决这一问题,Qi等人(2017)提出PointNet网络,该网络通过直接处理点云数据,利用最大池化提取特征,成功解决了点云的置换不变性和无序性等问题,为深度学习在点云配准中的应用奠定了基础。随后,基于PointNet的深度学习点云配准方法不断涌现。例如,Aoki等人(2019)提出PointNetLK算法,该算法通过PointNet提取点云特征,并结合改进的Lucas-Kanade算法迭代优化转换参数,提升了配准精度。Sarode等人(2019)的PCRNet算法则利用孪生结构的PointNet编码点云形状信息,通过数据驱动估计姿态,对噪声具有鲁棒性且配准速度快。Wang和Solomon(2019)的DCP算法则利用动态图卷积网络和注意力机制获取点云的局部和全局信息,实现了高精度的点云配准。

尽管这些深度学习点云配准方法在提高配准精度和效率方面取得了显著成果,但仍面临一些挑战。例如,它们对噪声和奇异值等因素较为敏感,配准准确率有待进一步提升。此外,在部分点云配准任务中,由于部分点云的缺失,可能导致配准精度下降。因此,未来的研究需要进一步探索更加鲁棒和高效的深度学习点云配准方法。

总之,点云配准作为计算机视觉和三维数据处

理中的重要任务，已经经历了从传统方法到深度学习方法的转变。尽管传统方法在某些应用场景中仍有优势，但深度学习方法凭借其强大的特征提取能力和端到端学习模式，展示了更广阔的应用前景和发展潜力。未来，随着算法和计算资源的不断进步，点云配准技术将会在更多实际场景中得到应用和验证。

3 场景分割

随着无人移动系统的广泛应用，对复杂动态场景的理解和分析变得日益重要。场景分割作为计算机视觉领域的关键技术，对无人移动系统在复杂环境中的感知和决策具有至关重要的作用（汪文靖等，2024）。本节将综述近年来在面向复杂动态场景的无人移动视觉技术中，场景分割领域的研究进展。

场景分割旨在将场景图像中的每个像素归类到特定的语义类别，如道路、建筑物和行人等。在无人驾驶、智能监控和工业自动化等应用中，场景分割技术为无人移动系统提供了丰富的环境感知信息，是实现自主导航、避障和目标跟踪等功能的基础。近年来，深度学习技术为场景分割领域带来了显著的突破。特别是全卷积网络（Long 等，2015）的提出，使场景分割任务能够以端到端的方式进行训练和学习，极大地提高了分割的准确性和效率。随后，多种基于全卷积网络的改进模型相继提出。这些模型通过引入跳层连接、多尺度特征融合以及空间上下文信息捕获等策略，进一步提升了场景分割的精度。具体而言，它们利用多个分辨率的特征图，并在不同分辨率之间进行信息交互和融合，从而实现对场景中细小物体的精确分割。此外，一些研究工作还通过构建多尺度特征金字塔（Zhao 等，2017）、引入膨胀卷积（Chen 等，2018）、通道以及空间上的注意力机制（Zhao 等，2018a；Guo 等，2022）等方法，捕获不同尺度的上下文信息，增强场景分割的准确性和鲁棒性。

尽管基于深度学习的场景分割方法取得了显著的成果，但在处理复杂动态场景时仍面临诸多挑战。例如，动态场景中物体的快速移动、遮挡和光照变化等因素都可能对场景分割的准确性产生不利影响。同时，实时性和计算资源的限制也是实际应用中需要考虑的重要问题。针对这些挑战，现有研究主要围绕高效实时、基于多模态融合和低可见度条件下的场景分割等方向展开。

3.1 高效实时的场景分割

表 10 为高效实时的场景分割方法对比。在无人移动系统中，要求系统能够在复杂动态场景中快速处理并做出响应，因此实现高效实时的场景分割至关重要。在无人移动系统中，如何在保证实时性的同时保持较高的分割精度是高效实时场景分割技术需要解决的关键问题。为了满足无人移动系统对实时性的高要求，研究者致力于开发更高效、更快速的场景分割算法，包括设计轻量级网络结构、基于知识蒸馏训练提取轻量级模型等。

3.1.1 基于轻量级网络架构设计的实时场景分割

轻量级网络结构是高效实时场景分割的重要研究方向之一。通过减少网络参数、降低网络深度或宽度等方式，轻量级网络结构能够显著降低计算复杂度，提高算法的运行速度。一些研究者通过引入深度可分离卷积、逐点卷积等轻量化技术，设计了高效的轻量级网络结构，在保持较高分割精度的同时，具有较低的计算复杂度和较快的运行速度，使其适用于无人移动系统中的场景分割任务。

为解决传统深度神经网络在实时场景分割任务中计算量大、运行时间长的问题。Paszke 等人（2016）专为实时场景分割设计了一种深度神经网络架构 ENet，通过减小特征图分辨率来降低计算量，但考虑到场景分割任务对细节信息的需求，采用了空洞卷积来扩大感受野，收集更多的上下文信息，在保持较高精度的同时，显著提高了处理速度。

为解决实时场景分割任务中空间信息保留和感受野扩大之间的矛盾，Yu 等人（2018）设计了一种双分支网络结构 BiSeNet（bilateral segmentation network），包括空间路径（spatial path，SP）和上下文路径（context path，CP）。其中，空间路径采用小步长卷积，以保留原始输入图像的空间尺寸，并编码丰富的空间信息。上下文路径使用轻量级模型（进行快速下采样，以获得较大的感受野。最终融合空间路径和上下文路径的特征用于最终的分割，能够在不损失过多计算成本的前提下提高分割精度。进一步地，Yu 等人（2021）提出 BiSeNet V2，设计了一个双边引导聚合层，用于增强两个分支（细节分支和语义分支）之间的连接，并更有效地融合两种类型的特征表示。此外，还引入了一种增强训练策略，该策略在训

表10　高效实时的场景分割方法对比

Table 10　Comparison of efficient and real-time scene segmentation methods

数据集	方法	PQ	FPS/(帧/s)	FLOPs/G	参数量/M
Cityscapes(Cordts等2016)	Mask2Former(Cheng等,2022)	62.1	4.1	519	44.0
	UPSNet(Xiong等,2019)	59.3	7.5	–	–
	LPSNet(Hong等,2021)	59.7	7.7	–	–
	PanDeepLab(Cheng等,2020a)	59.7	8.5	–	–
	FPSNet(De Geus等,2020)	55.1	8.8	–	–
	RealTimePan(Hou等,2020)	58.8	10.1	–	–
	YOSO(Hu等,2023)	59.7	11.1	265	42.6
	PEM(Cavagnero等,2024)	61.1	13.8	237	35.6
ADE20k(Zhou等,2017)	BGRNet(Wu等,2020b)	31.8	–	–	–
	MaskFormer(Cheng等,2021)	34.7	29.7	86	45.0
	Mask2Former(Cheng等,2022)	39.7	19.5	103	44.0
	kMaxDeepLab(Yu等,2022)	42.3	–	–	–
	YOSO(Hu等,2023)	38.0	35.4	52	42.0
	PEM(Cavagnero等,2024)	38.5	35.7	47	35.6

注:“–”表示无相应数据。

练阶段可以增强特征表示,而在推理阶段可以丢弃,几乎不增加额外的计算复杂度。后续,Fan等人(2021)对BiSeNet进行重新思考和改进,通过引入短时密集连接模块和细节引导模块,进一步提高了BiSeNet的性能。

尽管双分支网络结构在实时场景分割任务中显示出其高效性和有效性,但直接融合高分辨率细节和低频上下文的方法存在一个问题,即细节特征容易被周围的上下文信息所淹没,限制了分割精度的提高。Xu等人(2023a)通过引入比例—积分—微分(proportional-integral-derivative,PID)控制器的概念,揭示了双分支网络本质上等价于一个比例—积分(proportional-integral,PI)控制器,存在类似的超调问题。设计了一种新颖的实时场景分割网络架构,包含3个分支,分别用于解析细节、上下文和边界信息,并利用边界注意力来指导细节和上下文分支的融合,实现了推理速度和准确性之间的最佳权衡。

Shi等人(2024)提出一种轻量级的上下文感知网络,通过优化网络结构和减少模型参数,在网络中引入了部分通道变换技术,通过对通道进行有选择性的变换和组合,有效降低了计算复杂度和模型参数量,实现了实时场景分割。

近年来,基于Transformer的架构在场景分割领域取得了显著成果,但这些架构通常需要大量的计算资源,尤其是在处理大输入特征时效率低下,这限制了它们在边缘设备上的应用。为了降低计算成本同时保持高性能,Zhang等人(2022)提出Topformer,通过将输入图像分割成不同尺度的图像块,并构建图像块金字塔结构,能够高效提取图像中的多尺度特征,同时保持较低的计算成本。这使得Topformer在移动设备和资源受限的环境中能够实时进行高质量的语义分割任务。此外,Cavagnero等人(2024)通过引入基于原型的交叉注意力机制和高效的多尺度特征金字塔网络(feature pyramid network,FPN)(如图11所示)成功降低了基于Transformer的场景分割架构的计算成本(表10),同时保持了高性能。

3.1.2　基于知识蒸馏的实时场景分割模型

近年来在场景分割领域,知识蒸馏作为一种有效的模型压缩和加速技术展现了其独特的优势,通过将大型教师模型(teacher model)的知识传递给小型学生模型(student model),从而在保持较高分割精度的同时,大幅降低计算复杂度和运行时间。

知识蒸馏首先训练一个高精度的大型教师模型,随后利用该模型的预测结果或中间层特征作为

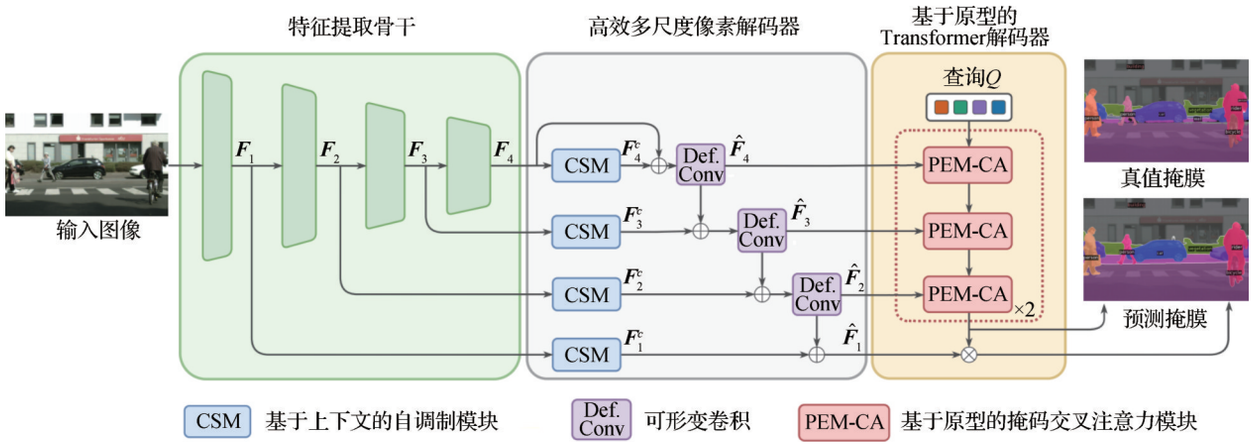


图 11　高效实时的场景分割模型图（Cavagnero 等，2024）

Fig. 11　Graph of efficient and real-time scene segmentation model (Cavagnero et al. , 2024)

监督信息，指导轻量级学生模型的训练。Liu 等人（2019）在这一框架下，引入了成对蒸馏与整体蒸馏的概念，不仅关注像素级别的精细匹配，还强调了结构级别和高阶特征的一致性，显著提升了场景分割任务中知识迁移的效果。

传统方法虽然能有效传递全局上下文知识，但往往忽视了类内特征变化的重要性。Wang 等人（2020）通过深入分析教师模型中同类实例间的特征差异，将这些变化作为关键知识传递给学生模型，增强了学生模型处理类内多样性的能力，特别是在面对具有显著类内差异的类别时，显著提升了分割精度。随后，Shu 等人（2021）提出通道级激活图归一化与对齐的新思路，通过最小化教师与学生模型通道概率图之间的 KL（Kullback-Leibler）散度，使学生模型能够更精准地学习关键区域的特征，优化了密集预测任务的性能，突破了以往仅在空间域对齐激活图的局限。Ji 等人（2022）的工作强调了低层次纹理知识在表征局部结构模式和全局统计属性中的关键作用，通过整合图像的结构性和统计性纹理信息，进一步提升了学生模型的分割性能，证明了多层次特征融合的重要性。为了充分利用教师模型中的全局语义关系，Yang 等人（2022a）提出跨图像关系知识蒸馏方法。该方法不仅关注单个图像内的结构化信息，还通过构建和传递不同图像间的像素与区域关系，结合内存库机制建模更广泛的像素依赖，显著提升了学生模型的分割性能，拓展了知识蒸馏的边界。针对传统蒸馏方法在处理高差异区域时的不足，Gou 等人（2024）引入了差异感知和双重掩码机制。该机制能够有效识别并强化学生模型在教师模型预测结果中的高差异区域的学习，显著提高了学生模型的分割准确性和鲁棒性，在多个基准数据集上取得了优异表现。

3.2　基于多模态融合的场景分割

表 11 为多模态融合的场景分割方法对比。在复杂动态场景的理解与解析领域内，多模态融合技术已跃居为研究前沿的热点，通过集成来自多元传感器的丰富数据资源，包括但不限于摄像头捕捉的 RGB 图像、雷达探测的深度信息、激光雷达生成的点云数据，以及热红外相机记录的红外图像等，构建了一个全方位、多维度的信息感知网络。这种跨模态的信息融合策略，极大地丰富了场景解析的维度，显著提升了场景分割的精度与鲁棒性，使之能够更准确地捕捉并解析复杂环境中的细微变化与动态特征。

在深度图像（RGB image with depth，RGB-D）场景分割领域，研究者通过整合 RGB 图像与深度图信息，实现了分割性能的大幅提升。Hazirbas 等人（2017）的 FuseNet 模型开创了编码器—解码器架构在 RGB-D 融合中的先河，证明了多模态信息融合的有效性。随后，Wang 和 Neumann（2018）提出的深度感知的卷积神经网络进一步挖掘了深度信息对于空间结构理解的贡献，显著增强了模型的感知能力。Hu 等人（2019a）在此基础上引入注意力机制，使得模型能够更高效地利用 RGB-D 数据中的互补特征，进一步优化了分割结果。Zhou 等人（2023）的 BCI-Net（bilateral cross-modal interaction network）网络作为专为 RGB-D 场景设计的框架，如图 12 所示，在复杂环境下展现出了卓越的性能。

表 11　多模态融合的场景分割方法对比

Table 11　Comparison of scene segmentation methods based on multi-modal fusion

数据集	方法	mAcc/%	PixAcc/%	mIoU/%
NYUv2(Silberman 等,2012)	DMFNet(Yuan 等,2019)	59.27	74.42	46.79
	TSNet(Zhou 等,2021c)	59.60	73.50	46.10
	CANet(Zhou 等,2021a)	63.80	76.60	51.20
	Shapeconv(Cao 等,2021)	63.50	76.40	51.30
	BCINet(Zhou 等,2023)	66.78	77.17	52.95
SUN(Song 等,2015)	RedNet(Jiang 等,2018)	60.30	81.30	47.80
	RDFNet(Lee 等,2017)	60.10	81.50	47.70
	CANet(Zhou 等,2021a)	60.50	82.50	49.30
	Shapeconv(Cao 等,2021)	59.20	82.20	48.60
	BCINet(Zhou 等,2023)	61.44	87.62	49.57

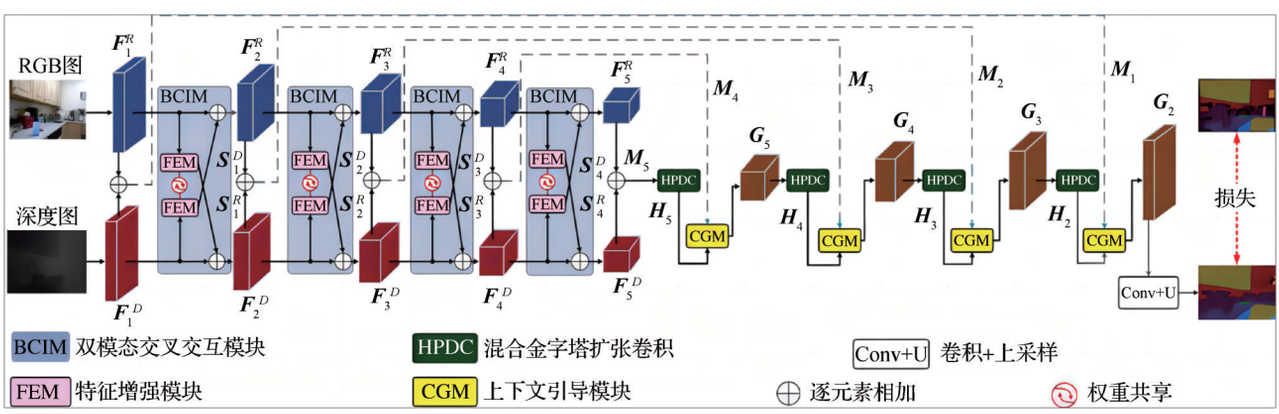


图 12　基于多模态融合的场景分割模型图(Zhou 等,2023)

Fig. 12　Scene segmentation based on multi-modal fusion (Zhou et al., 2023)

单一的 RGB-D 信息往往难以满足分割任务的需求。为此,研究者开始探索 RGB-T 多模态融合方法,利用热红外图像的光照不敏感特性作为补充。Ha 等人(2017)率先建立了 RGB-T 场景分割的基准,验证了热信息在提升分割准确性方面的关键作用。随后,Sun 等人(2019)、Shivakumar 等人(2020)及 Sun 等人(2021)的研究将编码器—解码器理念应用于 RGB-T 融合,开发了高效的融合架构,显著提高了城市场景的分割精度。Zhang 等人(2021c)提出的桥接—融合架构,通过优化跨模态特征的选择与融合策略,进一步推动了分割性能的提升。Zhou 等人(2022)则聚焦于多尺度特征融合与多监督学习,为 RGB-T 场景分割提供了更加全面的解决策略,而 Zhou 等人(2021b)提出的 GMNet(graded-feature multilabel-learning network)网络则展示了多监督架构在促进特征融合方面的巨大潜力。此外,Fu 等人(2022)提出的交叉引导融合注意力网络,通过注意力机制引导不同模态特征的有效融合,实现了更为精细的分割效果。Dong 等人(2022b)则利用图形分支捕捉高级线索和图像上下文,为分割任务提供了丰富的语义信息,进一步提升了分割的精度与鲁棒性。

3.3　低可见度条件下的场景分割

表 12 为低可见度条件下的场景分割方法对比。在低可见度条件下,尤其是夜间或低光照环境中,场景分割任务面临巨大挑战。为解决这一问题,近年来研究者探索了多种创新方法,显著提升了在低光照条件下的场景分割性能。这些方法主要集中在多尺度特征融合、多模态信息增强、无监督域自适应及特定算法框架设计等方面。

表 12　低可见度条件下的场景分割方法对比

Table 12　Comparison of scene segmentation methods under low visibility conditions

数据集	方法	mIoU/%
MFNet（Ha等，2017）	RTFNet（Sun等，2019）	54.8
	GMNet（Zhou等，2021b）	57.7
	ABMDRNet（Zhang等，2021d）	55.5
	LASNet（Li等，2023a）	58.7
	TokenFusion（Wang等，2022b）	58.7
	CMX（Zhang等，2023b）	57.8
	SMMCL（Dong和Yokoya，2024）	60.0
LLRGBD（Zhang等，2021b）	SA-Gate（Chen等，2020）	61.79
	ShapeConv（Cao等，2021）	63.26
	CEN（Wang等，2023d）	62.15
	TokenFusion（Wang等，2022b）	64.75
	CMX（Zhang等，2023b）	66.52
	SMMCL（Dong和Yokoya，2024）	68.76

3.3.1　多尺度特征融合与多模态信息增强

为了有效应对低可见度条件下信息缺失的问题，研究者提出多尺度特征融合策略，通过整合不同尺度下的图像信息，提升模型对复杂光照条件的适应性。同时，融合多模态信息成为另一重要方向，如利用红外图像、热成像等技术，不仅增强了场景的信息量，还显著提高了分割的准确性和鲁棒性。Wang等人（2022a）的SFNET-N（semantic flow network-N）框架便是这一思路的典范，其通过光照增强网络补偿曝光不足导致的信息损失，结合语义分割网络，实现了在低光照条件下的高效分割。

3.3.2　数据集合成与扩增

鉴于夜间场景分割数据集的稀缺性，研究者利用仿真平台和生成对抗网络（GAN）等技术进行数据集扩展。Wang等人（2022a）及Testolina等人（2023）分别通过自主驾驶仿真平台和CARLA（car learning to act）模拟器构建了大规模合成数据集，这些数据集不仅丰富了场景和类别的多样性，还通过风格迁移等方法提升了数据的真实性和适用性，为模型训练提供了有力支持。

3.3.3　特定算法框架设计

为了进一步提高夜间场景分割的精度，研究者设计了多种专用算法框架。Deng等人（2022）提出的NightLab框架采用两级架构，先通过初步分割识别困难区域，再针对这些区域进行精细分割，有效提升了分割效果。Xie等人（2023）则另辟蹊径，利用频域信息，通过动态调整频率成分重要性和捕捉空间频率长期依赖关系，增强了模型对夜视场景的理解能力。

3.3.4　无监督域自适应方法

针对夜间场景标注数据稀缺的问题，无监督域自适应方法成为研究热点。Gao等人（2022）提出的CCDistill（cross-domain correlation distillation）框架通过构建多源多目标域适应网络，探索图像间的照明或内在差异不变性，有效解决了夜间图像缺少标签的问题。Wang等人（2023c）进一步提出OICS（online informative class sampling）策略和InforMS（informativeness-based cross-domain mixed sampling）框架，通过在线信息类采样和基于信息度的跨域混合采样，实现了更加均匀的预测效果。Brüggemann等人（2023）的Refign方法则通过引入参考域概念和非参数修正模块，提升了伪标签质量，优化了迁移学习效果。

3.3.5　跨模态域适应与多模态对比学习

针对多模态数据在夜间语义分割中的应用，Xia等人（2023）提出利用事件相机图像和原相机图像进行跨模态域适应，有效提升了夜间语义分割性能。Dong和Yokoya（2024）则通过有监督的多模态对比学习，如图13所示，在类相关性监督下联合执行跨模态和模态内对比，增强了多模态特征空间的语义可辨别性，进一步提升了低光照场景的理解能力。

4　目标检测识别

4.1　无人移动视觉目标检测的重要性和挑战

目标检测是计算机视觉领域中的一项基础任务，旨在从图像或视频中识别和定位感兴趣的目标。与传统的图像分类任务不同，目标检测不仅要判断图像中是否含有特定类别的物体，还需要确定这些物体在图像中的具体位置，通常用边界框（bounding box）来表示。无人移动视觉目标检测是计算机视觉和人工智能领域的关键技术之一，它对于无人系统如自动驾驶汽车、无人机、机器人等具有至关重要的作用。无人移动视觉目标检测面临许多挑战，主要包括以下几个方面：1）环境复杂性：多变的光照条

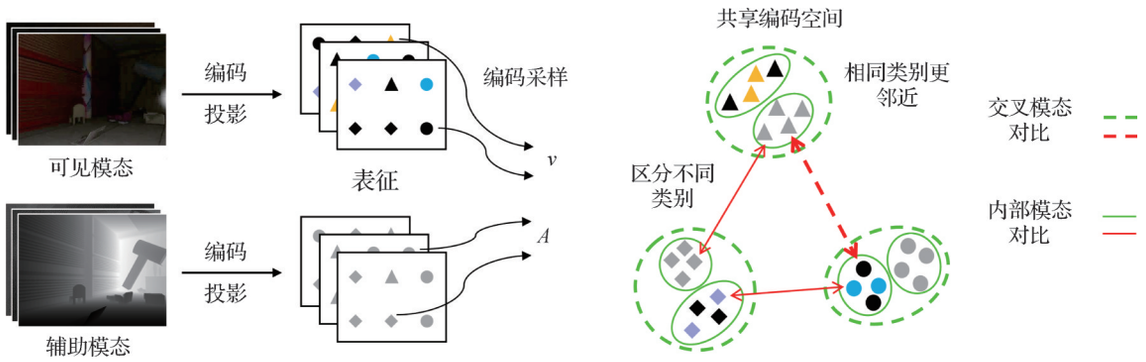


图 13　光照受限条件下的场景分割模型图(Dong 和 Yokoya, 2024)

Fig. 13　Image of scene segmentation model under limited illumination conditions (Dong and Yokoya, 2024)

件,包括光线的强弱、方向和颜色的变化等降低了图像的质量。恶劣的天气条件,如雨、雪和雾,会降低图像的清晰度,进一步增加目标检测的难度。2)目标特性:小目标由于在图像中占据的像素较少,常常易被检测系统忽略或错误识别,导致检测不准确。此外,目标遮挡问题,即目标物体部分被其他物体遮挡或与其他目标重叠,会显著增加检测任务的难度。同时,目标的多样性,包括不同类别和尺度的变化,要求检测算法必须具备广泛的适应性和精细的区分能力。3)实时性要求:由于无人移动设备的计算资源相对有限,系统需要在极短的时间内完成图像的获取、处理以及对目标的识别和定位,以保证在动态环境中的及时响应。4)数据不平衡:在现实世界的数据集中,某些类别的目标物体可能远多于其他类别,对于无人移动视觉某些特定的应用场景,获取足够的标注数据可能非常困难,即使在数据量充足的情况下,不同类别之间的实例数量也可能存在显著差异,这会影响检测模型的泛化能力和公平性。

4.2　无人移动视觉目标检测的关键技术

4.2.1　目标检测算法

基于卷积神经网络的目标检测算法自 2013 年以来经历了快速发展,可以分为一阶段(one-stage)和两阶段(two-stage)检测器。

两阶段检测器,如区域卷积神经网络(region-based convolutional neural network, R-CNN)系列(Girshick 等, 2014; Girshick, 2015; He 等, 2017; Ren 等, 2017; Cai 和 Vasconcelos, 2018; Lu 等, 2019; Pang 等, 2019)首先使用区域建议网络(region proposal network, RPN)或选择性搜索算法提取候选区域,然后对这些区域进行分类和边界框回归。R-CNN(Girshick 等, 2014)通过在不同尺度下生成图像金字塔来处理多尺度问题,尽管其速度较慢,但开启了基于深度学习的目标检测新篇章。紧接着,Fast R-CNN(Girshick, 2015)和 Faster R-CNN(Ren 等, 2017)相继提出,通过引入感兴趣区域(region of interest, RoI)池化层和 RPN,提高了计算效率,并能够同时处理多尺度目标。FPN(Lin 等, 2017)在 Faster R-CNN 的基础上,使用多层特征融合技术,增强了多尺度特征表示,提高了对小目标的检测能力。两阶段目标检测算法首先生成候选区域,然后对这些区域进行分类和边界框回归。这种方法通常在精度上优于一阶段检测器,但速度相对较慢。

表 13 为一阶段经典目标检测器在 COCO(common objects in context)(Lin 等, 2014)数据集上定量

表 13　一阶段经典目标检测器在 COCO(Lin 等, 2014)数据集上定量实验结果

Table 13　The one-stage classical object detector quantifies the experimental results on the COCO (Lin et al., 2014) dataset

研究方法	mAP@IoU = 0.5:0.95	参数量/M	Latency	FLOPs
YOLOv4-tiny	24.9	6.1	–	6.9
YOLOv5-N	28.0	1.9	–	4.5
YOLOv6-N	37.0	4.7	2.69	11.4
YOLOX-S	39.6	9.0	9.80	26.8
YOLOv7-tiny	35.2	6.2	–	5.8
Gold-YOLO-N	39.6	5.6	2.92	12.1
YOLOv8-N	37.3	3.2	6.16	8.7
YOLOv9-S	46.8	7.1	6.16	26.4
YOLOv10-N	38.5	2.3	1.84	6.7

注:"–"表示源文献未提供。

实验结果。

一阶段检测器，如 YOLO（you only look once）系列（Redmon 等，2016；Redmon 和 Farhadi，2017，2018；Bochkovskiy 等，2020；Ge 等，2021；Li 等，2022a；Xu 等，2022；Wang 等，2023b，2024a，b）和 SSD（single shot multi-box detector）（Liu 等，2016）将目标检测任务视为一个回归问题，直接在输入图像上预测边界框和类别概率，省去了区域建议的步骤，从而显著提高了检测速度。2016 年，YOLOv1（Redmon 等，2016）以其快速的检测速度和简化的检测流程，为单阶段目标检测算法树立了标杆。同年，SSD（Liu 等，2016）通过特征金字塔网络（FPN）结构在多个尺度的特征图上进行检测，能够同时识别不同大小的目标。YOLOv1 到 YOLOv10（Wang 等，2024a）的各个版本都对速度和精度进行了持续优化。YOLO 系列在速度和精度上都取得了显著的进展，广泛应用于实时视频监控、自动驾驶以及无人机监控等多个领域。CNN 模型通过其强大的特征提取能力，为无人系统提供了稳定的目标检测解决方案。

然而，随着目标检测任务的复杂性增加，CNN 模型在处理大范围上下文信息方面存在局限。Transformer（Vaswani 等，2017）模型通过自注意力机制能够处理长距离依赖关系，在理解图像全局上下文方面表现出色，在目标检测任务中展现出处理不同类型和尺度对象的灵活性，尤其在复杂的视觉场景中。DETR（detection Transformer）模型是 Transformer 应用于目标检测的开创性工作，它完全基于注意力机制来实现端到端的目标检测，无需任何手工设计的组件。之后，一系列基于 Transformer 的模型相继提出，如 Deformable DETR（Zhu 等，2021）在 DETR 的基础上引入了可变形卷积，以增强模型对目标形状的适应性，适用于无人系统的目标检测。Conditional DETR（Meng 等，2021）通过引入条件变分自编码器（conditional variational autoencoder，CVAE）的思想，为 DETR 模型增加了条件生成能力。DAB-DETR（dynamic anchor boxes detection Transformer）（Liu 等，2022b）将锚框的概念引入到 DETR，为模型提供了位置先验，加速了模型的收敛。DN-DETR（denoising detection Transformer）（Li 等，2022b）是一种在 Transformer 中引入噪声的模型，旨在通过噪声增强来提高模型的泛化能力。DINO（detection Transformer with improved denoising anchor boxes）（Zhang 等，2023a）采用一种对比去噪训练方法，这种方法通过考虑硬负样本来提高模型的训练效率和最终的检测性能。DQ-DETR（dual query detection Transformer）（Huang 等，2024）解决了 DETR 系列不适合航空数据集的问题，提出动态调整用于目标检测查询数量，以解决不同航拍图像之间实例数量不平衡和微小目标检测的问题。Zhao 等人（2023）使用轻量级网络 MobileViT 作为骨干网络，结合了 CNN 的空间归纳偏置和 Transformer 的全局建模能力，减少了模型参数和复杂度。表 14 为基于 Transformer 的经典目标检测器在 COCO（Lin 等，2014）数据集上的定量实验结果。

表 14　基于 Transformer 的经典目标检测器在 COCO（Lin 等，2014）数据集上定量实验结果

Table 14　Quantitative experimental results of the classical object detector based on Transformer on COCO（Li et al.，2014）dataset

研究方法	mAP@IoU = 0.5∶0.95	参数量/M	FPS/（帧/s）	Gflops/G
DETR（Zhu 等，2021）	42	41	–	86
Deformable DETR（Zhu 等，2021）	43.8	40	28	–
Conditional DETR（Meng 等，2021）	40.9	44	19	–
DAB-DETR（Liu 等，2022b）	45.7	44	–	216
DN-DETR（Li 等，2022b）	43.4	48	–	265
DINO（Zhang 等，2023a）	50.9	47	23	279
RT-DETR（Zhao 等，2024b）	53.1	42	5	136
EfficientDETR（Yao，2021）	45.1	35	108	210
LW-DETR-xlarge（Chen，2024c）	58.3	118	–	–

注：“–”表示源方法未提供。

4.2.2 数据增强

在无人移动视觉目标检测中，数据增强技术通过生成多样化的训练数据，可以提高模型的泛化能力和鲁棒性。常见的几何变换（Simard等，2003）方法包括旋转、平移、缩放和翻转，可以帮助模型学习不同角度、位置和大小下的目标特征。颜色变换（Nguyen等，2014）则通过调整亮度、对比度、色调和饱和度，增强模型在不同光照条件下的适应能力。随机裁剪（Srivastava等，2014）技术通过裁剪图像的一部分，模拟遮挡场景，提高模型在部分目标缺失情况下的检测能力。此外，添加高斯噪声、椒盐噪声（Zhang等，2016）等方法，能够增强模型对噪声的鲁棒性。混合增强技术如Mixup（Zhang等，2018a）和CutMix（Yun等，2019），通过将两幅图像进行线性混合或部分粘贴，增强模型对图像间过渡和部分遮挡的理解能力。

生成对抗网络（Goodfellow等，2020）近年来在数据增强领域中显示出巨大潜力。通过训练一个生成器生成逼真的图像数据，并通过一个判别器来提高生成图像的质量，GAN能够创建大量高质量的合成图像，用于训练目标检测模型。Park等人（2020）使用CycleGAN生成高质量的无人机野火图像，有助于克服数据稀缺和不平衡的挑战。Cores等人（2023）提出一种新的生成模型，通过降采样GAN来增强视频数据集中小目标的训练集。该方法针对现有数据集中小目标数量不足的问题，生成高质量的小目标合成图像。与通过插值方法生成的小目标相比，这种方法能够更好地保持对象的细节和特征，显著提高了目标检测模型在小目标检测任务中的表现。Lee等人（2022）提出一种名为RDAGAN（residual dense module and attention-guided generative adversarial networks）的模型，包含两个主要部分：目标生成网络和图像翻译网络，对象生成网络生成在输入数据集中目标的图像，而图像翻译网络则将这些生成的图像与干净的图像合并。Zhang等人（2021d）提出一种无监督的数据增强方法，使用WGAN（Wasserstein GAN）框架生成高质量的合成图像。模型通过两个主要步骤实现：首先训练GAN生成与目标图像域无区别的假图像，然后进一步训练GAN生成带有目标对象的图像，条件是输入位置的约束。

Fang等人（2024）针对目标检测任务中的数据增强需求，该研究提出一种基于可控扩散模型的数据增强管道，通过CLIP实现合成图像的后过滤，以提高目标检测的性能。Chen等人（2024b）设计了一种新颖的文本提示嵌入过程，利用冻结的CLIP文本编码器对每个对象的类别名称进行单独编码。随后，这些编码被堆叠，并由一个新引入的可训练文本嵌入编码器进一步处理，以生成用于控制图像生成的文本条件。GEODiffusion能够编码边界框以及额外的几何条件，如自动驾驶场景中的摄像机视图。Wang等人（2024d）提出一种新颖的框架，通过结合生成模型和感知模型来增强数据生成和感知能力。该框架利用感知模型的信息来增强生成模型的控制生成能力，并通过定制的数据增强来提升特定感知模型的性能。Feng等人（2024）通过集成预训练的扩散模型和实例级定位头部，自动生成具有精确边界框的合成图像。它采用两步训练策略，先在现有数据集上进行监督学习以识别基础类别，然后通过自训练扩展到新类别，从而提高目标检测器在开放词汇和数据稀疏场景下的性能。这种方法不仅增强了目标检测的准确性和鲁棒性，还显著提升了在复杂环境中的检测能力，尤其是在真实数据受限的情况下。

4.2.3 多模态数据融合

由于环境光线变化、遮挡、恶劣天气以及复杂场景等因素的影响，单一模态的图像数据往往难以满足复杂场景下的目标检测需求。为了克服这些挑战，多模态图像融合技术应运而生，通过整合来自不同传感器（如可见光相机、红外相机、雷达等）的数据，增强目标检测的性能（Dasgupta等，2022）。无人移动视觉目标检测的多模态数据融合技术是一项关键的研究领域，旨在利用来自不同传感器或数据源的多种数据类型，如图像、视频、激光雷达和红外图像等结合和融合，以提高目标检测的性能和鲁棒性。这些技术不仅扩展了数据来源的多样性，还能够弥补各种数据的局限性，为无人移动视觉系统提供更全面的感知能力（Chen等，2022b；Sun等，2022；许可等，2024）。

在处理多模态数据融合时，模态差异问题尤为突出，因为不同模态的数据可能在光照条件、分辨率和感知范围等方面存在本质上的差异。Yuan和Wei（2024）提出一种校准和互补的变压器结构，用于RGB-IR目标检测。C2Former通过跨模态交叉注意力模块（intermodality cross-attention，ICA）和自适应特征采样模块（adaptive feature sampling，AFS），解决

了模态错位和融合不精确的问题。Wang 等人(2024c) 提出一种基于图像特征的无人机跨模态目标检测方法，如图 14 所示，旨在通过设计光照感知模块和不确定性感知模块来增强无人机航空图像的跨模态目标检测能力。这种方法特别关注于利用不同模态在不同光照和环境条件下的优势特征，并通过结合 YOLOv5 的轻量级结构和旋转检测头来提高检测性能和保留物体的方向信息。Zhao 等人(2024a) 提出一种粗到细的融合策略，通过频域冗余频谱移除模块(redundant spectrum removal，RSR) 和动态特征选择模块(dynamic feature selection，DFS)，在 RGB 和 IR 图像的融合过程中减少冗余信息，并提高多尺度特征的选择性，从而提升物体检测的准确性。Zhao 等人(2024a) 提出冗余信息抑制网络 (redundant information suppression network，RISNet)，通过最小化 RGB 外观特征和红外辐射特征之间的互信息，减少跨模态冗余信息，增强多模态信息的互补优势。Zhang 等人(2023c) 提出一种融合 RGB 和热成像(RGB-T)技术的 YOLO 网络结构，用于无人机场景下的目标检测。该方法在白天和夜晚均能有效检测目标，并具有较低的误报率，表明其在实际应用中的鲁棒性。

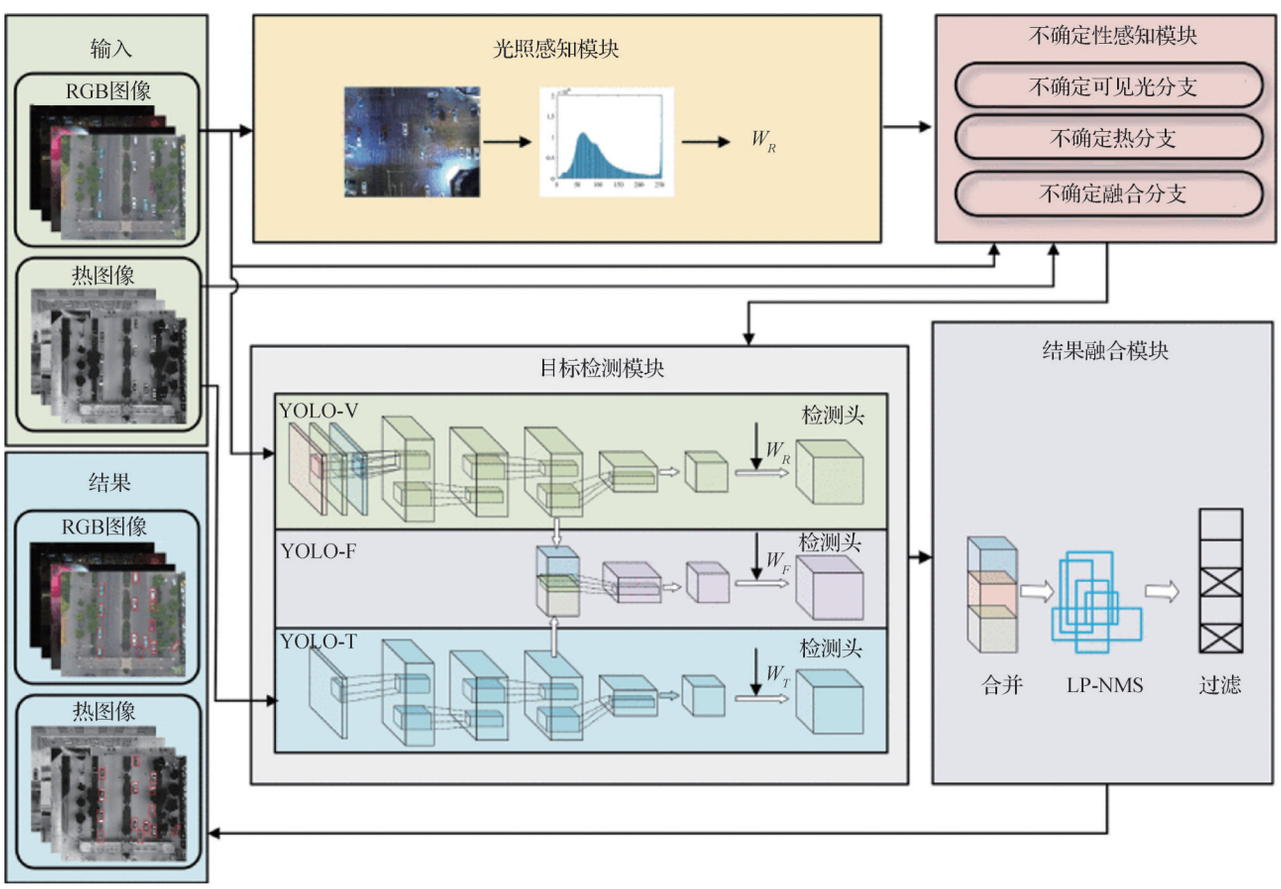


图 14　基于多模态融合的目标检测模型图(Wang 等，2024c)

Fig. 14　Object detection based on multi-modal fusion(Wang et al. ，2024c)

弱对齐问题是由于多模态图像对之间的分辨率和视场差异引起的多模态图像对之间的弱未对准问题限制了其在目标检测中的性能。Yuan 等人(2024) 使用模态选择策略，选择高质量的标注作为参考边界框，然后对两种模态的区域特征进行模态校准过程，以执行特征对齐。大多数现有方法往往忽略模态差距并强调严格对齐。针对此问题，Chen 等人(2024a) 通过解耦多模态特征到模态不变特征和模态特定特征，并预测一个在共同子空间中优越的空间偏移。这种分布对齐的映射建立一个跨模态的共同子空间，以估计精确的特征级偏移，可以用于更准确地评估多模态特征之间的空间一致性。

4.2.4　开放世界目标检测

尽管现有技术在封闭集目标检测上取得了显著

进展，但在面对开放世界中的未知类别和复杂场景时，其性能往往受限。开放世界目标检测是指在没有预先定义类别的情况下，能够识别和检测出图像中所有可能的目标。这与传统的目标检测不同，后者通常只关注于特定的、已知的类别。在自动驾驶、机器人导航等领域，系统必须能够理解和适应不断变化的环境。开放目标检测为这些系统提供了识别未知障碍物和目标的能力。

Joseph等人(2021)首次定义了开放世界目标检测问题，并提出一个名为ORE(open world object detector)的新颖方法来解决这一挑战，该方法通过两阶段训练策略，首先构建无偏特征表示以检测已知类别，然后利用多视图自标签和一致性约束来优化对未知目标的检测。结合知识蒸馏和典型复现技术，有效减轻了在增量学习过程中对已知类别的灾难性遗忘。作为一个端到端的框架，Dong等人(2022a)集成了目标检测、新类别分类和目标性评分，并通过联合损失函数进行训练，支持模型在接收新标注数据时进行有效的增量学习，使其在持续变化的环境中保持高性能目标检测能力。GLIP模型通过统一对象检测和短语定位任务，构建了一个能够处理多模态信息的预训练框架。它利用深度融合策略，将图像的视觉特征与语言的语义信息紧密结合，从而学习到对象级、语言感知且语义丰富的视觉表征。这种表征使得Li等人(2022c)在零样本和少样本学习方面表现出色，能够识别和定位训练数据中未出现过的新类别目标。Wang等人(2023e)提出一个通用目标检测框架，能够利用多源图像和异构标签空间进行训练，并通过解耦训练和概率校准进一步提升对新类别的泛化能力。Wang等人(2024e)在多个大型词汇数据集上展现出强大的零样本泛化能力，超越了传统有监督方法的性能，同时在封闭世界中也保持了优异的检测效果。Xu等人(2023c)在传统文本驱动的目标检测框架之上进行了创新性扩展，集成了视觉提示(visual prompt)查询机制。该模型使用了即插即用门控感知结构(gated class-scalable perceiver，GCP)，该结构允许模型在维持对已知类别的高泛化能力的同时，增强对细粒度特征的检测能力。

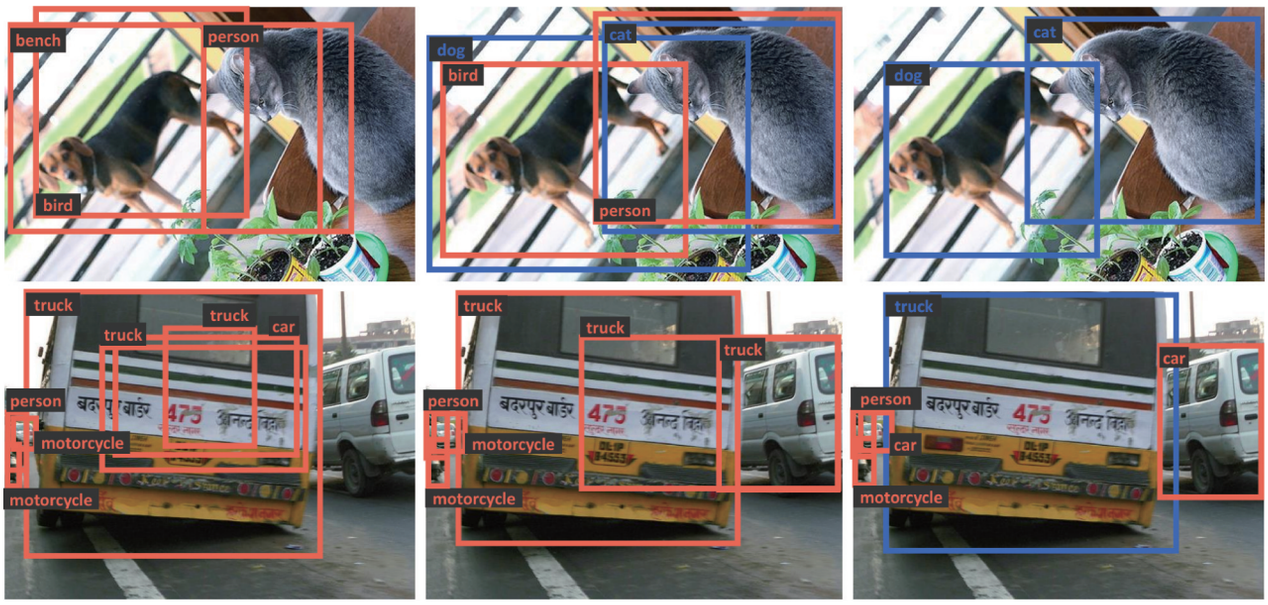
随着时间的推移，开放世界目标检测领域出现了多种模型和方法。这些模型通常专注于提高模型的泛化能力，使其能够处理未知类别，并通过增量学习不断适应新的类别。

4.2.5 实时性能优化

无人系统，包括但不限于自动驾驶汽车、无人机监控系统以及工业自动化机器人，均需依托于高效能的目标检测机制以实现即时决策与反应。实时性是确保系统在多变环境中迅速识别目标并作出相应动作的关键，对于预防碰撞、保障操作安全以及提升任务执行效率具有决定性影响。而效率则涉及算法的执行速率与资源消耗，一个高效的目标检测算法能够有效降低对计算资源的依赖，延长系统运行周期，并在资源受限的设备上实现更为复杂的模型部署。

在过去的数年中，研究者致力于开发轻量级目标检测模型，旨在适应资源受限的计算平台。这些模型通过采用多种网络架构配置，实现了对计算资源的优化。Zhao等人(2024b)设计了一个高效的混合编码器来处理多尺度特征，通过解耦同尺度交互和跨尺度融合来提高处理速度。为了提供高质量的初始查询给解码器，其提出一种新的查询选择机制，通过显式优化不确定性来选择特征，从而提高准确性。Chen等人(2024c)提出一种轻量级的DETR模型，通过减少模型复杂度和参数数量来降低计算需求，使得模型更适合于边缘设备或资源受限的环境。Wang等人(2023a)引入一种先进的GD(gather and distribute)机制，通过卷积和自注意力操作实现高效的信息交换。该机制包含浅层和深层的GD分支，分别通过基于卷积的块和基于注意力的块提取和融合特征信息。为了进一步促进信息流动，其引入了一个轻量级的邻层融合模块，该模块在局部尺度上结合邻近层级的特征，该模型在保持低延迟的同时，实现了高准确率，如图15所示。

Junos等人(2022)采用EfficientNet架构中提出的优化MBConv模块的增强版本。优化后的MBConv用Swish函数取代了ReLU(rectified linear unit)函数，并向网络引入了挤压和激励模块。为了在不影响目标检测精度的情况下满足水下无人平台的轻量化要求，Sun等人(2023)提出一种基于移动视觉变换器(MobileViT)和YOLOX的水下目标检测模型，该模型利用MobileViT作为算法骨干网络，提高了算法的全局特征提取能力，减少了算法参数量。同时使用DCA(dual channel attention)机制，其可以使用最少的参数提高浅层特征的提取以及检测精度。



(a) Faster R-CNN　　(b) Detic　　(c) UniDetector

图 15　开放世界目标检测效果图(Wang 等,2023e)

Fig. 15　Comparison of open-world object detection(Wang et al. ,2023e)((a) Faster R-CNN; (b) Detic; (c) UniDetector)

Du 等人(2023)通过整合稀疏卷积技术和自适应多层掩膜策略(adaptive multi-layer masking strategy,AMM),实现了目标检测模型的轻量化。稀疏卷积限制了卷积运算仅在特征图的稀疏采样区域执行,减少了不必要的计算。Chen 等人(2023b)通过使用深度可分离卷积(depth-wise separable convolution)结构重构骨干网络,有效减少了模型参数。Ye 等人(2023)结合了卷积神经网络和 Transformer,使用轻量级特征提取模块构建网络的骨干部分,通过同质多分支架构减少了计算量和参数数量,从而加快了模型的推理速度;高效卷积 Transformer 块引入了卷积多头自注意力,这种机制通过卷积投影替代传统的多头自注意力中的位置线性投影,有效降低了计算负担,同时保持了性能。Zhou 等人(2024)提出一种创新的双路径轻量级网络,如图 16 所示。它通过并行提取高级语义特征和低级细节,利用轻量级自相关模块和轻量级交叉相关模块捕获全局和局部特征交互,显著提升了检测精度。图 17 为不同目标检测算法的检测对比图(Chen 等,2023b)。

此外,研究人员探索了模型压缩技术,该技术通过减少模型参数和加速推理过程,进一步降低了深度学习模型在实际部署时所需的资源消耗。Zhang 等人(2019)采用一种增量式的模型修剪策略,包括评估每个组件的重要性、移除不重要的组件、微调修剪后的模型以补偿可能的性能下降,并评估微调后的模型以确定其是否适合部署。Mahaur 等人(2023)提出一种基于泰勒准则的网络剪枝技术,有效减少了模型的计算负担,同时保持了高检测精度。Yang 等人(2024a)提出一种进化通道修剪(evolutionary channel pruning,ECP)方法来减少网络中的冗余参数,该方法可以通过为卷积层分配适当的通道数来快速识别优秀的网络架构。

5　异常检测与行为分析

5.1　背景介绍

现代智能城市情境下,监控摄像头、车载记录仪和摄像无人机(Bozcan 和 Kayacan,2021)的普及使得获取的视频片段数量激增。一方面,这些设备获取的视频片段能够让人们可以获取和保存更多的信息,为侦破案件、遏制犯罪以及处理交通事故等提供巨大便利;另一方面,为了更好地利用这些视频片段,及时响应和处理各种突发事件和突发情况,需要花费巨大的人力成本在监视器前进行监视,带来了巨大的成本和人力资源的浪费(Wu 等,2020a,2023a)。随着互联网技术、物联网技术、人工智能与深度学习技术的快速发展,为了能够更好地利用这些视频片段,降低人力资源的成本,提高响应和处理

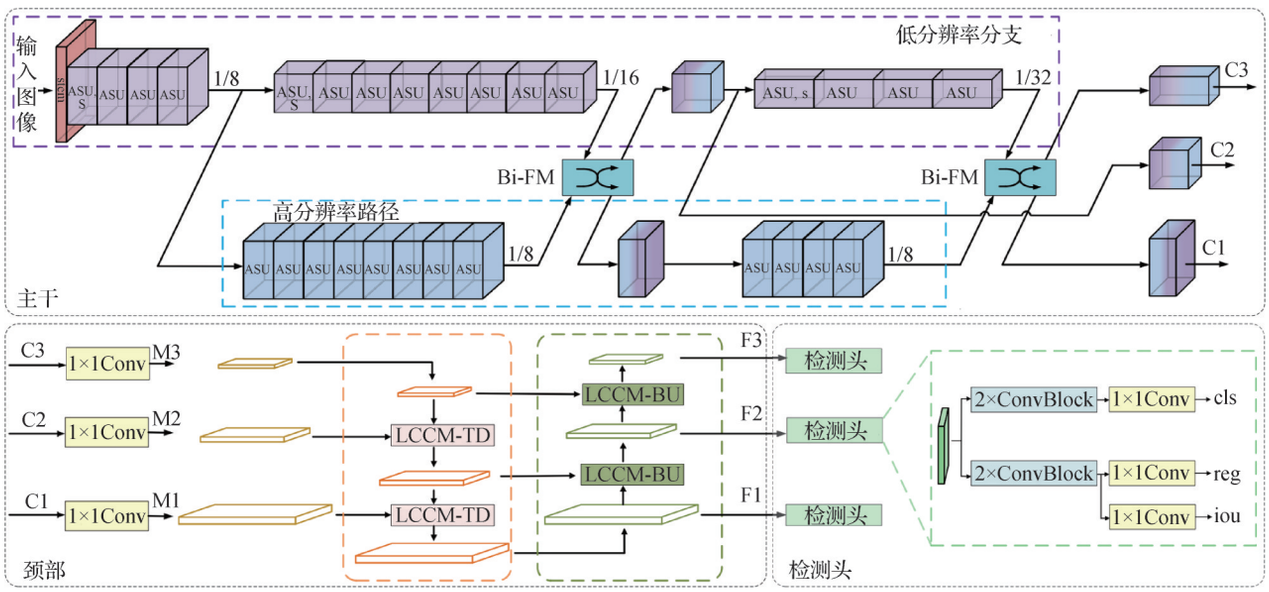


图 16　高效实时的目标检测模型图(Zhou 等,2024)

Fig. 16　Graph of efficient and real-time object detection model(Zhou et al. ,2024)

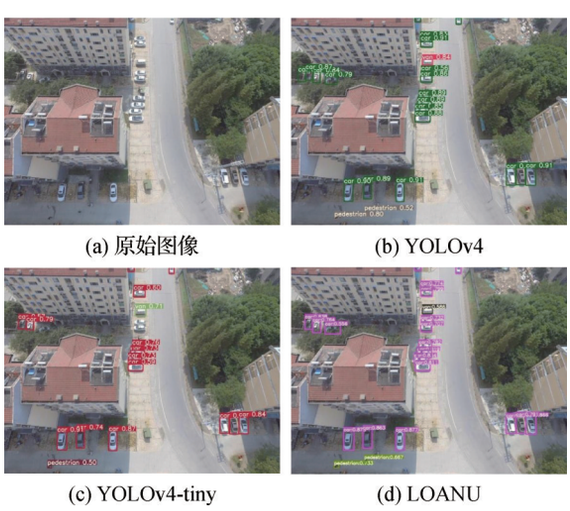


图 17　不同目标检测算法的检测对比图(Chen 等,2023b)

Fig. 17　Detection comparison plots of different object detection algorithms (Chen et al. ,2023b) ((a) original image; (b) YOLOv4; (c) YOLOv4-tiny; (d) LOANU)

各类异常事件的准确度和效率,更好地分析场景中发生的情况和预测场景中可能发生的情况,针对无人移动视觉系统的异常检测和行为分析任务逐渐成为人工智能领域非常重要的研究任务。

无人移动视觉系统中的异常检测主要涉及实时监控和分析视觉数据,以识别和定位异常事件。异常事件可以包括各种犯罪活动、爆炸、火灾、道路上的障碍物、突然出现的行人或车辆以及非正常的驾驶行为等(Srivastava 等,2021)。先进的异常检测算法通常结合了机器学习和深度学习技术,利用大量的正常数据进行训练,以识别异常模式。常用的方法包括卷积神经网络(CNN)、长短期记忆网络(long short-term memory,LSTM)以及生成对抗网络(GAN)等。这些方法能够有效地处理视频流中的时空信息,提高检测的准确性和实时性。

行为分析在无人移动视觉系统中同样关键。它涉及对环境中动态元素(如行人、其他车辆等)的行为进行监控和预测,以便系统能够做出适当的响应。例如,预测行人的运动轨迹、分析其他车辆的驾驶意图等。行为分析通常依赖于时序数据的建模和分析,常用的方法包括轨迹预测模型、行为分类模型和意图推断模型。这些模型可以基于历史行为数据进行训练,利用时空特征对未来行为进行准确预测。

5.2　应用分类介绍

本文提到的无人移动视觉系统包括无人机、自动驾驶、移动机器人、智能工业系统以及智慧城市系统等,主要应用场景有无人机视频异常检测与行为分析、自动驾驶中的交通异常检测与行为分析、智能工业系统中的异常检测与行为分析、智慧城市系统中的异常检测与行为分析。表 15 为几种关键技术的详情对比。

5.2.1　无人机视频异常检测与行为分析

无人机在复杂动态场景中进行视频异常检测和行为分析具有广泛的应用。例如,识别飞行路径中的异常事件,如非法入侵、突发事故等,以及对视频

表 15　几种关键技术的详情对比

Table 15　Detailed comparison of several key technologies

关键技术	经典方法	优点	缺点
异常检测	重构	采用半监督训练方法，训练时无需标签	模型学习能力过强导致可能重构异常
	单分类器	简单易实现，计算效率高，泛化能力强	局限于特定类型的数据，可能无法检测较为复杂的异常
	弱监督异常检测	使用视频级别异常标签，可以提供更加充分的监督信息	缺乏具体的异常定位标签，可能导致模型无法定位异常，出现随机猜测现象
行为分析与识别	弱监督行为识别	提供具体的行为类别标签，便于对行为进行分类和识别	对给定标签外的行为缺乏泛化能力
行为预测	预测	可以更好地利用视频的上下文逻辑信息	人类行为难以预测，误报率较高

中的暴力行为进行检测（Nguyen 等，2022；Srivastava 等，2022）。

5.2.2　自动驾驶中的交通异常检测与行为分析

自动驾驶系统需要检测道路上的异常情况，如障碍物、突然出现的行人或车辆、非正常的驾驶行为等。此外，系统还需要对其他车辆和行人的行为进行分析和预测，以便做出适当的响应（Santhosh 等，2020）。

5.2.3　移动视觉机器人上的异常检测与行为分析

与自动驾驶系统主要关注室外道路场景不同，移动视觉机器人则专注于室内复杂环境中的异常检测和行为分析。其目的是优化在室内环境中的导航能力，并根据特定功能进行行为异常检测、分析和预测。例如，针对老年人的养护机器人可以检测、分析和预测异常行为，以便及时发出预警。

5.2.4　智能工业系统中的异常检测与行为分析

在智能工业系统中，异常检测主要用于识别工业产品和工厂中的异常情况，例如设备故障、产品缺陷等。同时，行为分析用于监控和预测工人和工业机器人的行为，以提高生产效率和安全性（Salhaoui 等，2019；白艳峰 等，2024）。

5.2.5　智慧城市系统中的异常检测与行为分析

智慧城市系统综合了无人机、自动驾驶和智能工业系统等多个应用领域的技术，用于进行城市管理和安全监控。例如，城市交通中的异常检测、公共场所中的暴力行为检测等。图 18 展示了 Liu 等人（2024）提出的无人移动视觉系统。

5.3　应用技术方法

在无人移动视觉系统上进行的异常检测和行为分析与传统的异常检测和行为分析任务既有共性又有差异，需要根据无人移动视觉系统的特点针对性地对传统方法进行改进，以便能够适应不同的工作环境。

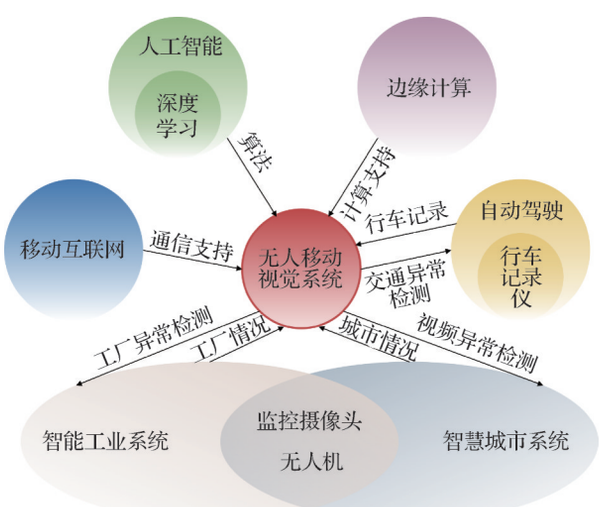


图 18　无人移动视觉系统（Liu 等，2024）

Fig. 18　Unmanned mobile vision system (Liu et al., 2024)

5.3.1　与传统方法的差异

传统方法所使用的数据集一般为固定角度的单摄像头在相对固定的环境所采集到的视频或是从电影等网络视频中采集到的视频或是使用技术模拟生成的视频，视频角度位置固定并且视频质量相对比较清晰，环境较为简单，而无人移动视觉系统视角不固定（如无人机）或摄像头视角差异比常规摄像头视角差异更大，并且面临室外复杂多变的环境（如雨天、大雾、阳光等），视频质量无法保证；并且无人移动视觉系统对于及时响应和处理突发事件，进行行

为分析和预测有着很高的时间要求，并且可能需要在边缘设备上进行检测，这就对模型提出很高的要求，要求模型必须轻量级并且运行速度快，同时也对准确率提出较高的要求。

5.3.2　方法流程概述

为无人机或其他无人移动视觉系统设计的暴力检测系统也遵循着一般暴力检测方法的基本流程，将暴力事件视为异常事件，采用异常检测的方法对视频片段进行二分类任务，确定视频片段中是否存在暴力事件，主要使用RGB图像作为输入，经过提取特征，学习外观/运动模型或交互关系，训练判别器最后输出异常标签进行分类的流程，在中间会使用一些预训练模型提取特征，使用LSTM网络、CNN网络等作为网络架构，使用特征融合、时空建模和关系学习作为方法策略，使用CNN的全连接层或支持向量机（supported vector machine，SVM）作为判别器，如图19所示。

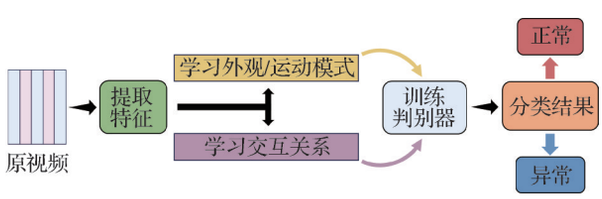


图19　暴力检测方法流程图

Fig. 19　Flow chart of violence detection method

无人移动视觉系统上的视频异常检测和交通异常检测大多遵循传统方法中的无监督或半监督的视频异常检测方法，采用重构的方法进行视频异常检测，重构方法流程为输入RGB视频，使用预训练模型提取特征，使用编码器进行编码，经过重构学习之后进行解码操作，计算重构损失并判断是否异常，主要使用CNN卷积神经网络，并可能在中间加入记忆模块等作为策略，如图20所示。

5.3.3　研究现状概述

在无人机视频异常检测方面，Singh等人（2018）使用特征金字塔网络从航空图像中检测人体，并提出ScatterNet混合深度学习（ScatterNet hybrid deep learning，SHDL）网络，使用人体图像区域进行人体姿势估计并通过姿势方向来识别暴力个体；Srivastava等人（2021）创建了新的无人机暴力检测数据集，并使用2D骨架作为输入，使用SVM和随机森林作为分类模型识别动作进行暴力检测；Srivastava等人（2022）采用结合图像处理技术的CNN模型，在测试中，针对非受限环境开发了无人机捕获的视频数据集，采用包含Inception模块和残差块的混合模型与LSTM架构提取的特征进行暴力检测，最终取得了97.33%的准确率。

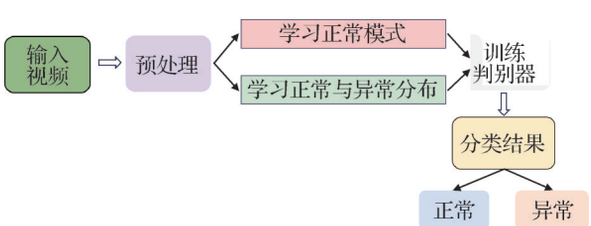


图20　视频异常检测流程图

Fig. 20　Video anomaly detection flow chart

6　问题与发展

6.1　图像增强处理

图像增强处理技术在无人移动视觉技术中的应用愈加广泛，但其在复杂动态场景中的应用仍面临诸多挑战。主要问题包括噪声、模糊和低分辨率。噪声可能源于低光照条件和传感器缺陷，导致图像细节模糊，影响特征提取和目标检测的准确性。模糊通常由运动或焦距引起，导致图像边缘不清晰，影响物体识别和跟踪。低分辨率图像由于传感器限制或远距离拍摄缺乏细节，不利于精细分析。单帧图像增强的信息有限，难以同时解决多种质量问题，而多帧图像增强利用冗余信息能够更好地提升图像质量，但需要有效处理多帧数据的复杂性。未来图像增强处理的发展趋势主要体现在深度学习和实时处理技术的进步上。深度学习方法，如卷积神经网络（CNN）和生成对抗网络（GAN），在图像去噪、去模糊和超分辨率方面展现了强大能力。自监督学习和无监督学习方法正在减少对标注数据的依赖，进一步提高了模型的适应性和泛化能力。随着硬件计算能力的提升，实时图像增强处理成为可能，这对于无人移动设备在复杂动态场景中的应用至关重要。未来的发展还包括融合多种技术，如联合去噪和去模糊的模型，及多模态信息的融合，以提升感知能力。图像增强处理技术的应用领域也在不断拓展，从无人移动设备到医学影像、安防监控、虚拟现实和增强现实等，展示出广泛的应用前景。随着技术的不断进步和应用需求的增加，图像增强处理将在更多领域发挥更大的作用。

6.2 三维重建

基于神经网络在复杂动态场景三维重建方法的性能很大程度上依赖于高质量的标注训练数据，而在动态场景中，获取高质量的三维数据往往非常困难。尽管一些少样本的方法被提出，但是复杂动态场景中存在的大量遮挡和运动模糊限制了神经网络准确捕捉和预测物体的动态行为。未来，进一步研究有效结合预训练大模型和多模态数据融合等技术的动态三维重建方法，可以弥补高质量标注训练数据的不足，整合来自不同传感器和数据源的信息，从而增强模型对复杂动态场景的理解能力，将能实现更高效、更精确的动态三维重建；三维点云配准重建技术在计算机视觉和三维数据处理中扮演着重要角色，但在粗配准和精配准两个阶段面临诸多挑战。粗配准方法如 4PCS 和 Super4PCS 在提供初始变换值时效率不高，精配准方法如 ICP 则对噪声和离群点敏感，容易陷入局部最优解。基于特征的配准方法在特征描述符的稳定性上存在问题，而深度学习方法虽然展示了较强的特征提取能力，但对噪声和部分点云缺失的情况仍有局限。未来的发展方向包括优化粗配准和精配准算法、开发高效特征描述符生成方法、结合传统方法与深度学习技术以及增强算法的鲁棒性，以应对复杂环境中的点云配准任务，提升其精度和效率。

6.3 开放场景下的场景分割

传统场景分割方法通常基于预设的、有限的类别集合进行训练和优化，但在实际应用中，特别是无人移动系统在复杂环境中的运行时，往往会遇到大量未知或罕见的目标类别。这些未知类别可能具有独特的视觉特征，且其数量庞大，远远超出训练数据集所能覆盖的范围。这要求场景分割算法模型需要具备较强的泛化能力和自适应学习能力，能够在线学习新的类别，并快速更新其分割模型，以适应不断变化的场景需求。因此，未来的研究应聚焦于探索开放词汇分割、增量学习以及连续学习等前沿技术，使得模型能在线学习新的类别，并迅速调整其分割模型，以灵活适应不断变化的场景需求。这一能力对于确保无人移动系统在未知环境中的稳定运行至关重要。

6.4 目标检测识别

传统目标检测算法通常在特定条件下训练完成，因此在复杂和多变的开放环境中，其适应性往往较差。尤其在无人系统操作的动态环境中，如无人车在城市交通或无人机在复杂气候条件下的应用，背景和目标的快速变化可能导致这些传统方法的效果大打折扣。为了解决这些问题，开放世界目标检测和类增量持续学习目标检测成为应对无人系统中开放场景目标检测挑战的两种先进方法。开放世界目标检测不仅能够识别已知类别的目标，还具备发现并学习新目标类别的能力，极大地拓展了模型的应用范围。这种能力使得无人系统能够在遇到前所未见的目标时，通过在线学习逐渐适应和识别新的目标类型，而不需要人工重新标注和训练。类增量持续学习则通过在保留已有知识的基础上逐步引入新类别信息，有效地提升了模型的持续学习能力和泛化性。这种方法主要通过技术如经验回放、知识蒸馏或弹性权重合并等策略来缓解模型在学习新类别时出现的灾难性遗忘问题。这使得模型在不断增加新的检测类别的同时，仍能保持对旧类别的高效识别。这些策略显著增强了模型在面对动态变化和未知目标时的适应性和鲁棒性。

6.5 异常检测与行为分析

尽管在异常检测和行为分析方面已经取得了显著进展，但仍然存在许多挑战。环境的复杂性和多变性、数据的高维度和多样性、实时性的要求等都对算法提出更高的要求。此外，如何在保持高精度的同时降低计算开销，以及在不同应用场景中实现算法的泛化能力，仍是当前研究的热点和难点。未来的研究方向可能包括多模态数据融合、增强学习与主动学习的应用、边缘计算的优化，以及更为鲁棒和高效的模型设计。这些发展将进一步提升无人移动视觉系统的智能化水平，使其在更广泛的应用场景中发挥重要作用。

7 结 语

本文全面回顾了面向复杂动态场景的无人移动视觉技术的最新研究进展。无人移动视觉技术作为无人系统的关键技术，对于复杂动态应用场景的理解和适应至关重要。本文从图像增强处理、三维重建、场景分割、目标检测识别以及异常检测与行为分析 5 个关键任务出发，详细介绍了各领域关键技术的最新进展，展示了这些技术在应对复杂动态场景中的潜力和面临的挑战。

总体来看，面向复杂动态场景的无人移动视觉技术是一个快速发展且充满潜力的领域，预计随着技术的不断进步，无人移动视觉技术能够在复杂动态场景中实现更广泛的应用，并为未来的长远发展和实际落地提供坚实的基础。

参考文献（References）

Abuolaim A and Brown M S. 2020. Defocus deblurring using dual-pixel data//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 111-126 [DOI: 10.1007/978-3-030-58607-2_7]

Achlioptas P, Diamanti O, Mitliagkas I and Guibas L. 2018. Learning representations and generative models for 3D point clouds//Proceedings of the 35th International Conference on Machine Learning. Stockholmsmässan, Stockholm, Sweden: PMLR: 40-49

Aiger D, Mitra N J and Cohen-Or D. 2008. 4-points congruent sets for robust pairwise surface registration. ACM Transactions on Graphics (TOG), 27(3): 1-10 [DOI: 10.1145/1360612.1360684]

Aoki Y, Goforth H, Srivatsan R A and Lucey S. 2019. PointNetLK: robust and efficient point cloud registration using PointNet//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 7156-7165 [DOI: 10.1109/CVPR.2019.00733]

Bai Y F, Wang L B, Gao W D and Ma Y L. 2024. Multi-modal hierarchical classification for power equipment defect detection. Journal of Image and Graphics, 29(7): 2011-2023 (白艳峰, 王立彪, 高卫东, 马应龙. 2024. 面向电力设备缺陷检测的多模态层次化分类. 中国图象图形学报, 29(7): 2011-2023) [DOI: 10.11834/jig.230269]

Barnes C, Shechtman E, Finkelstein A and Goldman D B. 2009. PatchMatch: a randomized correspondence algorithm for structural image editing. ACM Transactions on Graphics (TOG), 28(3): #24 [DOI: 10.1145/1576246.1531330]

Bochkovskiy A, Wang C Y and Liao H Y M. 2020. YOLOv4: optimal speed and accuracy of object detection [EB/OL]. [2024-08-05]. https://arxiv.org/pdf/2004.10934.pdf

Bouaziz S, Tagliasacchi A and Pauly M. 2013. Sparse iterative closest point. Computer Graphics Forum, 32(5): 113-123 [DOI: 10.1111/cgf.12178]

Bozcan I and Kayacan E. 2021. Context-dependent anomaly detection for low altitude traffic surveillance//Proceedings of 2021 IEEE International Conference on Robotics and Automation. Xi'an, China: IEEE: 224-230 [DOI: 10.1109/ICRA48506.2021.9562043]

Brüggemann D, Sakaridis C, Truong P and Van Gool L. 2023. Refign: align and refine for adaptation of semantic segmentation to adverse conditions//Proceedings of 2023 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 3173-3183 [DOI: 10.1109/WACV56688.2023.00319]

Cai Z W and Vasconcelos N. 2018. Cascade R-CNN: delving into high quality object detection//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 6154-6162 [DOI: 10.1109/CVPR.2018.00644]

Campbell N D F, Vogiatzis G, Hernández C and Cipolla R. 2008. Using multiple hypotheses to improve depth-maps for multi-view stereo//Proceedings of the 10th European Conference on Computer Vision on Computer Vision. Marseille, France: Springer: 766-779 [DOI: 10.1007/978-3-540-88682-2_58]

Canelhas D R, Schaffernicht E, Stoyanov T, Lilienthal A J and Davison A J. 2016. An eigenshapes approach to compressed signed distance fields and their utility in robot mapping [EB/OL]. [2024-08-05]. https://arxiv.org/pdf/ 1609.02462.pdf

Cao J M, Leng H C, Lischinski D, Cohen-Or D, Tu C H and Li Y Y. 2021. ShapeConv: shape-aware convolutional layer for indoor RGB-D semantic segmentation//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 7068-7077 [DOI: 10.1109/ICCV48922.2021.00700]

Cavagnero N, Rosi G, Cuttano C, Pistilli F, Ciccone M, Averta G and Cermelli F. 2024. PEM: prototype-based efficient maskformer for image segmentation//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 15804-15813 [DOI: 10.1109/CVPR52733.2024.01496]

Chang L H, Feng X C, Li X P and Zhang R. 2016. A fusion estimation method based on fractional Fourier transform. Digital Signal Processing, 59: 66-75 [DOI: 10.1016/j.dsp.2016.07.016]

Chen C, Qi J H, Liu X Y, Bin K C, Fu R G, Hu X K and Zhong P. 2024a. Weakly misalignment-free adaptive feature alignment for UAVs-based multimodal object detection//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 26826-26835 [DOI: 10.1109/CVPR52733.2024.02534]

Chen H Y, Gu J J, Liu Y H, Magid S A, Dong C, Wang Q, Pfister H and Zhu L. 2023a. Masked image training for generalizable deep image denoising//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 1692-1703 [DOI: 10.1109/CVPR52729.2023.00169]

Chen J, Yang Z F, Chan T N, Li H, Hou J H and Chau L P. 2022a. Attention-guided progressive neural texture fusion for high dynamic range image restoration. IEEE Transactions on Image Processing, 31: 2661-2672 [DOI: 10.1109/TIP.2022.3160070]

Chen K, Xie E Z, Chen Z, Wang Y B, Hong L Q, Li Z G and Yeung D T. 2024b. GeoDiffusion: text-prompted geometric control for object detection data generation//Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria: ICLR

Chen L C, Papandreou G, Kokkinos I, Murphy K and Yuille A L. 2018. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transac-

tions on Pattern Analysis and Machine Intelligence, 40(4): 834-848[DOI: 10.1109/TPAMI.2017.2699184]

Chen L Y, Lu X, Zhang J, Chu X J and Chen C P. 2021a. HINet: half instance normalization network for image restoration//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Nashville, USA: IEEE: 182-192[DOI: 10.1109/CVPRW53098.2021.00027]

Chen N Y, Li Y, Yang Z M, Lu Z S, Wang S and Wang J N. 2023b. LODNU: lightweight object detection network in UAV vision. The Journal of Supercomputing, 79(9): 10117-10138[DOI: 10.1007/s11227-023-05065-x]

Chen Q, Su X B, Zhang X Y, Wang J, Chen J H, Shen Y P, Han C C, Chen Z L, Xu W X, Li F R, Zhang S, Yao K, Ding E, Zhang G and Wang J D. 2024c. LW-DETR: a transformer replacement to YOLO for real-time detection[EB/OL].[2024-08-05]. https://arxiv.org/pdf/2406.03459.pdf

Chen W Z, Litalien J, Gao J, Wang Z, Tsang C F, Khamis S, Litany O and Fidler S. 2021b. DIB-R++: learning to predict lighting and material with a hybrid differentiable renderer//Proceedings of the 35th International Conference on Neural Information Processing Systems. [s.l.]: Curran Associates Inc.: #1749

Chen X K, Lin K Y, Wang J B, Wu W, Qian C, Li H S and Zeng G. 2020. Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 561-577[DOI: 10.1007/978-3-030-58621-8_33]

Chen X Y, Wang X T, Zhou J T, Qiao Y and Dong C. 2023c. Activating more pixels in image super-resolution transformer//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 22367-22377[DOI: 10.1109/CVPR52729.2023.02142]

Chen Y T, Shi J H, Ye Z L, Mertz C, Ramanan D and Kong S. 2022b. Multimodal object detection via probabilistic ensembling//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 139-158[DOI: 10.1007/978-3-031-20077-9_9]

Cheng B W, Collins M D, Zhu Y K, Liu T, Huang T S, Adam H and Chen L C. 2020a. Panoptic-DeepLab: a simple, strong, and fast baseline for bottom-up panoptic segmentation//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 12472-12482[DOI: 10.1109/CVPR42600.2020.01249]

Cheng B W, Misra I, Schwing A G, Kirillov A and Girdhar R. 2022. Masked-attention mask transformer for universal image segmentation//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 1280-1289[DOI: 10.1109/CVPR52688.2022.00135]

Cheng B W, Schwing A G and Kirillov A. 2021. Per-pixel classification is not all you need for semantic segmentation//Proceedings of the 35th International Conference on Neural Information Processing Systems. [s.l.]: Curran Associates Inc.: #1367

Cheng S, Xu Z X, Zhu S L, Li Z W, Li L E and Ramamoorthi R. 2020b. Deep stereo using adaptive thin volume representation with uncertainty awareness//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 2521-2531[DOI: 10.1109/CVPR42600.2020.00260]

Cho S J, Ji S W, Hong J P, Jung S W and Ko S J. 2020. Rethinking coarse-to-fine approach in single image deblurring//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 4621-4630[DOI: 10.1109/ICCV48922.2021.00460]

Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, Franke U, Roth S and Schiele B. 2016. The cityscapes dataset for semantic urban scene understanding//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 3213-3223[DOI: 10.1109/CVPR.2016.350]

Cores D, Brea V M, Mucientes M, Seidenari L and Del Bimbo A. 2023. Downsampling GAN for small object data augmentation//Proceedings of the 20th International Conference on Computer Analysis of Images and Patterns. Limassol, Cyprus: Springer: 89-98[DOI: 10.1007/978-3-031-44237-7_9]

Cornillère V, Djelouah A, Wang Y F, Sorkine-Hornung O and Schroers C. 2019. Blind image super-resolution with spatially variant degradations. ACM Transactions on Graphics (TOG), 38(6): #166[DOI: 10.1145/3355089.3356575]

Curless B and Levoy M. 1996. A volumetric method for building complex models from range images//Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques. [s.l.]: ACM: 303-312[DOI: 10.1145/237170.237269]

Dabov K, Foi A, Katkovnik V and Egiazarian K. 2007. Image denoising by sparse 3-D transform-domain collaborative filtering. IEEE Transactions on Image Processing, 16(8): 2080-2095[DOI: 10.1109/TIP.2007.901238]

Dalal N and Triggs B. 2005. Histograms of oriented gradients for human detection//Proceedings of 2005 IEEE Conference on Computer Vision and Pattern Recognition. San Diego, CA, USA: IEEE: 886-893[DOI: 10.1109/CVPR.2005.177]

Dasgupta K, Das A, Das S, Bhattacharya U and Yogamani S. 2022. Spatio-contextual deep network-based multimodal pedestrian detection for autonomous driving. IEEE Transactions on Intelligent Transportation Systems, 23(9): 15940-15950[DOI: 10.1109/TITS.2022.3146575]

De Geus D, Meletis P and Dubbelman G. 2020. Fast panoptic segmentation network. IEEE Robotics and Automation Letters, 5(2): 1742-1749[DOI: 10.1109/LRA.2020.2969919]

Deng X Q, Wang P, Lian X C and Newsam S. 2022. NightLab: a dual-level architecture with hardness detection for segmentation at night//

Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 16917-16927 [DOI: 10.1109/CVPR52688.2022.01643]

Deng Y P, Liu Q and Ikenaga T. 2020. Multi-scale contextual attention based HDR reconstruction of dynamic scenes//Proceedings of the 12th International Conference on Digital Image Processing. Osaka, Japan: SPIE: #115191F [DOI: 10.1117/12.2572977]

Ding Y K, Yuan W T, Zhu Q T, Zhang H T, Liu X Y, Wang Y J and Liu X. 2022. TransMVSNet: global context-aware multi-view stereo network with transformers//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 8575-8584 [DOI: 10.1109/CVPR52688.2022.00839]

Dong N, Zhang Y Q, Ding M L and Lee G H. 2022a. Open world DETR: transformer based open world object detection [EB/OL]. [2024-08-05]. https://arxiv.org/pdf/ 2212.02969.pdf

Dong S H, Zhou W J, Qian X H and Yu L. 2022b. GEBNet: graph-enhancement branch network for RGB-T scene parsing. IEEE Signal Processing Letters, 29: 2273-2277 [DOI: 10.1109/LSP.2022.3219350]

Dong X Y and Yokoya N. 2024. Understanding dark scenes by contrasting multi-modal observations//Proceedings of 2024 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 829-839 [DOI: 10.1109/WACV57701.2024.00089]

Dou M S, Davidson P, Fanello S R, Khamis S, Kowdle A, Rhemann C, Tankovich V and Izadi S. 2017. Motion2fusion: real-time volumetric performance capture. ACM Transactions on Graphics (TOG), 36(6): #246 [DOI: 10.1145/3130800.3130801]

Dou M S, Khamis S, Degtyarev Y, Davidson P, Fanello S R, Kowdle A, Escolano S O, Rhemann C, Kim D, Taylor J, Kohli P, Tankovich V and Izadi S. 2016. Fusion4D: real-time performance capture of challenging scenes. ACM Transactions on Graphics (TOG), 35(4): #114 [DOI: 10.1145/2897824.2925969]

Du B W, Huang Y C, Chen J X and Huang D. 2023. Adaptive sparse convolutional networks with global context enhancement for faster object detection on drone images//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE: 13435-13444 [DOI: 10.1109/CVPR52729.2023.01291]

Erkan U, Engınoğlu S and Thanh D N H. 2019. A recursive mean filter for image denoising//Proceedings of 2019 International Artificial Intelligence and Data Processing Symposium (IDAP). Malatya, Turkey: IEEE: 1-5 [DOI: 10.1109/IDAP.2019.8875957]

Esteban C H and Schmitt F. 2004. Silhouette and stereo fusion for 3D object modeling. Computer Vision and Image Understanding, 96(3): 367-392 [DOI: 10.1016/j.cviu.2004.03.016]

Fan H Q, Su H and Guibas L. 2017. A point set generation network for 3D object reconstruction from a single image//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 2463-2471 [DOI: 10.1109/CVPR.2017.264]

Fan M Y, Lai S Q, Huang J S, Wei X M, Chai Z H, Luo J F and Wei X L. 2021. Rethinking BiSeNet for real-time semantic segmentation//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 9711-9720 [DOI: 10.1109/CVPR46437.2021.00959]

Fang H Y, Han B R, Zhang S, Zhou S, Hu C X and Ye W M. 2024. Data augmentation for object detection via controllable diffusion models//Proceedings of 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Waikoloa, USA: IEEE: 1246-1255 [DOI: 10.1109/WACV57701.2024.00129]

Feng C J, Zhong Y J, Jie Z Q, Xie W D and Ma L. 2024. InstaGen: enhancing object detection by training on synthetic dataset//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 14121-14130 [DOI: 10.1109/CVPR52733.2024.01339]

Fridovich-Keil S, Meanti G, Warburg F R, Recht B and Kanazawa A. 2023. K-Planes: explicit radiance fields in space, time, and appearance//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 12479-12488 [DOI: 10.1109/CVPR52729.2023.01201]

Fu Y P, Chen Q Q and Zhao H F. 2022. CGFNet: cross-guided fusion network for RGB-thermal semantic segmentation. The Visual Computer, 38(9): 3243-3252 [DOI: 10.1007/s00371-022-02559-2]

Fu Z Q, Yang Y, Tu X T, Huang Y, Ding X H and Ma K K. 2023. Learning a simple low-light image enhancer from paired low-light instances//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 22252-22261 [DOI: 10.1109/CVPR52729.2023.02131]

Furukawa Y and Ponce J. 2009. Accurate, dense, and robust multiview stereopsis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32 (8): 1362-1376 [DOI: 10.1109/TPAMI.2009.161]

Galliani S, Lasinger K and Schindler K. 2015. Massively parallel multiview stereopsis by surface normal diffusion//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 873-881 [DOI: 10.1109/ICCV.2015.106]

Gao H, Guo J C, Wang G L and Zhang Q. 2022. Cross-domain correlation distillation for unsupervised domain adaptation in nighttime semantic segmentation//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 9903-9913 [DOI: 10.1109/CVPR52688.2022.00968]

Gao H Y, Tao X, Shen X Y and Jia J Y. 2019. Dynamic scene deblurring with parameter selective sharing and nested skip connections//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 3843-3851 [DOI: 10.1109/CVPR.2019.00397]

Ge Z, Liu S T, Wang F, Li Z M and Sun J. 2021. YOLOX: exceeding

YOLO Series in 2021 [EB/OL]. [2024-08-05]. https://arxiv.org/pdf/ 2107.08430.pdf

Girshick R. 2015. Fast R-CNN//Proceedings of 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE: 1440-1448 [DOI: 10.1109/ICCV.2015.169]

Girdhar R, Fouhey D F, Rodriguez M, Gupta A, Leibe B, Matas J, Sebe N and Welling M. 2016. Learning a predictable and generative vector representation for objects//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 484-499 [DOI: 10.1007/978-3-319-46466-4_29]

Girshick R, Donahue J, Darrell T and Malik J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation// Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE: 580-587 [DOI: 10.1109/CVPR.2014.81]

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y. 2020. Generative adversarial networks. Communications of the ACM, 63(11): 139-144 [DOI: 10.1145/3422622]

Gou J P, Zhou X B, Du L, Zhan Y B, Chen W and Yi Z. 2024. Difference-aware distillation for semantic segmentation. IEEE Transactions on Multimedia, 26: 10069-10080 [DOI: 10.1109/TMM.2024.3405619]

Gu J J, Lu H N, Zuo W M and Dong C. 2019. Blind super-resolution with iterative kernel correction//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 1604-1613 [DOI: 10.1109/CVPR. 2019.00170]

Gu X D, Fan Z W, Zhu S Y, Dai Z Z, Tan F T and Tan P. 2020. Cascade cost volume for high-resolution multi-view stereo and stereo matching//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: 2492-2501 [DOI: 10.1109/CVPR42600.2020.00257]

Guo C L, Li C Y, Guo J C, Loy C C, Hou J H, Kwong S and Cong R M. 2020. Zero-reference deep curve estimation for low-light image enhancement//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 1777-1786 [DOI: 10.1109/CVPR42600.2020.00185]

Guo M H, Lu C Z, Hou Q B, Liu Z N, Cheng M M and Hu S M. 2022. SegNeXt: rethinking convolutional attention design for semantic segmentation//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: #84

Ha Q S, Watanabe K, Karasawa T, Ushiku Y and Harada T. 2017. MFNet: towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes//Proceedings of 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Vancouver, Canada: IEEE: 5108-5115 [DOI: 10.1109/IROS.2017.8206396]

Han D, Li L, Guo X J and Ma J Y. 2022. Multi-exposure image fusion via deep perceptual enhancement. Information Fusion, 79: 248-262 [DOI: 10.1016/j.inffus.2021.10.006]

Hasselgren J, Hofmann N and Munkberg J. 2022. Shape, light, and material decomposition from images using monte carlo rendering and denoising//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: #1661

Hazirbas C, Ma L N, Domokos C and Cremers D. 2017. FuseNet: incorporating depth into semantic segmentation via fusion-based CNN architecture//Proceedings of the 13th Asian Conference on Computer Vision. Taipei, China: Springer: 213-228 [DOI: 10.1007/978-3-319-54181-5_14]

He K M, Gkioxari G, Dollár P and Girshick R. 2017. Mask R-CNN//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 2980-2988 [DOI: 10.1109/ICCV.2017.322]

Hernández C, Vogiatzis G and Cipolla R. 2007. Probabilistic visibility for multi-view stereo//Proceedings of 2007 IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis, USA: IEEE: 1-8 [DOI: 10.1109/CVPR.2007.383193]

Hong W X, Guo Q P, Zhang W, Chen J D and Chu W. 2021. LPSNet: a lightweight solution for fast panoptic segmentation//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 16741-16749 [DOI: 10.1109/CVPR46437.2021.01647]

Hou R, Li J, Bhargava A, Raventos A, Guizilini V, Fang C, Lynch J and Gaidon A. 2020. Real-time panoptic segmentation from dense detections//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 8520-8529 [DOI: 10.1109/CVPR42600.2020.00855]

Hu J, Gallo O, Pulli K and Sun X B. 2013. HDR deghosting: how to deal with saturation?//Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland, USA: IEEE: 1163-1170 [DOI: 10.1109/CVPR.2013.154]

Hu J, Huang L Y, Ren T H, Zhang S C, Ji R R and Cao L J. 2023. You only segment once: towards real-time panoptic segmentation//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 17819-17829 [DOI: 10.1109/CVPR52729.2023.01709]

Hu X X, Yang K L, Fei L and Wang K W. 2019a. ACNET: attention based network to exploit complementary features for RGBD semantic segmentation//Proceedings of 2019 IEEE International Conference on Image Processing (ICIP). Taipei, China: IEEE: 1440-1444 [DOI: 10.1109/ICIP.2019.8803025]

Hu Y T, Zhen R W and Sheikh H. 2019b. CNN-based deghosting in high dynamic range imaging//Proceedings of 2019 IEEE International Conference on Image Processing (ICIP). Taipei, China: IEEE: 4360-4364 [DOI: 10.1109/ICIP.2019.8803532]

Huang Y X, Liu H I, Shuai H H and Cheng W H. 2024. DQ-DETR: DETR with dynamic query for tiny object detection//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 290-305 [DOI: 10.1007/978-3-031-73116-7_17]

Ji D J, Wang H R, Tao M Y, Huang J Q, Hua X S and Lu H T. 2022. Structural and statistical texture knowledge distillation for semantic segmentation//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 16855-16864 [DOI: 10.1109/CVPR52688.2022.01637]

Jiang J D, Zheng L N, Luo F and Zhang Z. 2018. RedNet: residual encoder-decoder network for indoor RGB-D semantic segmentation [EB/OL]. [2024-08-05]. https://arxiv.org/pdf/1806.01054.pdf

Joseph K J, Khan S, Khan F S and Balasubramanian V N. 2021. Towards open world object detection//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2021). Nashville, USA: IEEE: 5826-5836 [DOI: 10.1109/CVPR46437.2021.00577]

Junos M H, Mohd Khairuddin A S and Dahari M. 2022. Automated object detection on aerial images for limited capacity embedded device using a lightweight CNN model. Alexandria Engineering Journal, 61(8): 6023-6041 [DOI: 10.1016/j.aej.2021.11.027]

Kalantari N K and Ramamoorthi R. 2017. Deep high dynamic range imaging of dynamic scenes. ACM Transactions on Graphics (TOG), 36(4): #144 [DOI: 10.1145/3072959.3073609]

Kar A, Häne C and Malik J. 2017. Learning a multi-view stereo machine//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 364-375

Kerbl B, Kopanas G, Leimkuehler T and Drettakis G. 2023. 3D Gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics (TOG), 42(4): #139 [DOI: 10.1145/3592433]

Kuhn A, Lin S and Erdler O. 2019. Plane completion and filtering for multi-view stereo reconstruction//Proceedings of the 41st DAGM German Conference on Pattern Recognition. Dortmund, Germany: Springer: 18-32 [DOI: 10.1007/978-3-030-33676-9_2]

Kumari R and Mustafi A. 2022. Denoising of images using fractional Fourier transform//Proceedings of the 2nd International Conference on Emerging Frontiers in Electrical and Electronic Technologies. Patna, India: IEEE: 1-6 [DOI: 10.1109/ICEFEET51821.2022.9848244]

Lee H, Kang S and Chung K. 2022. Robust data augmentation generative adversarial network for object detection. Sensors, 23(1): #157 [DOI: 10.3390/s23010157]

Lee J, Son H, Rim J, Cho S and Lee S. 2021. Iterative filter adaptive network for single image defocus deblurring//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 2034-2042 [DOI: 10.1109/CVPR46437.2021.00207]

Lee S, Park S and Hong K. 2017. RDFNet: RGB-D multi-level residual feature fusion for indoor semantic segmentation//Proceedings of 2017 IEEE/CVF International Conference on Computer Vision. Venice, Italy: IEEE: 4990-4999 [DOI: 10.1109/ICCV.2017.533]

Li C Y, Li L L, Jiang H L, Weng K H, Geng Y F, Li L, Ke Z D, Li Q Y, Cheng M, Nie W Q, Li Y D, Zhang B, Liang Y F, Zhou L Y, Xu X M, Chu X X, Wei X M and Wei X L. 2022a. YOLOv6: a single-stage object detection framework for industrial applications [EB/OL]. [2024-08-05]. https://arxiv.org/pdf/2209.02976.pdf

Li F, Zhang H, Liu S L, Guo J, Ni L M and Zhang L. 2022b. DN-DETR: accelerate DETR training by introducing query DeNoising//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 13609-13617 [DOI: 10.1109/CVPR52688.2022.01325]

Li G Y, Wang Y K, Liu Z, Zhang X P and Zeng D. 2023a. RGB-T semantic segmentation with location, activation, and sharpening. IEEE Transactions on Circuits and Systems for Video Technology, 33(3): 1223-1235 [DOI: 10.1109/TCSVT.2022.3208833]

Li H, Wu X J and Durrani T. 2020. NestFuse: an infrared and visible image fusion architecture based on nest connection and spatial/channel attention models. IEEE Transactions on Instrumentation and Measurement, 69(12): 9645-9656 [DOI: 10.1109/TIM.2020.3005230]

Li L H, Zhang P C, Zhang H T, Yang J W, Li C Y, Zhong Y W, Wang L J, Yuan L, Zhang L, Hwang J N, Chang K W and Gao J F. 2022c. Grounded language-image pre-training//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 10955-10965 [DOI: 10.1109/CVPR52688.2022.01069]

Li M D, Liu J Y, Yang W H, Sun X Y and Guo Z M. 2018. Structure-revealing low-light image enhancement via robust retinex model. IEEE Transactions on Image Processing. 27(6): 2828-2841 [DOI: 10.1109/TIP.2018.2810539]

Li S, Zhang G X, Luo Z X, Liu J, Zeng Z and Zhang S W. 2022d. From general to specific: online updating for blind super-resolution. Pattern Recognition, 127: #108613 [DOI: 10.1016/j.patcog.2022.108613]

Li X, Li B C, Jin X, Lan C L and Chen Z B. 2023b. Learning distortion invariant representation for image restoration from a causality perspective//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 1714-1724 [DOI: 10.1109/CVPR52729.2023.00171]

Liang Z X, Li C Y, Zhou S C, Feng R C and Loy C C. 2023. Iterative prompt learning for unsupervised backlit image enhancement//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 8060-8069 [DOI: 10.1109/ICCV51070.2023.00743]

Liao J L, Ding Y K, Shavit Y, Huang D H, Ren S H, Guo J, Feng W S and Zhang K. 2022. WT-MVSNet: window-based transformers for

multi-view stereo. Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: #623

Lin T Y, Dollár P, Girshick R, He K M, Hariharan B and Belongie S. 2017. Feature pyramid networks for object detection//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 936-944 [DOI: 10.1109/CVPR.2017.106]

Lin T Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D and Zitnick C L. 2014. Microsoft COCO: common objects in context//Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer: 740-755 [DOI: 10.1007/978-3-319-10602-1_48]

Liu J, Liu Y, Lin J Y, Li J L, Cao L, Sun P, Hu B, Song L, Boukerche A and Leung V C M. 2024. Networking systems for video anomaly detection: a tutorial and survey [EB/OL]. [2024-08-05]. http://arxiv.org/pdf/2405.10347.pdf

Liu J Y, Fan X, Huang Z B, Wu G Y, Liu R S, Zhong W and Luo Z X. 2022a. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 5792-5801 [DOI: 10.1109/CVPR52688.2022.00571]

Liu R S, Ma L, Zhang J A, Fan X and Luo Z X. 2021. Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 10556-10565 [DOI: 10.1109/CVPR46437.2021.01042]

Liu S L, Li F, Zhang H, Yang X, Qi X B, Su H, Zhu J and Zhang L. 2022b. DAB-DETR: dynamic anchor boxes are better queries for DETR//Proceedings of the 10th International Conference on Learning Representations. [s.l.]: ICLR

Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C Y, Berg A C. 2016. SSD: single shot MultiBox detector//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 21-37 [DOI: 10.1007/978-3-319-46448-0_2]

Liu Y F, Chen K, Liu C, Qin Z C, Luo Z B and Wang J D. 2019. Structured knowledge distillation for semantic segmentation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 2599-2608 [DOI: 10.1109/CVPR.2019.00271]

Long J, Shelhamer E and Darrell T. 2015. Fully convolutional networks for semantic segmentation//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 3431-3440 [DOI: 10.1109/CVPR.2015.7298965]

Lowe D G. 2004. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60(2): 91-110 [DOI: 10.1023/B:VISI.0000029664.99615.94]

Lu X, Li B Y, Yue Y X, Li Q Q and Yan J J. 2019. Grid R-CNN//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 7355-7364 [DOI: 10.1109/CVPR.2019.00754]

Lugmayr A, Danelljan M, Timofte R, Ahn N, Bai D, Cai J, Cao Y, Chen J Y, Cheng K H, Chun S, Deng W, El-Khamy M, Ho C M, Ji X Z, Kheradmand A, Kim G, Ko H, Lee K, Lee J, Li H, Liu Z L, Liu Z S, Liu S, Lu Y H, Meng Z B, Michelini P N, Micheloni C, Prajapati K, Ren H Y, Seo Y H, Siu W C, Sohn K A, Tai Y, Umer B M, Wang S Q, Wang H B, Wu T H, Wu H N, Yang B, Yang F Z, Yoo J, Zhao T T, Zhou Y B, Zhuo H J, Zong Z Y and Zou H Y. 2020. NTIRE 2020 challenge on real-world image super-resolution: methods and results//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Seattle, USA: IEEE: 2058-2076 [DOI: 10.1109/CVPRW50498.2020.00255]

Luo K, Guan T, Ju L L, Wang Y, Chen Z and Luo Y W. 2020a. Attention-aware multi-view stereo//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 1587-1596 [DOI: 10.1109/CVPR42600.2020.00166]

Luo Z X, Huang Y, Li S, Wang L and Tan T N. 2020b. Unfolding the alternating optimization for blind super resolution//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: #473

Ma J Y, Tang L F, Xu M L, Zhang H and Xiao G B. 2021. STDFusionNet: an infrared and visible image fusion network based on salient target detection. IEEE Transactions on Instrumentation and Measurement, 70: #5009513 [DOI: 10.1109/TIM.2021.3075747]

Ma L, Ma T Y, Liu R S, Fan X and Luo Z X. 2022. Toward fast, flexible, and robust low-light image enhancement//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 5627-5636 [DOI: 10.1109/CVPR52688.2022.00555]

Mahaur B, Mishra K K and Kumar A. 2023. An improved lightweight small object detection framework applied to real-time autonomous driving. Expert Systems with Applications, 234: #121036 [DOI: 10.1016/j.eswa.2023.121036]

Mao X T, Liu Y M, Liu F Z, Li Q L, Shen W and Wang Y. 2023. Intriguing findings of frequency selection for image deblurring//Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, USA: AAAI: 1905-1913 [DOI: 10.1609/aaai.v37i2.25281]

Martin D, Fowlkes C, Tal D and Malik J. 2001. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics//Proceedings of the 8th IEEE International Conference on Computer Vision. Vancouver, Canada: IEEE: 416-423 [DOI: 10.1109/ICCV.2001.937655]

Mellado N, Aiger D and Mitra N J. 2014. Super 4PCS fast global point-

cloud registration via smart indexing. Computer Graphics Forum, 33(5): 205-215 [DOI: 10.1111/cgf.12446]

Meng D P, Chen X K, Fan Z J, Zeng G, Li H Q, Yuan Y H, Sun L and Wang J D. 2021. Conditional DETR for fast training convergence//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE: 3631-3640 [DOI: 10.1109/ICCV48922.2021.00363]

Mescheder L, Oechsle M, Niemeyer M, Nowozin S and Geiger A. 2019. Occupancy networks: learning 3D reconstruction in function space//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE: 4455-4465 [DOI: 10.1109/CVPR.2019.00459]

Mildenhall B, Srinivasan P P, Tancik M, Barron J T, Ramamoorthi R and Ng R. 2020. NeRF: representing scenes as neural radiance fields for view synthesis//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 405-421 [DOI: 10.1007/978-3-030-58452-8_24]

Munkberg J, Chen W Z, Hasselgren J, Evans A, Shen T C, Müller T, Gao J and Fidler S. 2022. Extracting triangular 3D models, materials, and lighting from images//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 8270-8280 [DOI: 10.1109/CVPR52688.2022.00810]

Murali V and Sudeep P V. 2020. Image denoising using DnCNN: an exploration study//Jayakumari J, Karagiannidis G K, Ma M D and Hossain S A, eds. Advances in Communication Systems and Networks: Select Proceedings of ComNet 2019. Singapore, Singapore: Springer: 847-859 [DOI: 10.1007/978-981-15-3992-3_72]

Nah S, Kim T H and Lee K M. 2017. Deep multi-scale convolutional neural network for dynamic scene deblurring//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 257-265 [DOI: 10.1109/CVPR.2017.35]

Newcombe R A, Izadi S, Hilliges O, Molyneaux D, Kim D, Davison A J, Kohi P, Shotton J, Hodges S and Fitzgibbon A. 2011. KinectFusion: real-time dense surface mapping and tracking//Proceedings of the 10th IEEE International Symposium on Mixed and Augmented Reality. Basel, Switzerland: IEEE: 127-136 [DOI: 10.1109/ISMAR.2011.6092378]

Nguyen K, Fookes C, Sridharan S, Tian Y L, Liu F, Liu X M and Ross A. 2022. The state of aerial surveillance: a survey [EB/OL]. [2024-08-05]. https://arxiv.org/pdf/2201.03080.pdf

Nguyen R M H, Kim S J and Brown M S. 2014. Illuminant aware gamut-based color transfer. Computer Graphics Forum, 33(7): 319-328 [DOI: 10.1111/cgf.12500]

Niu Y Z, Wu J B, Liu W X, Guo W Z and Lau R W H. 2021. HDR-GAN: HDR image reconstruction from multi-exposed LDR images with large motions. IEEE Transactions on Image Processing, 30: 3885-3896 [DOI: 10.1109/TIP.2021.3064433]

Pan L Y, Chowdhury S, Hartley R, Liu M M, Zhang H G and Li H D. 2021. Dual pixel exploration: simultaneous depth estimation and image restoration//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 4338-4347 [DOI: 10.1109/CVPR46437.2021.00432]

Pang J M, Chen K, Shi J P, Feng H J, Ouyang W L and Lin D H. 2019. Libra R-CNN: towards balanced learning for object detection//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE: 821-830 [DOI: 10.1109/CVPR.2019.00091]

Park J J, Florence P, Straub J, Newcombe R and Lovegrove S. 2019. DeepSDF: learning continuous signed distance functions for shape representation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 165-174 [DOI: 10.1109/CVPR.2019.00025]

Park K, Sinha U, Barron J T, Bouaziz S, Goldman D B, Seitz S M and Martin-Brualla R. 2021. Nerfies: deformable neural radiance fields//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 5845-5854 [DOI: 10.1109/ICCV48922.2021.00581]

Park M, Tran D Q, Jung D and Park S. 2020. Wildfire-detection method using DenseNet and CycleGAN data augmentation-based remote camera imagery. Remote Sensing, 12(22): #3715 [DOI: 10.3390/rs12223715]

Paszke A, Chaurasia A, Kim S and Culurciello E. 2016. ENet: a deep neural network architecture for real-time semantic segmentation [EB/OL]. [2024-08-05]. https://arxiv.org/pdf/1606.02147.pdf

Pumarola A, Corona E, Pons-Moll G and Moreno-Noguer F. 2021. D-NeRF: neural radiance fields for dynamic scenes//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 10313-10322 [DOI: 10.1109/CVPR46437.2021.01018]

Qi C R, Su H, Kaichun M and Guibas L J. 2017. PointNet: deep learning on point sets for 3D classification and segmentation//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 77-85 [DOI: 10.1109/CVPR.2017.16]

Redmon J, Divvala S, Girshick R and Farhadi A. 2016. You only look once: unified, real-time object detection//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE 779-788 [DOI: 10.1109/CVPR.2016.91]

Redmon J and Farhadi A. 2017. YOLO9000: better, faster, stronger//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 6517-6525 [DOI: 10.1109/CVPR.2017.690]

Redmon J and Farhadi A. 2018. YOLOv3: an incremental improvement [EB/OL]. [2024-08-05]. https://arxiv.org/pdf/1804.02767.pdf

Ren C L, Xu Q S, Zhang S K and Yang J Q. 2023. Hierarchical prior mining for non-local multi-view stereo//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France:

IEEE：3588-3597［DOI：10.1109/ICCV51070.2023.00334］

Ren S Q，He K M，Girshick R and Sun J. 2017. Faster R-CNN：towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence，39(6)：1137-1149［DOI：10.1109/TPAMI.2016.2577031］

Romanoni A and Matteucci M. 2019. Tapa-MVS：textureless-aware patchmatch multi-view stereo//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul，Korea (South)：IEEE：10412-10421［DOI：10.1109/ICCV.2019.01051］

Salhaoui M，Guerrero-González A，Arioua M，Ortiz F J，El Oualkadi A and Torregrosa C L. 2019. Smart industrial IoT monitoring and control system based on UAV and cloud computing applied to a concrete plant. Sensors，19(15)：#3316［DOI：10.3390/s19153316］

Santhosh K K，Dogra D P and Roy P P. 2020. Anomaly detection in road traffic using visual surveillance：a survey. ACM Computing Surveys (CSUR)，53(6)：#119［DOI：10.1145/3417989］

Sarode V，Li X Q，Goforth H，Aoki Y，Srivatsan R A，Lucey S and Choset H. 2019. PCRNet：point cloud registration network using PointNet encoding［EB/OL］.［2024-08-05］. https://arxiv.org/pdf/1908.07906.pdf

Satapathy L M，Das P，Shatapathy A and Patel A K. 2019. Bio-medical image denoising using wavelet transform. International Journal of Recent Technology and Engineering，8(1)：2479-2484

Schnabel R，Wahl R and Klein R. 2007. Efficient RANSAC for point-cloud shape detection. Computer Graphics Forum，26(2)：214-226［DOI：10.1111/j.1467-8659.2007.01016.x］

Sen P，Kalantari N K，Yaesoubi M，Darabi S，Goldman D B and Shechtman E. 2012. Robust patch-based hdr reconstruction of dynamic scenes. ACM Transactions on Graphics (TOG)，31(6)：#203［DOI：10.1145/2366145.2366222］

Sharp G C，Lee S W and Wehe D K. 2002. ICP registration using invariant features. IEEE Transactions on Pattern Analysis and Machine Intelligence，24(1)：90-102［DOI：10.1109/34.982886］

Shen S H. 2013. Accurate multiple view 3D reconstruction using patch-based stereo for large-scale scenes. IEEE Transactions on Image Processing，22(5)：1901-1914［DOI：10.1109/TIP.2013.2237921］

Shi M，Lin S W，Yi Q M，Weng J，Luo A W and Zhou Y C. 2024. Lightweight context-aware network using partial-channel transformation for real-time semantic segmentation. IEEE Transactions on Intelligent Transportation Systems，25(7)：7401-7416［DOI：10.1109/TITS.2023.3348631］

Shivakumar S S，Rodrigues N，Zhou A，Miller I D，Kumar V and Taylor C J. 2020. PST900：RGB-thermal calibration，dataset and segmentation network//Proceedings of 2020 IEEE International Conference on Robotics and Automation (ICRA). Paris，France：IEEE：9441-9447［DOI：10.1109/ICRA40945.2020.9196831］

Shocher A，Cohen N and Irani M. 2018. Zero-shot super-resolution using deep internal learning//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City，USA：IEEE：3118-3126［DOI：10.1109/CVPR.2018.00329］

Shu C Y，Liu Y F，Gao J F，Yan Z and Shen C H. 2021. Channel-wise knowledge distillation for dense prediction//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal，Canada：IEEE：5291-5300［DOI：10.1109/ICCV48922.2021.00526］

Silberman N，Hoiem D，Kohli P and Fergus R. 2012. Indoor Segmentation and Support Inference from RGBD Images//Proceedings of the 12th European Conference on Computer Vision. Florence，Italy：Springer：746-760［DOI：10.1007/978-3-642-33715-4_54］

Simard PY，Steinkraus D and Platt JC. 2003. Best practices for convolutional neural networks applied to visual document analysis//Proceedings of the 7th International Conference on Document Analysis and Recognition. Edinburgh，UK：IEEE：958-963［DOI：10.1109/ICDAR.2003.1227801］

Singh A，Patil D and Omkar S N. 2018. Eye in the sky：real-time drone surveillance system (DSS) for violent individuals identification using scatternet hybrid deep learning network//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Salt Lake City，USA：IEEE：1710-17108［DOI：10.1109/CVPRW.2018.00214］

Song L C，Chen A P，Li Z，Chen Z，Chen L L，Yuan J S，Xu Y and Geiger A. 2023. NeRFPlayer：a streamable dynamic scene representation with decomposed neural radiance fields. IEEE Transactions on Visualization and Computer Graphics，29(5)：2732-2742［DOI：10.1109/TVCG.2023.3247082］

Song S R，Lichtenberg S P and Xiao J X. 2015. SUN RGB-D：a RGB-D scene understanding benchmark suite//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston，USA：IEEE：567-576［DOI：10.1109/CVPR.2015.7298655］

Srivastava A，Badal T，Garg A，Vidyarthi A and Singh R. 2021. Recognizing human violent action using drone surveillance within real-time proximity. Journal of Real-Time Image Processing，18(5)：1851-1863［DOI：10.1007/s11554-021-01171-2］

Srivastava A，Badal T，Saxena P，Vidyarthi A and Singh R. 2022. UAV surveillance for violence detection and individual identification. Automated Software Engineering，29(1)：#28［DOI：10.1007/s10515-022-00323-3］

Srivastava N，Hinton G，Krizhevsky A，Sutskever I and Salakhutdinov R. 2014. Dropout：a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research，15(1)：1929-1958

Sun Y，Zheng W X，Du X and Yan Z P. 2023. Underwater small target detection based on YOLOX combined with MobileViT and double coordinate attention. Journal of Marine Science and Engineering，11(6)：#1178［DOI：10.3390/jmse11061178］

Sun Y M，Cao B，Zhu P F and Hu Q H. 2022. Drone-based RGB-infrared cross-modality vehicle detection via uncertainty-aware learning. IEEE Transactions on Circuits and Systems for Video

Technology, 32(10): 6700-6713 [DOI: 10.1109/TCSVT.2022.3168279]

Sun Y X, Zuo W X and Liu M. 2019. RTFNet: RGB-thermal fusion network for semantic segmentation of urban scenes. IEEE Robotics and Automation Letters, 4(3): 2576-2583 [DOI: 10.1109/LRA.2019.2904733]

Sun Y X, Zuo W X, Yun P, Wang H L and Liu M. 2021. FuseSeg: semantic segmentation of urban scenes based on RGB and thermal data fusion. IEEE Transactions on Automation Science and Engineering, 18(3): 1000-1011 [DOI: 10.1109/TASE.2020.2993143]

Tang L F, Huang H, Zhang Y F and Li F. 2022. Spatial aware channel attention guided high dynamic image reconstruction. Journal of Image and Graphics, 27(12): 3581-3595 (唐凌峰, 黄欢, 张亚飞, 李凡. 2022. 空间感知通道注意力引导的高动态图像重建. 中国图象图形学报, 2022, 27(12): 3581-3595) [DOI: 10.11834/jig.211039]

Tao X, Gao H Y, Shen X Y, Wang J and Jia J Y. 2018. Scale-recurrent network for deep image deblurring//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 8174-8182 [DOI: 10.1109/CVPR.2018.00853]

Tatarchenko M, Dosovitskiy A and Brox T. 2017. Octree generating networks: efficient convolutional architectures for high-resolution 3D outputs//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 2107-2115 [DOI: 10.1109/ICCV.2017.230]

Testolina P, Barbato F, Michieli U, Giordani M, Zanuttigh P and Zorzi M. 2023. SELMA: semantic large-scale multimodal acquisitions in variable weather, daytime and viewpoints. IEEE Transactions on Intelligent Transportation Systems, 24(7): 7012-7024 [DOI: 10.1109/TITS.2023.3257086]

Tikhonov A N. 1963. On the solution of ill-posed problems and the method of regularization. Doklady Akademii Nauk SSSR, 151(3): 501-504

Trinidad M C, Martin-Brualla R, Kainz F and Kontkanen J. 2019. Multi-view image fusion//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE: 4100-4109 [DOI: 10.1109/ICCV.2019.00420]

Tulsiani S, Efros A A and Malik J. 2018. Multi-view consistency as supervisory signal for learning shape and pose prediction//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 2897-2905 [DOI: 10.1109/CVPR.2018.00306]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6000-6010

Wang A, Chen H, Liu L H, Chen K, Lin Z J, Han J G and Ding G G. 2024a. YOLOva10: real-time end-to-end object detection [EB/OL]. [2024-08-05]. https://arxiv.org/pdf/2405.14458.pdf

Wang C C, He W, Nie Y, Guo J Y, Liu C J, Han K and Wang Y H. 2023a. Gold-YOLO: efficient object detector via gather-and-distribute mechanism//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: #2224

Wang C Y, Bochkovskiy A and Liao H Y M. 2023b. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE: 7464-7475 [DOI: 10.1109/CVPR52729.2023.00721]

Wang C Y, Yeh I H and Liao H Y M. 2024b. YOLOv9: learning what you want to learn using programmable gradient information//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 1-21 [DOI: 10.1007/978-3-031-72751-1_1]

Wang F J H, Galliani S, Vogel C, Speciale P and Pollefeys M. 2021a. PatchmatchNet: learned multi-view patchmatch stereo//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 14189-14198 [DOI: 10.1109/CVPR46437.2021.01397]

Wang H, Chen Y Y, Cai Y F, Chen L, Li Y C, Sotelo M A and Li Z X. 2022a. SFNet-N: an improved SFNet algorithm for semantic segmentation of low-light autonomous driving road scenes. IEEE Transactions on Intelligent Transportation Systems, 23(11): 21405-21417 [DOI: 10.1109/TITS.2022.3177615]

Wang H Y, Wang C P, Fu Q, Zhang D D, Kou R K, Yu Y and Song J. 2024c. Cross-modal oriented object detection of UAV aerial images based on image feature. IEEE Transactions on Geoscience and Remote Sensing, 62: #5403021 [DOI: 10.1109/TGRS.2024.3367934]

Wang N Y, Zhang Y D, Li Z W, Fu Y W, Liu W and Jiang Y G. 2018. Pixel2Mesh: generating 3D mesh models from single RGB images//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 55-71 [DOI: 10.1007/978-3-030-01252-6_4]

Wang P S, Liu Y and Guo Y X. 2017. O-CNN: octree-based convolutional neural networks for 3D shape analysis. ACM Transactions on Graphics (TOG), 36(4): #72 [DOI: 10.1145/3072959.3073608]

Wang S Q, Xu X, Ma X Z, Jiang K and Wang Z. 2023c. Informative classes matter: towards unsupervised domain adaptive nighttime semantic segmentation//Proceedings of the 31st ACM International Conference on Multimedia. Ottawa, Canada: ACM: 163-172 [DOI: 10.1145/3581783.3611956]

Wang T X, Jiang Z, Jiang D G, Li B L and Guo C L. 2022. Improved ICP laser point cloud registration algorithm based on PCA. Remote Sensing Information, 37(2): 70-76 (王太学, 江智, 江德港, 李柏林, 郭彩玲. 2022b. 融合PCA的改进ICP激光点云配准算法. 遥感信息, 37(2): 70-76). [DOI: 10.3969/j.issn.1000-3177.2022.02.009]

Wang W J, Yang W H, Fang Y M, Huang H and Liu J Y. 2024. Visual perception and understanding in degraded scenarios. Journal of Image and Graphics, 29(6): 1667-1684 (汪文靖, 杨文瀚, 方玉明, 黄华, 刘家瑛. 2024. 恶劣场景下视觉感知与理解综述. 中国图象图形学报, 29(6): 1667-1684) [DOI: 10.11834/jig.240041]

Wang W Y and Neumann U. 2018. Depth-aware CNN for RGB-D segmentation//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 144-161 [DOI: 10.1007/978-3-030-01252-6_9]

Wang X T, Xie L B, Dong C and Shan Y. 2021b. Real-ESRGAN: training real-world blind super-resolution with pure synthetic data//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision Workshops. Montreal, Canada: IEEE: 1905-1914 [DOI: 10.1109/ICCVW54120.2021.00217]

Wang Y and Solomon J. 2019. Deep closest point: learning representations for point cloud registration//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 3522-3531 [DOI: 10.1109/ICCV.2019.00362]

Wang Y B, Gao R Y, Chen K, Zhou K Q, Cai Y J, Hong L Q, Li Z G, Jiang L H, Yeung D Y, Xu Q and Zhang K. 2024d. DetDiffusion: synergizing generative and perceptive models for enhanced data generation and perception//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 7246-7255 [DOI: 10.1109/CVPR52733.2024.00692]

Wang Y K, Chen X H, Cao L L, Huang W B, Sun F C and Wang Y H. 2022b. Multimodal token fusion for vision transformers//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 12176-12185 [DOI: 10.1109/CVPR52688.2022.01187]

Wang Y K, Sun F C, Huang W B, He F X and Tao D C. 2023d. Channel exchanging networks for multimodal and multitask dense image prediction. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(5): 5481-5496 [DOI: 10.1109/TPAMI.2022.3211086]

Wang Y K, Zhou W, Jiang T, Bai X and Xu Y C. 2020. Intra-class feature variation distillation for semantic segmentation//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 346-362 [DOI: 10.1007/978-3-030-58571-6_21]

Wang Z D, Cun X D, Bao J M, Zhou W G, Liu J Z and Li H Q. 2022c. Uformer: a general u-shaped transformer for image restoration//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 17662-17672 [DOI: 10.1109/CVPR52688.2022.01716]

Wang Z Y, Li Y L, Chen X, Lim S N, Torralba A, Zhao H S and Wang S J. 2023e. Detecting everything in the open world: towards universal object detection//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 11433-11443 [DOI: 10.1109/CVPR52729.2023.01100]

Wang Z Y, Li Y L, Chen X, Lim S N, Torralba A, Zhao H S and Wang S J. 2024e. UniDetector: towards universal object detection with heterogeneous supervision. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1-18 [DOI: 10.1109/TPAMI.2024.3411595]

Wei P X, Xie Z W, Lu H N, Zhan Z Y, Ye Q X, Zuo W M and Lin L. 2020. Component divide-and-conquer for real-world image super-resolution//Proceedings of the 16th European Conference on Computer Vision (ECCV). Glasgow, UK: Springer: 101-117 [DOI: 10.1007/978-3-030-58598-3_7]

Wittner C, Schauerte B and Stiefelhagen R. 2015. What's the point? Frame-wise pointing gesture recognition with latent-dynamic conditional random fields [EB/OL]. [2024-08-05]. https://arxiv.org/pdf/1510.05879.pdf

Wu P, Liu J, Shi Y J, Sun Y J, Shao F T, Wu Z Y and Yang Z W. 2020a. Not only look, but also listen: learning multimodal violence detection under weak supervision//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 322-339 [DOI: 10.1007/978-3-030-58577-8_20]

Wu P, Liu X T and Liu J. 2023a. Weakly supervised audio-visual violence detection. IEEE Transactions on Multimedia, 25: 1674-1685 [DOI: 10.1109/TMM.2022.3147369]

Wu R, Nie J H, G H, Liu Y and Lu H T. 2022. SingleMatch: a point cloud coarse registration method with single match point and deep-learning describer. Multimedia Tools and Applications, 81(12): 16967-16986 [DOI: 10.1007/s11042-022-12704-7]

Wu S Z, Xu J R, Tai Y W and Tang C K. 2018. Deep high dynamic range imaging with large foreground motions//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 120-135 [DOI: 10.1007/978-3-030-01216-8_8]

Wu Y, Yuan Y Z, Xiang B H, Sheng J L, Lei J Y, Hu C Y, Gong M G, Ma W P and Miao Q G. 2023b. Overview of the computational intelligence method in 3D point cloud registration. Journal of Image and Graphics, 28(9): 2763-2787 (武越, 苑咏哲, 向本华, 绳金龙, 雷佳熠, 胡聪颖, 公茂果, 马文萍, 苗启广. 2023. 三维点云配准中的计算智能方法综述. 中国图象图形学报, 28(9): 2763-2787) [DOI: 10.11834/jig.220727]

Wu Y X, Zhang G W, Gao Y M, Deng X J, Gong K, Liang X D and Lin L. 2020b. Bidirectional graph reasoning network for panoptic segmentation//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 9077-9086 [DOI: 10.1109/CVPR42600.2020.00910]

Wu Z R, Song S R and Khosla A. 2015. 3D ShapeNets: a deep representation for volumetric shapes//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 1912-1920 [DOI: 10.1109/CVPR.2015.7298801]

Xia R H, Zhao C Q, Zheng M, Wu Z Y, Sun Q Y and Tang Y. 2023. CMDA: cross-modality domain adaptation for nighttime semantic

segmentation//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 21515-21524 [DOI: 10.1109/ICCV51070.2023.01972]

Xie H Z, Yao H X, Sun X S, Zhou S C and Zhang S P. 2019. Pix2Vox: context-aware 3D reconstruction from single and multi-view images//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE: 2690-2698 [DOI: 10.1109/ICCV.2019.00278]

Xie Z F, Wang S, Xu K, Zhang Z Z, Tan X, Xie Y and Ma L Z. 2023. Boosting night-time scene parsing with learnable frequency. IEEE Transactions on Image Processing, 32: 2386-2398 [DOI: 10.1109/TIP.2023.3267044]

Xiong Y W, Liao R J, Zhao H S, Hu R, Bai M, Yumer E and Urtasun R. 2019. UPSNet: a unified panoptic segmentation network//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 8810-8818 [DOI: 10.1109/CVPR.2019.00902]

Xu J C, Xiong Z X and Bhattacharyya S P. 2023a. PIDNet: a real-time semantic segmentation network inspired by PID controllers//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 19529-19539 [DOI: 10.1109/CVPR52729.2023.01871]

Xu K, Liu X P, Wang H Y, Wan J W and Guo Y L. 2024. Infrared-visible image object detection algorithm using feature dynamic selection. Journal of Image and Graphics, 29(8): 2350-2363 (许可, 刘心溥, 汪汉云, 万建伟, 郭裕兰. 2024. 红外与可见光图像特征动态选择的目标检测网络. 中国图象图形学报, 29(8): 2350-2363) [DOI: 10.11834/jig.230495]

Xu Q S, Kong W H, Tao W B and Pollefeys M. 2023b. Multi-scale geometric consistency guided and planar prior assisted multi-view stereo. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(4): 4945-4963 [DOI: 10.1109/TPAMI.2022.3200074]

Xu Q S and Tao W B. 2019. Multi-scale geometric consistency guided multi-view stereo//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 5478-5487 [DOI: 10.1109/CVPR.2019.00563]

Xu Q S and Tao W B. 2020. Planar prior assisted PatchMatch multi-view stereo//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA: AAAI: 12516-12523 [DOI: 10.1609/aaai.v34i07.6940]

Xu S L, Wang X X, Lv W Y, Chang Q Y, Cui C, Deng K P, Wang G Z, Dang Q Q, Wei S Y, Du Y N and Lai B H. 2022. PP-YOLOE: an evolved version of YOLO [EB/OL]. [2024-08-05]. https://arxiv.org/pdf/2203.16250.pdf

Xu Y F, Zhang M D, Fu C Y, Chen P X, Yang X S, Li K and Xu C S. 2023c. Multi-modal queried object detection in the wild//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: #198

Xu Z Y, Liu Y G, Shi X L, Wang Y and Zheng Y N. 2020. MARMVS: matching ambiguity reduced multiple view stereo for efficient large scale scene reconstruction//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 5980-5989 [DOI: 10.1109/CVPR42600.2020.00602]

Yan Q S, Gong D, Shi Q F, van den Hengel A, Shen C H, Reid I and Zhang Y N. 2019. Attention-guided network for ghost-free high dynamic range imaging//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 1751-1760 [DOI: 10.1109/CVPR.2019.00185]

Yan Q S, Zhang L, Liu Y, Zhu Y, Sun J Q, Shi Q F and Zhang Y N. 2020. Deep HDR imaging via a non-local network. IEEE Transactions on Image Processing, 29: 4308-4322 [DOI: 10.1109/TIP.2020.2971346]

Yang C C, Lin Z J, Lan Z Y, Chen R Q, Wei L F and Liu Y Z. 2024a. Evolutionary channel pruning for real-time object detection. Knowledge-Based Systems, 287: #111432 [DOI: 10.1016/j.knosys.2024.111432]

Yang C G, Zhou H L, An Z L, Jiang X, Xu Y J and Zhang Q. 2022a. Cross-image relational knowledge distillation for semantic segmentation//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 12309-12318 [DOI: 10.1109/CVPR52688.2022.01200]

Yang J L, Li H D, Campbell D and Jia Y D. 2016. Go-ICP: a globally optimal solution to 3D ICP point-set registration. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(11): 2241-2254 [DOI: 10.1109/TPAMI.2015.2513405]

Yang J Q, Chen J H, Quan S W, Wang W and Zhang Y N. 2022b. Correspondence selection with loose-tight geometric voting for 3-D point cloud registration. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-14 [DOI: 10.1109/TGRS.2022.3142074]

Yang J Q, Huang Z Q, Quan S W, Qi Z S and Zhang Y N. 2021. Sac-cot: sample consensus by sampling compati-bility triangles in graphs for 3-d point cloud registration. IEEE Transactions on Geoscience and Remote Sensing, 2021, 60: 1-15 [DOI: 10.1109/TGRS.2021.3058552]

Yang J Y, Mao W, Alvarez J M and Liu M M. 2020. Cost volume pyramid based depth inference for multi-view stereo//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 4876-4885 [DOI: 10.1109/CVPR42600.2020.00493]

Yang M J, Zhong Y Z, Guo B and Tian S Y. 2022. Point cloud registration based on improved grey wolf optimization algorithm. Computer Simulation, 39(12): 513-518 (杨沐杰, 钟羽中, 郭斌, 佃松宜. 2022c. 基于改进灰狼优化算法的点云配准. 计算机仿真, 39(12): 513-518) [DOI:10.3969/j.issn.1006-9348.2022.12.095]

Yang Y, Pan L Y, Liu L and Liu M M. 2023. K3DN: disparity-aware kernel estimation for dual-pixel defocus deblurring//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Rec-

ognition. Vancouver, Canada: IEEE: 13263-13272 [DOI: 10.1109/CVPR52729.2023.01274]

Yang Z Y, Gao X Y, Zhou W, Jiao S H, Zhang Y Q and Jin X G. 2024b. Deformable 3D Gaussians for high-fidelity monocular dynamic scene reconstruction//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 20331-20341 [DOI: 10.1109/CVPR52733.2024.01922]

Yao Y, Luo Z X, Li S W, Fang T and Quan L. 2018. MVSNet: depth inference for unstructured multi-view stereo//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 785-801 [DOI: 10.1007/978-3-030-01237-3_47]

Yao Y, Luo Z X, Li S W, Shen T W, Fang T and Quan L. 2019. Recurrent MVSNet for high-resolution multi-view stereo depth inference//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 5520-5529 [DOI: 10.1109/CVPR.2019.00567]

Yao Z Y, Ai J B, Li B X and Zhang C. 2021. Efficient DETR: improving end-to-end object detector with dense prior [EB/OL]. [2024-08-05]. https://arxiv.org/pdf/2104.01318.pdf

Ye T, Qin W, Zhao Z Y, Gao X Z, Deng X P and Ouyang Y. 2023. Real-time object detection network in UAV-vision based on CNN and transformer. IEEE Transactions on Instrumentation and Measurement, 72: #2505713 [DOI: 10.1109/TIM.2023.3241825]

Ye X R, Li Z P and Xu C. 2022. Ghost-free multi-exposure image fusion technology based on the multi-scale block LBP operator. Electronics, 11(19): #3129 [DOI: 10.3390/electronics11193129]

Yu C Q, Gao C X, Wang J B, Yu G, Shen C H and Sang N. 2021. BiSeNet V2: bilateral network with guided aggregation for real-time semantic segmentation. International Journal of Computer Vision, 129(11): 3051-3068 [DOI: 10.1007/s11263-021-01515-2]

Yu C Q, Wang J B, Peng C, Gao C X, Yu G and Sang N. 2018. BiSeNet: bilateral segmentation network for real-time semantic segmentation//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 334-349 [DOI: 10.1007/978-3-030-01261-8_20]

Yu Q H, Wang H Y, Qiao S Y, Collins M, Zhu Y K, Adam H, Yuille A and Chen L. 2022. *k*-means mask Transformer//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 288-307 [DOI: 10.1007/978-3-031-19818-2_17]

Yuan J Z, Zhou W J and Luo T. 2019. DMFNet: deep multi-modal fusion network for RGB-D indoor scene segmentation. IEEE Access, 7: 169350-169358 [DOI: 10.1109/ACCESS.2019.2955101]

Yuan M X, Shi X R, Wang N, Wang Y Y and Wei X X. 2024. Improving RGB-infrared object detection with cascade alignment-guided transformer. Information Fusion, 105: #102246 [DOI: 10.1016/j.inffus.2024.102246]

Yuan M X and Wei X X. 2024. C^{2}Former: calibrated and complementary transformer for RGB-infrared object detection. IEEE Transactions on Geoscience and Remote Sensing, 62: #5403712 [DOI: 10.1109/TGRS.2024.3376819]

Yun S D, Han D, Chun S, Oh S J, Yoo Y and Choe J. 2019. CutMix: regularization strategy to train strong classifiers with localizable features//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 6022-6031 [DOI: 10.1109/ICCV.2019.00612]

Zamir S W, Arora A, Khan S, Hayat M, Khan F S and Yang M H. 2022. Restormer: efficient transformer for high-resolution image restoration//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 5718-5729 [DOI: 10.1109/CVPR52688.2022.00564]

Zeng A, Song S R, Niessner M, Fisher M, Xiao J X and Funkhouser T. 2017. 3DMatch: learning local geometric descriptors from RGB-D reconstructions//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 199-208 [DOI: 10.1109/CVPR.2017.29]

Zhang H, Li F, Liu S L, Zhang L, Su H, Zhu J, Ni L M and Shum H Y. 2023a. DINO: DETR with improved DeNoising anchor boxes for end-to-end object detection//Proceedings of the 11th International Conference on Learning Representations. Kigali, Rwanda: ICLR

Zhang H Y, Cissé M, Dauphin Y N and Lopez-Paz D. 2018a. mixup: beyond empirical risk minimization//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada: ICLR

Zhang J M, Liu H Y, Yang K L, Hu X X, Liu R P and Stiefelhagen R. 2023b. CMX: cross-modal fusion for RGB-X semantic segmentation with transformers. IEEE Transactions on Intelligent Transportation Systems, 24(12): 14679-14694 [DOI: 10.1109/TITS.2023.3300537]

Zhang K, Liang J Y, Van Gool L and Timofte R. 2021a. Designing a practical degradation model for deep blind image super-resolution//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 4771-4780 [DOI: 10.1109/ICCV48922.2021.00475]

Zhang K, Zuo W M, Chen Y J, Meng D Y and Zhang L. 2017. Beyond a Gaussian denoiser: residual learning of deep CNN for image denoising. IEEE Transactions on Image Processing, 26(7): 3142-3155 [DOI: 10.1109/TIP.2017.2662206]

Zhang L, Wu X L, Buades A and Li X. 2011. Color demosaicking by local directional interpolation and nonlocal adaptive thresholding. Journal of Electronic Imaging, 20(2): #023016 [DOI: 10.1117/1.3600632]

Zhang N, Nex F, Kerle N and Vosselman G. 2021b. Towards learning low-light indoor semantic segmentation with illumination-invariant features. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLIII-B2-2021:

427-432 [DOI: 10.5194/isprs-archives-XLIII-B2-2021-427-2021]

Zhang P Y, Zhong Y X and Li X Q. 2019. SlimYOLOv3: narrower, faster and better for real-time UAV applications//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). Seoul, Korea (South): IEEE: 37-45 [DOI: 10.1109/ICCVW.2019.00011]

Zhang Q, Zhao S L, Luo Y J, Zhang D W, Huang N C and Han J G. 2021c. ABMDRNet: adaptive-weighted bi-directional modality difference reduction network for RGB-T semantic segmentation//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 2633-2642 [DOI: 10.1109/CVPR46437.2021.00266]

Zhang R, Isola P and Efros A A. 2016. Colorful image colorization//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 649-666 [DOI: 10.1007/978-3-319-46487-9_40]

Zhang S L, Xie T, Wang Y Q, Liu Y Y, Dou Y L and Wang S K. 2023c. SF-YOLO: RGB-T fusion object detection in UAV scenes//Proceedings of the 8th International Conference on Image, Vision and Computing (ICIVC). Dalian, China: IEEE: 51-59 [DOI: 10.1109/ICIVC58118.2023.10270358]

Zhang W Q, Huang Z L, Luo G Z, Chen T, Wang X G, Liu W Y, Yu G and Shen C H. 2022. TopFormer: token pyramid transformer for mobile semantic segmentation//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 12073-12083 [DOI: 10.1109/CVPR52688.2022.01177]

Zhang X Y, Yang J Q, Zhang S K and Zhang Y N. 2023d. 3D registration with maximal cliques.//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 17745-17754 [DOI: 10.1109/CVPR52729.2023.01702]

Zhang Y C, Song Z Y and Li W B. 2021d. Unsupervised data augmentation for object detection [EB/OL]. [2024-08-05]. https://arxiv.org/pdf/2104.14965.pdf

Zhang Y L, Li K P, Li K, Wang L C, Zhong B N and Fu Y. 2018b. Image super-resolution using very deep residual channel attention networks//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 286-301 [DOI: 10.1007/978-3-030-01234-2_18]

Zhao H S, Shi J P, Qi X J, Wang X G and Jia J Y. 2017. Pyramid scene parsing network//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 6230-6239 [DOI: 10.1109/CVPR.2017.660]

Zhao H S, Zhang Y, Liu S, Shi J P, Loy C C, Lin D H and Jia J Y. 2018a. PSANet: point-wise spatial attention network for scene parsing//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 270-286 [DOI: 10.1007/978-3-030-01240-3_17]

Zhao T Y, Ren W Q, Zhang C Q, Ren D W and Hu Q H. 2018b. Unsupervised degradation learning for single image super-resolution [EB/OL]. [2024-08-05]. https://arxiv.org/pdf/1812.04240.pdf

Zhao T Y, Yuan M X and Wei X X. 2024a. Removal and selection: improving RGB-infrared object detection via coarse-to-fine fusion [EB/OL]. [2024-08-05]. https://arxiv.org/pdf/2401.10731.pdf

Zhao X F, Xia Y T, Zhang W W, Zheng C and Zhang Z L. 2023. YOLO-ViT-based method for unmanned aerial vehicle infrared vehicle target detection. Remote Sensing, 15(15): #3778 [DOI: 10.3390/rs15153778]

Zhao Y A, Lv W Y, Xu S L, Wei J M, Wang G Z, Dang Q Q, Liu Y and Chen J. 2024b. DETRs beat YOLOs on real-time object detection//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 16965-16974 [DOI: 10.1109/CVPR52733.2024.01605]

Zhong L, Cho S, Metaxas D, Paris S and Wang J. 2013. Handling noise in single image deblurring using directional filters//Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Portland, USA: IEEE: 612-619 [DOI: 10.1109/CVPR.2013.85]

Zhou B L, Zhao H, Puig X, Fidler S, Barriuso A and Torralba A. 2017. Scene parsing through ADE20K dataset//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 5122-5130 [DOI: 10.1109/CVPR.2017.544]

Zhou H, Qi L, Wan Z L, Huang H and Yang X. 2021a. RGB-D co-attention network for semantic segmentation//Proceedings of the 15th Asian Conference on Computer Vision. Kyoto, Japan: Springer: 519-536 [DOI: 10.1007/978-3-030-69525-5_31]

Zhou Q, Shi H M, Xiang W K, Kang B and Latecki L J. 2024. DPNet: dual-path network for real-time object detection with lightweight attention. IEEE Transactions on Neural Networks and Learning Systems, 1-15 [DOI: 10.1109/TNNLS.2024.3376563]

Zhou Q Y, Park J and K V. 2016. Fast global registration//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 766-782 [DOI: 10.1007/978-3-319-46475-6_47]

Zhou W J, Lin X Y, Lei J S, Yu L and Hwang J N. 2022. MFFENet: multiscale feature fusion and enhancement network for RGB — thermal urban road scene parsing. IEEE Transactions on Multimedia, 24: 2526-2538 [DOI: 10.1109/TMM.2021.3086618]

Zhou W J, Liu J F, Lei J S, Yu L and Hwang J N. 2021b. GMNet: graded-feature multilabel-learning network for RGB — thermal urban scene semantic segmentation. IEEE Transactions on Image Processing, 30: 7790-7802 [DOI: 10.1109/TIP.2021.3109518]

Zhou W J, Yuan J Z, Lei J S and Luo T. 2021c. TSNet: three-stream self-attention network for RGB-D indoor semantic segmentation. IEEE Intelligent Systems, 36(4): 73-78 [DOI: 10.1109/MIS.2020.2999462]

Zhou W J, Yue Y C, Fang M X, Qian X H, Yang R W and Yu L. 2023.

BCINet: bilateral cross-modal interaction network for indoor scene understanding in RGB-D images. Information Fusion, 94: 32-42 [DOI: 10.1016/j.inffus.2023.01.016]

Zhu X Z, Su W J, Lu L W, Li B, Wang X G and Dai J F. 2021. Deformable DETR: deformable transformers for end-to-end object detection//Proceedings of the 9th International Conference on Learning Representations. [s.l.]: ICLR

Zuo W M, Ren D W, Gu S H, Lin L and Zhang L. 2015. Discriminative learning of iteration-wise priors for blind deconvolution//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 3232-3240 [DOI: 10.1109/CVPR.2015.7298943]

作者简介

张艳宁，女，教授，主要研究方向为计算机视觉。
E-mail：ynzhang@nwpu.edu.cn

张磊，通信作者，男，教授，主要研究方向为计算机视觉。
E-mail：nwpuzhanglei@nwpu.edu.cn

王昊宇，男，博士研究生，主要研究方向为计算机视觉。
E-mail：wanghaoyunwpu@mail.nwpu.edu.cn

闫庆森，男，教授，主要研究方向为图像重建、图像融合和持续学习。E-mail：qingsenyan@nwpu.edu.cn

杨佳琪，男，副教授，主要研究方向为三维视觉。
E-mail：jqyang@nwpu.edu.cn

刘婷，女，副教授，主要研究方向为图像语义分割。
E-mail：liuting@nwpu.edu.cn

符梦芹，女，硕士研究生，主要研究方向为视觉目标检测识别。E-mail：fumengqin2001@163.com

吴鹏，男，副教授，主要研究方向为计算机视觉和异常行为检测。E-mail：xdwupeng@gmail.com