中图法分类号:TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(2025)10-3319-16

论文引用格式:Shen Y X, Guo X, Wang H, Hu K Y and Tao C. 2025. Integrated generation-detection approach for pedestrian recognition in dark-light conditions. Journal of Image and Graphics, 30(10):3319-3334(沈羽翔, 郭鑫, 王昊, 胡柯彦, 陶超. 2025. 面向暗光条件下行人识别的生成检测一体化方法. 中国图象图形学报, 30(10):3319-3334)[DOI:10.11834/jig.240649]

面向暗光条件下行人识别的生成检测一体化方法

沈羽翔,郭鑫,王昊,胡柯彦,陶超*

中南大学地球科学与信息物理学院,长沙　410083

摘　要:**目的**　暗光环境下的行人检测是计算机视觉领域的一项重大挑战,其难点在于光照不充分会导致行人特征模糊不可辨识。传统方法一般通过图像增强或生成红外图像提供补充信息解决这一问题。但由于其增强过程与下游检测任务存在分离,限制其应用性能。针对这一问题,提出一种面向暗光条件下行人识别的生成检测一体化方法,旨在解决传统范式中暗光图像数据增强和下游检测任务间的割裂问题。**方法**　提出一种端到端生成检测一体化架构,利用条件扩散模型从暗光图像生成辅助红外模态图像,并通过生成的多尺度特征提升目标检测性能。此外,为了避免梯度无法经过VAE(variational autoencoder)解码器传递的问题,进一步提出生成检测端到端联合优化策略。**结果**　在LLVIP(low-light visible-infrared paired dataset)和VTMOT(visible-thermal multiple object tracking dataset)数据集上的实验结果表明,本文方法在总体精度指标上显著优于传统方法和其他先进方法。在LLVIP数据集上,本文方法的F1值为85.75%,平均精度均值(mAP)为75.35%,优于传统方法中结果最佳的SCI(self-calibrated illumination)方法(F1值82.05%,mAP值72.30%)和其他先进方法中结果最佳的Faster_RCNN_hrnet(F1值82.91%,mAP值70.02%)。在VTMOT数据集上,本文方法同样表现优异,F1值为90.01%,mAP为73.44%,优于SCI方法(F1值85.91%,mAP值72.19%)和Faster_RCNN_hrnet(F1值84.48%,mAP值71.16%)。此外,消融实验验证了生成模块和联合优化策略在整体框架中的有效性。**结论**　本文证明了生成检测一体化框架在复杂低光环境下的优越性,有效解决了生成过程与检测任务割裂的问题。未来研究将进一步优化生成效率,并扩展该方法至更多多模态应用领域。

关键词:暗光行人识别;生成检测一体化;条件扩散模型;目标检测;红外图像生成;端到端联合优化;多模态融合

Integrated generation-detection approach for pedestrian recognition in dark-light conditions

Shen Yuxiang, Guo Xin, Wang Hao, Hu Keyan, Tao Chao*

School of Geosciences and Info-physics, Central South University, Changsha　410083, China

Abstract: **Objective** Pedestrian detection in dark-light environments is a key challenge in the field of computer vision due to the lack of sufficient lighting, which results in blurry and low-contrast features of pedestrians. Traditional approaches to address this issue often focus on enhancing image quality through domain-specific visible light enhancement or generating

收稿日期:2024-11-07;**修回日期**:2025-02-13;**预印本日期**:2025-02-20

*通信作者:陶超　kingtaochao@csu.edu.cn

基金项目:国家自然科学基金项目(42171376,42471419);湖南省杰出青年科学基金项目(2022JJ10072);内蒙古大学青年学术人才科研启动项目(10000-A25206020)

Supported by: National Natural Science Foundation of China (42171376, 42471419); Natural Science Foundation of Hunan for Distinguished Young Scholars (2022JJ10072); High-level Talents Introduction Project of Inner Mongolia University (10000-A25206020)

infrared images to provide complementary information. However, these approaches are typically limited by the separation between the data enhancement process and the downstream detection tasks, which results in suboptimal performance. To overcome these limitations, this study proposes an integrated generation and detection framework tailored for pedestrian detection in dark-light conditions. The primary goal of the proposed framework is to bridge the gap between data enhancement and pedestrian detection by generating high-quality infrared images from dark-light images and jointly optimizing the generation and detection tasks in a unified architecture. **Method** The proposed framework is built upon a conditional diffusion model, which is used to generate auxiliary infrared modality images from dark-light RGB images. Unlike traditional approaches that first enhance or generate images and then input them into a detection network separately, this framework integrates the image generation and pedestrian detection tasks into a single end-to-end architecture. Inspired by ControlNet, the model leverages a conditional diffusion process to produce infrared images that provide enhanced pedestrian features. These infrared images are then used in conjunction with the original dark-light images to improve the detection performance. In addition, the framework extracts multi-scale features from the generated images through a UNet architecture, which replaces the traditional backbone network of the detection model. These multi-scale features are directly fed into the detection head, which allows the detection task to be optimized simultaneously with the image generation process. To ensure effective joint optimization, the framework incorporates a novel strategy to bypass the typical gradient flow limitations encountered when using variational autoencoders (VAEs). In standard designs, the VAE decoder is often a fixed module that prevents gradients from being propagated back during the training of the detection task. To address this limitation, the proposed framework allows the gradient to flow through the multi-scale feature maps generated by the UNet, which ensures that the generation and detection tasks benefit from the same optimization process. The detection task is supervised by a loss function that combines boundary box regression, classification, and confidence losses. This integration enables the model to learn object localization and classification. **Result** Extensive experiments were conducted on the LLVIP and VTMOT datasets, which together contain 26 193 paired visible-light and infrared images (a total 52 386 images), all annotated with pedestrian detection labels. These datasets present significant challenges, particularly due to images captured under extreme low-light conditions. The proposed method was evaluated against several state-of-the-art techniques, including single-domain visible light enhancement methods and cross-domain infrared image generation approaches. Experimental results on the LLVIP and VTMOT datasets show that the method proposed in this study significantly outperforms traditional and state-of-the-art methods in terms of overall accuracy metrics. For example, on the LLVIP dataset, the F1 value of the proposed method is 85.75%, and the mAP is 75.35%. It outperforms the SCI method, which is the best of the traditional methods (an F1 value of 82.05% and a mean average precision(mAP) value of 72.30%), and Faster_RCNN_hrnet, which is the best of the state-of-the-art methods (an F1 value of 82.91% and a mAP value of 70.02%). On the VTMOT dataset, the proposed method also performs effectively with an F1 value of 90.01% and a mAP of 73.44%. It is superior to the best SCI method among the traditional methods (an F1 value of 85.91% and a mAP value of 72.19%) and the best Faster_RCNN_hrnet among the advanced methods (an F1 value of 84.48% and a mAP value of 71.16%). Moreover, the proposed method outperforms other methods overall in terms of precision, recall, and F1 score, which highlights its robustness and reliability in detecting pedestrians in dark-light conditions. In addition to the quantitative improvements, qualitative analysis reveals that the generated infrared images, when fused with the original dark-light images, provide clearer and more distinguishable pedestrian features. This fusion process helps reduce false positives and enhances the accuracy of pedestrian detection. The proposed framework could successfully detect pedestrians that were missed by other methods, especially in challenging scenes with low contrast between the pedestrians and their backgrounds. These results demonstrate the advantages of the joint infrared generation and pedestrian detection approach in overcoming the limitations of traditional methods, particularly in extreme lighting conditions. Ablation studies were conducted to evaluate the contribution of different components within the proposed framework. The results showed that the infrared image generation module and the joint optimization strategy are crucial for improving detection performance. Removing the infrared generation module led to a notable performance drop, as the model struggled to detect pedestrians in low-light conditions using only visible-light images. Similarly, excluding the joint optimization strategy caused a decrease in accuracy due to the lack of gradient propagation between the image generation and detection tasks. **Conclusion** This study presents a novel integrated frame-

work for pedestrian detection in dark-light environments, which combines the generation of infrared modality images with the detection task in a unified architecture. By leveraging a conditional diffusion model to generate auxiliary infrared images and extracting multi-scale features to assist pedestrian detection, the framework successfully addresses the limitations of traditional methods that separate the data enhancement and detection tasks. The proposed end-to-end architecture not only simplifies the detection pipeline but also ensures optimization of the generated images to improve detection accuracy. The experimental results confirm that the proposed framework significantly improves pedestrian detection performance under dark-light conditions compared with single-domain visible light enhancement methods, cross-domain infrared generation methods, and current generalized pedestrian detection methods. The framework achieves higher precision, recall, F1 scores, and mAP while reducing the number of false detections and missed pedestrians. Furthermore, the ablation studies validate the importance of the infrared generation module and the joint optimization strategy in achieving superior detection performance. In summary, the integrated generation and detection framework proposed in this study offers a powerful solution for pedestrian detection in dark-light conditions, which provides qualitative and quantitative improvements. Future research may focus on addressing the computational complexity and inference time associated with diffusion models by exploring accelerated sampling techniques and lightweight architectures to enhance the real-time capabilities of the framework. The flexibility of the proposed architecture allows for easy adaptation to various detection models, which makes it a versatile tool for future developments in multimodal pedestrian detection.

Key words: dark light pedestrian detection; integrated generation and detection; conditional diffusion model; object detection; infrared image generation; end-to-end joint optimization; multimodal fusion

0 引　言

在暗光条件下，行人检测面临诸多挑战（Hwang等，2015）。由于光线不足，图像中的行人特征受噪点影响较为模糊，同时背景与目标之间对比度较低，这使得现有的视觉检测算法难以有效识别和定位行人（Li等，2018；Chen等，2018）。因此，如何提升暗光条件下的行人检测准确性成为计算机视觉领域一个重要研究课题（Kruthiventi等，2017；汪文靖等，2024）。

为了应对这一挑战，研究人员主要提出3种策略提高暗光条件下的行人检测性能：同域可见光的图像增强（Wei等，2018；张航和颜佳，2024）、多模态融合方法和跨域红外模态的图像生成（Güzel和Yavuz，2022）。这些策略都在一定程度上增强了目标特征与背景噪声之间的对比度，推动了暗光条件行人检测任务的进展。

在同域可见光的图像增强策略中，直方图均衡化和Retinex（Rahman等，1996）方法得到广泛应用以改善图像质量。通过调整图像的亮度分布，这些方法能够提高图像的可见度，但往往引入伪影和噪声（Pizer等，1987），从而影响检测算法的性能。为了解决这些问题，深度学习方法逐渐被引入（Chen等，2018）。同域图像增强策略中的自校准照明（self-calibrated illumination，SCI）方法（Ma等，2022），通过特定的损失函数和网络设计，能够更好地保留原始图像的色彩和细节，使增强后的图像更加自然，符合人眼感知，因此在目标检测任务中表现出显著的提升。但是该方法主要关注图像的视觉质量提升，而未针对目标检测任务进行特定的优化，没有考虑到增强后的图像是否有利于检测算法识别目标（Guo等，2021）。

多模态融合方法是指结合来自多个传感器或模态（如可见光、红外、雷达等）的数据，以提高行人检测的准确性和鲁棒性（Cao等，2023）。这种方法利用了不同模态数据的优势，弥补单一模态的局限性。例如，可见光图像提供了丰富的纹理和颜色信息，而红外图像则能在低光照条件下提供稳定的热辐射信息。多模态融合可以通过早期融合（early fusion）、晚期融合（late fusion）或特征级融合（feature-level fusion）等方式实现。研究表明，多模态融合方法能够在复杂环境下显著提升行人的检测效果。特别地，在暗光条件下，红外模态提供的额外信息可以显著增强行人特征，有助于提高检测精度（Wang等，2023）。但红外模态的获取以及与可见光模态的配准都为该方法的应用带来了挑战。

跨域红外模态的图像生成策略充分利用了红外

模态在暗光条件下能够突出高对比度行人特征的独特优势(Wu等,2017;Ye等,2022),通过生成辅助红外数据为暗光条件下的行人识别提供了创新性的解决方案。在跨域生成过程中,生成对抗网络(generative adversarial network,GAN)广泛应用于图像生成任务(Zhu等,2017;陈佛计 等,2021)。GAN通过引入生成器和判别器的对抗训练机制,能够有效学习数据的分布,从而生成更为真实和自然的图像(Goodfellow等,2014;余佩伦 等,2021)。针对同域图像增强策略的不足,一些研究者提出可见光到红外的图像转换方法(Isola 等,2017;Ma 等,2020;Fang,2023)。这些方法不仅能够提升图像的亮度和对比度,同时通过生成高质量的可见光图像,克服了传统方法中伪影和噪声引入的问题。此外,扩散模型作为一种新兴的生成模型,在图像生成领域展现出了强大的能力(Rombach 等,2022;闫志浩 等,2024)。与GAN不同,扩散模型通过逐步添加和去除噪声的过程,逐步生成高质量图像(Ho等,2020)。这种生成过程不仅能够保持图像的细节和纹理,还能更有效地捕捉原始图像的结构信息(Song等,2021)。通过对红外图像进行扩散建模,能够生成具有高视觉质量的可见光图像,有效降低了图像增强过程中噪声的影响。

虽然生成的辅助数据在一定程度上提升了暗光条件下图像的质量,但生成过程与检测模型训练的割裂限制了高质量生成图像对检测器学习效果的显著提升。最新的扩散模型研究表明,扩散模型中U-Net提取的多尺度特征可以很好地适配多种视觉下游任务(Zhao等,2023),在语义分割、参考图像分割和深度估计等任务上取得了最先进的成果(state-of-the-art,SOTA),展示了预训练扩散模型在下游任务中的快速适应性和高效性。因此,本文将预训练扩散模型强大的图像生成能力和特征提取能力分别应用于辅助红外模态数据的生成和下游目标检测任务,为改善暗光条件下的行人检测提供新的思路和方法。在目标检测任务中,需要对物体的边界位置进行精确回归(Redmon和Farhadi,2018;赵永强 等,2020)。然而,如果首先生成辅助红外图像并将其输入检测头,由于在扩散模型的生成过程中,VAE(variational autoencoder)解码器通常视为固定且不可训练的模块,这导致在解码器处计算图被截断,梯度无法通过解码器进行反向传播。因此,这种设计仍然导致生成过程与检测任务之间的割裂,影响整体的协同优化效果。为了解决这一问题,本文提出一种新颖的端到端架构,利用条件扩散模型从暗光图像生成辅助红外模态的对应图像,并利用其多尺度特征进行目标检测。在整个流程中,通过目标检测任务的损失函数对生成过程进行约束,从而实现生成与检测的联动优化。

本文的主要贡献包括以下两个方面:1)提出一种生成任务和目标检测一体化的框架,这能够在进行行人检测的同时,利用生成模型生成的图像对原始图像的检测任务进行辅助。具体而言,该框架将生成模型和检测模型结合,通过生成模型为原始图像生成高质量的辅助图像,提升行人特征的清晰度和对比度,进而提高检测算法的精度。相比于传统的单独进行图像增强和检测的策略,这种一体化方法不仅简化了处理流程,还可以针对性地完成辅助图像生成,最终提升了检测性能。2)提出一种生成检测端到端联合优化策略,采用联合训练的方式,将扩散模型生成红外图像的任务与行人检测任务结合在一个流水线中,避免了梯度无法经过VAE解码器传递的问题。具体而言,在联合训练过程中,扩散模型和目标检测模型共享梯度信息。扩散模型通过学习生成具有更高对比度和清晰度的辅助红外图像,增强了行人特征的显著性,同时允许梯度从检测模型反向传播到生成模型,解决了梯度无法传递的问题。目标检测模型则利用这些增强的特征进行行人检测,提高对行人特征的敏感度和检测的准确性。这种联合训练方法避免了独立进行数据增强和目标检测时的割裂问题,提高整体检测性能。

1 方 法

本文介绍了一种用于暗光条件下的行人目标检测增强方法。需要强调的是,提出的方法并非追求对人类视觉直观的图像增强,而是旨在突出图像中目标行人的特征,充分区分其与背景,从而有助于提升下游视觉任务,即目标检测的性能。

1.1 生成检测一体化框架

如图1所示,本文提出一种生成检测一体化的框架,主要包括两个阶段:辅助红外生成阶段和下游目标检测阶段。将这两个阶段结合到一个框架内,进行端到端的联合优化。具体而言,在辅助红外生

成阶段，受到 ControlNet(Zhang 等，2023)的启发，利用条件扩散模型跨模态生成有助于检测任务的辅助红外图像。在检测阶段，借鉴了 VPD(visual perception with a pre-trained diffusion model)(Zhao 等，2023)的思想，充分利用在生成辅助红外图像过程中产生的多尺度特征图，以此替代传统检测模型骨干网络提取的特征，直接输入到检测头中进行目标检测的学习。

将生成与检测过程一体化是为了克服传统生成检测框架中的计算冗余和误差传递问题。在传统方法中，辅助红外图像首先被生成出来，然后再输入到独立的检测网络中，这一流程往往导致多次重复计算，并且会带来不可避免的误差累积，从而对检测性能产生负面影响。而通过将生成和检测统一到一个框架中，可以实现端到端的联合优化，消除独立阶段之间的中间信息传递障碍。具体而言，这种一体化设计不仅减少了计算冗余，还有效降低了由于生成与检测过程独立导致的误差传递，从而提升了整个系统的精度和效率。此外，由于多尺度特征图的共享，本文框架可以更好地适配不同的检测模型，具有较高的灵活性和可扩展性。

在实现生成和检测一体化的过程中，面临的主要挑战在于如何同时保证生成阶段和检测阶段的有效协同优化。具体而言，一方面，生成过程必须产生对检测任务有利的辅助红外图像，而不仅仅是逼真性高的红外图像；另一方面，如何将生成过程中提取的特征无缝融入检测阶段也是一个关键问题。为了解决这些挑战，采用了基于条件扩散模型的生成过程，并通过对多尺度特征的提取与传递实现检测学习的优化。在辅助红外生成阶段，两个具有相同架构的 U-Net 网络并行工作。将输入的 RGB 图像表示为 $\boldsymbol{I}_{\text{vis}} \in \mathbf{R}^{512\times512\times3}$，输入的红外图像表示为 $\boldsymbol{I}_{\text{inf}} \in \mathbf{R}^{512\times512\times3}$，生成的潜空间辅助红外图表示为 $\boldsymbol{z}'_{\text{inf}} \in \mathbf{R}^{64\times64\times4}$，提取的多尺度特征图表示为 $\boldsymbol{F}_{\text{det}} = \left[\boldsymbol{F}^{i}_{\text{inf}} \in \mathbf{R}^{w_i\times h_i\times c_i}\right]^{2}_{i=0}$，即

$$\boldsymbol{z}'_{\text{inf}}, \boldsymbol{F}_{\text{det}} = U(F(V_E(\boldsymbol{I}_{\text{inf}}) + z_t, V_E(\boldsymbol{I}_{\text{vis}}))) \tag{1}$$

式中，V_E、F 和 U 分别为辅助红外生成阶段的 VAE 编码器、跨模态融合模块以及 U-Net，z_t 代表步长为 t 的随机噪声。

本文通过优化生成的潜空间红外图 $\boldsymbol{z}'_{\text{inf}}$ 与经过 VAE 编码前的空间红外图 $\boldsymbol{z}_{\text{inf}}$ 之间的均方误差损失 L_{gen}，以实现从可见光域到红外域的有效映射。与此同时，在目标检测阶段，充分利用生成阶段提取的多尺度特征图 $\boldsymbol{F}_{\text{det}}$，这些特征经过检测模型的 Neck 和 Head 网络，最终输出边界框及目标类别，并通过 L_{det} 损失函数进行行人目标检测任务的学习。

通过上述设计，避免了使用传统 VAE 解码器导致的梯度无法有效反向传播的问题，从而实现了真正意义上的端到端训练。这种方式不仅能够简化训练流程，还可以显著提升生成与检测之间的协同效应，使得整个框架可以轻松适应不同的检测模型，例如本文使用的 YOLOv3(you only look once)(Redmon 和 Farhadi，2018)。此外，在联合优化过程中并未对检测网络的各项损失手动分配权重，而是依据扩散过程中采样步长 t 动态优化生成与检测过程的损失关系，使得整体系统的性能更加稳健和高效。

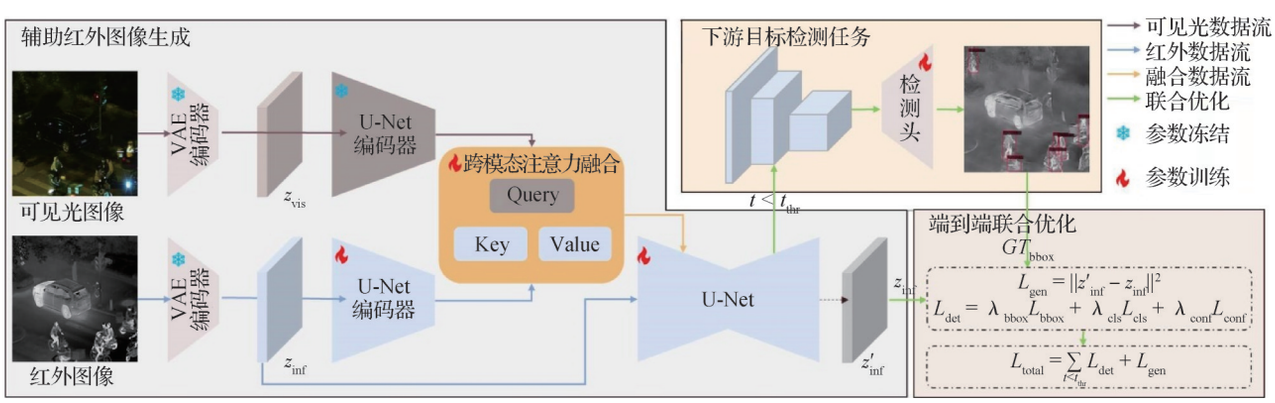


图 1　本文提出的生成检测一体化框架

Fig. 1　The integration of generation and detection framework proposed in this paper

1.2 辅助红外图像生成

本文采用了一种类似ControlNet(Zhang等，2023)结构的跨模态特征融合方法，以实现从可见光域到红外域的特征转换，旨在增强目标行人在辅助红外图像中的可见性与对比度。如图1所示，辅助红外图像生成网络的整体结构由以下3个主要模块组成。

1)VAE编码器。VAE编码器模块主要用于将输入图像映射到潜在空间中，从而获得低维度的特征表示。通过这种降维操作，不仅降低了后续计算的复杂性，也为生成网络提供了丰富的图像特征，便于更有效地进行多模态特征的融合。

2)U-Net。U-Net通过跳跃连接保留了输入图像的精细信息，并通过多层卷积结构提取多尺度特征。在本文中，输入为可见光域的暗光图像，输出为生成的辅助红外图像。跳跃连接的引入，使得网络能够更好地整合不同尺度的上下文信息，进而提升辅助红外图像生成的细节表现力。

3)跨模态注意力融合。为实现跨模态特征的充分融合，引入了交叉注意力机制与零卷积相结合的模块。交叉注意力机制用于在潜在空间的多模态特征间建立关联，从而实现高效的信息交换，而零卷积用于融合过程中保留原始特征的多样性和稳定性，最终提高辅助红外图像的生成质量。

给定RGB图像$\boldsymbol{I}_{\text{vis}}$，辅助红外图像生成阶段不再预测添加的噪声，而是直接输出生成的辅助红外图像的潜在空间采样$\boldsymbol{z}'_{\text{inf}} \in \mathbf{R}^{64\times64\times4}$，用于计算图像生成损失。具体而言，首先将红外模态和可见光模态的图像分别输入参数冻结的VAE编码器，得到潜空间中的图像表示$\boldsymbol{z}_{\text{inf}}$、$\boldsymbol{z}_{\text{vis}} \in \mathbf{R}^{64\times64\times4}$。随后使用两个U-Net编码器分别提取$\boldsymbol{z}_{\text{inf}}$和$\boldsymbol{z}_{\text{vis}}$中的特征，其中作用于红外模态的U-Net参数是可训练的而作用于可见光模态的U-Net参数被冻结，从而得到两组多尺度特征图$\boldsymbol{F}_{\text{inf}} = \left[\boldsymbol{F}_{\text{inf}}^{i} \in \mathbf{R}^{w_i\times h_i\times c_i}\right]_{i=0}^{12}$和$\boldsymbol{F}_{\text{vis}} = \left[\boldsymbol{F}_{\text{vis}}^{i} \in \mathbf{R}^{w_i\times h_i\times c_i}\right]_{i=0}^{12}$。接下来，将红外模态的多尺度特征图作为键(Key)和值(Value)，将可见光模态的多尺度特征图作为查询(Query)，输入特征融合模块。通过交叉注意力机制，计算得到的注意力图被用做U-Net编码器下采样过程中的额外残差$\boldsymbol{F}_{\text{res}} = \left[\boldsymbol{F}_{\text{inf}}^{i} \in \mathbf{R}^{w_i\times h_i\times c_i}\right]_{i=0}^{12}$。这些额外残差进一步输入到红外模态生成U-Net的主干网络中，从而增强了网络对跨模态特征的表达能力。最后，利用引入的额外残差$\boldsymbol{F}_{\text{res}}$与红外模态的潜空间表示$\boldsymbol{z}_{\text{inf}}$经过U-Net直接生成采样结果$\boldsymbol{z}'_{\text{inf}}$。本文通过均方误差(mean squared error，MSE)损失函数L_{gen}优化红外模态图像的生成质量，从而实现从可见光域到红外域的有效转换。具体为

$$L_{\text{gen}} = \left\| \boldsymbol{z}'_{\text{inf}} - \boldsymbol{z}_{\text{inf}} \right\|^2 \tag{2}$$

通过上述方法，本文在辅助红外图像生成过程中充分利用了可见光和红外模态的多尺度特征信息，实现了跨域信息的有效融合。这种设计不仅提高生成图像的视觉质量，还为后续的目标检测与识别任务提供了更具判别性的特征。

1.3 下游目标检测任务

为了实现与下游目标检测任务的有效衔接，本文从负责红外模态生成的主干U-Net解码器的中间输出中提取3个尺度的特征图，形成红外模态的多尺度特征集合$\boldsymbol{F}_{\text{det}} = \{\boldsymbol{F}_{\text{inf}}^{1}, \boldsymbol{F}_{\text{inf}}^{2}, \boldsymbol{F}_{\text{inf}}^{3}\}$，这些红外模态特征图被输入到检测头中，用于目标检测任务的学习，并通过计算检测损失优化检测器的性能。具体而言，U-Net解码器会输出5个尺度的特征图组合。为了适配检测头的需求，从中选取尺寸为(320, 64, 64)、(640, 32, 32)、(1 280, 16, 16)的3个尺度特征图，作为$\boldsymbol{F}_{\text{det}}$输入到检测头中。这些特征图包含了不同尺度下的丰富信息，有助于检测不同大小的目标。

在目标检测任务中，损失函数由4个主要部分组成：边界框中心点损失$L_{\text{bbox}_{xy}}$、边界框宽高损失$L_{\text{bbox}_{wh}}$、分类损失L_{cls}和置信度损失L_{conf}。具体如下：

1)边界框中心点损失$L_{\text{bbox}_{xy}}$。使用的是带有sigmoid激活的二元交叉熵损失来衡量预测边界框与真实边界框之间的差异，以确保目标定位的准确性。损失函数定义为

$$L_{\text{bbox}_{xy}} = -\sum_{i=1}^{N}\left(y_i^{xy}\log(p_i^{xy}) + (1-y_i^{xy})\log(1-p_i^{xy})\right) \tag{3}$$

式中，y_i^{xy}表示第i个预测框的中心点真实标签，p_i^{xy}是预测的中心点坐标概率。

2)边界框宽高损失$L_{\text{bbox}_{wh}}$。使用均方误差来衡量预测边界框宽高与真实宽高之间的差异。损失函数定义为

$$L_{\text{bbox}_{wh}} = \sum_{i=1}^{N}\left((w_i - w_i^*)^2 + (h_i - h_i^*)^2\right) \tag{4}$$

式中，w_i和h_i是第i个预测框的宽高，w_i^*和h_i^*是对应

的真实宽高。

3)分类损失L_{cls}。采用多分类交叉熵损失评估预测类别与实际类别之间的一致性，确保检测器能够正确识别各类目标。损失函数定义为

$$L_{\text{cls}} = -\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(p_{i,c}) \tag{5}$$

式中，$y_{i,c}$是第i个样本在第c类的真实标签，$p_{i,c}$是预测的类别概率，C为类别总数。

4)置信度损失L_{conf}。评估预测目标存在性的置信度，通常采用二元交叉熵损失度量预测值与真实值之间的误差。损失函数定义为

$$L_{\text{conf}} = -\sum_{i=1}^{N}(y_i^{\text{obj}}\log(p_i^{\text{obj}}) + (1 - y_i^{\text{obj}})\log(1 - p_i^{\text{obj}})) \tag{6}$$

式中，y_i^{obj}表示第i个预测框是否包含目标的真实标签，p_i^{obj}是预测的置信度概率。

总的检测损失L_{det}为上述4个损失的加权和，即

$$L_{\text{det}} = \lambda_{\text{bbox}_{xy}}L_{\text{bbox}_{xy}} + \lambda_{\text{bbox}_{wh}}L_{\text{bbox}_{wh}} + \lambda_{\text{conf}}L_{\text{conf}} \tag{7}$$

式中，$\lambda_{\text{bbox}_{xy}}$、$\lambda_{\text{bbox}_{wh}}$、$\lambda_{\text{cls}}$、$\lambda_{\text{conf}}$为损失项的权重系数，分别取值2.0、2.0、1.0、1.0，用于平衡各部分对总损失的贡献。

通过将多尺度特征图$\boldsymbol{F}_{\text{det}}$输入到检测头，并基于上述损失函数进行优化，能够有效地提升检测器的性能。值得注意的是，由于本文直接利用U-Net解码器的中间特征作为检测器的输入，避免了传统方法中额外的特征提取过程，减少了计算开销，同时也有利于实现端到端的训练。此外，这种设计使得生成和检测过程紧密结合，在生成高质量辅助红外图像的同时，确保了目标检测的精度。这种一体化的框架不仅提高模型的效率，而且在暗光条件下显著提升了行人检测的性能。

1.4 端到端联合优化

在提出的框架中，辅助红外图像生成与目标检测任务通过端到端的方式进行联合优化。然而，生成的辅助红外图像直接输入检测器训练时存在梯度无法有效回传的问题，这是由于传统方法中需要通过VAE解码器恢复图像，从而导致梯度无法有效地反向传播。为了解决这一问题，采用了一种创新的策略，即直接利用生成过程中的多尺度特征，绕开VAE解码阶段，从而确保梯度可以顺利地从目标检测任务回传到生成阶段。具体而言，在辅助红外生成阶段，从负责红外模态生成的主干 U-Net 解码器的中间输出中提取多尺度特征图$\boldsymbol{F}_{\text{det}}$。这些特征图不经过VAE解码器，而是直接输入到目标检测头（如本文使用的 YOLOv3 的 Detect Neck 和 Detect Head）中进行目标边界框和类别的预测，从而有效地避免了由于使用VAE解码器而导致的梯度无法回传问题。

与传统方法不同，本文不再为检测网络的各项损失手动分配权重，而是利用扩散过程中的采样步长t自适应地联合优化生成与检测过程。在扩散过程中，对随机选择的采样步长t进行判断。当t大于预设的阈值t_{thr}时，认为图像中噪声较多，生成的特征不利于目标检测任务的训练；相反，当t小于t_{thr}时，认为图像清晰度较高，或经过了适当的数据增强，适合进行目标检测任务的训练。因此，对于每一个满足小于t_{thr}的采样步长t，本文从 U-Net 解码器的中间输出中提取多尺度特征图$\boldsymbol{F}_{\text{det}}$，这些特征图直接输入到目标检测头中，用于预测最终的目标边界框和类别。具体的联合优化过程如下：

1)生成阶段。利用扩散模型生成辅助红外图像，并提取对应的多尺度特征图$\boldsymbol{F}_{\text{det}}$。如式(2)所示，生成损失$L_{\text{gen}}$用于衡量生成图像与真实红外图像之间的差异。

2)检测阶段。将提取的多尺度特征图$\boldsymbol{F}_{\text{det}}$输入到检测模型中，预测目标边界框*bbox*和目标类别*cls*。如式(7)所示，检测损失L_{det}包括边界框中心点损失、边界框宽高损失、分类损失和置信度损失。

3)联合损失函数。通过对生成损失和检测损失的联合优化，最小化以下总损失，具体为

$$L_{\text{total}} = \sum_{t < t_{\text{thr}}} L_{\text{det}} + L_{\text{gen}} \tag{8}$$

该损失函数在满足小于t_{thr}的采样步长t上对检测损失进行求和，确保在图像清晰或噪声适中的情况下同时优化生成和检测任务。

通过这种端到端的联合优化策略，本文在扩散模型的生成过程和目标检测任务之间建立了紧密的联系，实现了两个任务的协同优化。这种方法在提升暗光条件下行人检测性能的同时，还为跨模态目标检测提供了新的思路。通过直接利用生成过程中的特征而不是经过VAE解码，可以确保梯度能够从检测网络顺利地回传到生成网络，从而实现有效的联合训练。

2 实验与结果分析

2.1 数据集描述和实验设计

2.1.1 数据集

实验主要采用LLVIP(low-light visible-infrared paired dataset)数据集(Jia等,2021)和VTMOT(visible-thermal multiple object tracking dataset)数据集(Zhu等,2024)进行分析和验证。LLVIP数据集包含来自可见光和红外两个模态的15 488对图像,总计30 976幅图像。其中,部分目标处于极度黑暗的光照条件下,同时数据集中还提供了行人目标检测的标注标签。VTMOT数据集包括来自监控、无人机和手持平台的582个视频序列对,共计可见光和红外两个模态的401 000对图像。本研究根据可见光模态数据筛选了其中夜间采集的71组视频帧序列对,共9 400对配对图像。同时数据集中还提供了行人目标检测的标注标签,其中部分目标存在与背景对比度低、目标尺度小等问题。这使得这两个数据集在评估暗光条件下的行人检测模型性能具有重要意义。

2.1.2 实验设计

首先根据整幅图像的平均亮度以及行人目标与背景的亮度对比度从LLVIP和VTMOT数据集中挑选出具有高度挑战性的可见光图像作为测试集,以确保测试样本能够充分体现暗光和复杂环境下的检测难度。图像平均亮度计算为

$$Y = 0.299R + 0.587G + 0.114B \tag{9}$$

$$L_{\text{avg}} = \frac{1}{N}\sum_{i=1}^{N} Y(i) \tag{10}$$

式中,R、G、B为图像的3个通道,考虑人眼对不同颜色的敏感度采用加权平均的方法将RGB图像转为亮度图,N是图像中所有像素的数量。

行人目标与背景的亮度对比度计算为

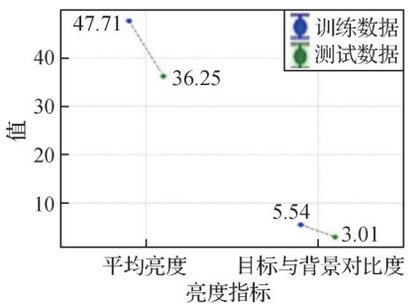
$$C = \frac{1}{M}\sum_{i=1}^{M} \frac{B(i)_{\text{avg}}}{L_{\text{avg}}} \tag{11}$$

式中,M是图像中所有边界框的数量,$B(i)_{\text{avg}}$与L_{avg}为通过式(10)计算的边界框平均亮度和图像平均亮度。

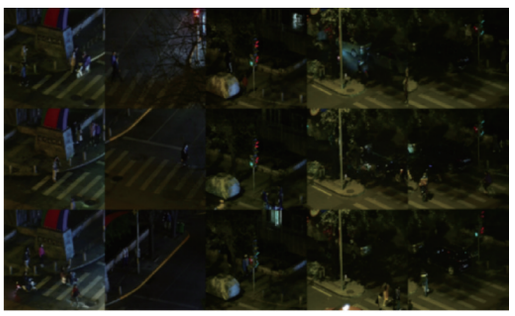
LLVIP数据集所选测试集的平均亮度为36.25,低于训练集的47.71,训练集和测试集的行人与背景对比度分别为5.54和3.01,均处于较低水平。VTMOT数据集所选测试集的平均亮度为52.05,远低于训练集的25.91,训练集和测试集的行人与背景对比度为分别为10.61和9.91,同样处于较低水平。LLVIP和VTMOT测试集与整个数据集的亮度评估指标对比结果与测试集样例如图2所示。

同时,在保证不包含测试集数据的条件下,随机打乱原始数据集中的所有图像,并将其中的80%作为训练集,剩余20%作为验证集。数据划分时,确保训练集、验证集与测试集中包含数据集中所有可能的场景,以保证模型的泛化能力。具体而言,LLVIP数据集中包含12 304对可见光和红外训练样本、3 076对可见光和红外验证样本以及108幅可见光图像作为测试样本;VTMOT数据集中包含8 537对可见光和红外训练样本、2 107对可见光和红外验证样本以及61幅可见光图像作为测试样本。

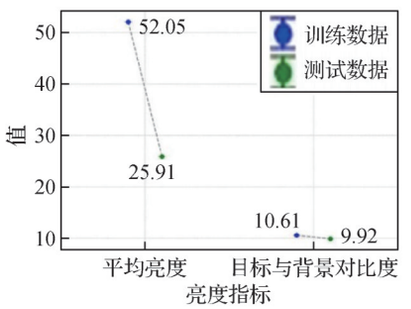
在2.2节中,本文对增强与检测割裂的传统范式及所提出的生成检测一体化范式进行了系统性的定量和定性对比实验。传统范式通常先对原始数据进行增强处理,随后再训练检测模型。具体而言,在同域增强范式中,选择了直方图均衡化和自校准照明(SCI)两种图像增强方法进行对比。直方图均衡化通过调整图像亮度分布来提升对比度;而SCI方法则引入了自校准照明模块,动态调整图像的亮度和对比度,并采用特定损失函数优化网络,以确保增强后图像在视觉上更为自然。对于跨域增强范式,选用了pix2pix和扩散模型(diffusion)两种生成模型进行对比。pix2pix是一种基于生成对抗网络(GAN)的图像到图像转换方法,而扩散模型则利用可见光模态信息作为条件引导去噪过程,从而生成红外模态图像。本文在常用的检测指标和特定评估体系上进行了定量分析,并针对可视化检测结果及生成的红外图像进行了定性评估。在2.3节中,本文将当前广泛使用的行人检测模型与所提出的方法在常规检测指标上进行了定量对比。所比较的行人检测模型均来自Pedestron框架(Hasan等,2021),每个模型均使用其默认配置并加载预训练权重进行训练,这使得本文的实验能够在一个公平的基础上评估不同方法的有效性。在2.4节中,为了验证所提方法中各个模块的有效性,进行了详细的消融实验。这些实验旨在单独评估每个组件对整体性能的影响,从而证明各模块在提升行人检测准确性方面的重要作用。



(a) LLVIP 测试样本与训练样本的亮度指标对比



(b) LLVIP 测试集样例



(c) VTMOT 测试样本与训练样本的亮度指标对比

(d) VTMOT 测试集样例

图 2　测试样本与训练样本的亮度指标对比与测试集样例展示

Fig. 2　Comparison of luminance metrics between test samples and training samples with test datasets sample demonstration

((a) luminance metrics of LLVIP test samples vs. training samples; (b) samples of LLVIP test dataset; (c) luminance metrics of VTMOT test samples vs. training samples; (d) samples of VTMOT test dataset)

2.1.3　训练细节

实验在 Ubuntu 20.04 平台上进行，使用 PyTorch 框架实现，硬件配置为 RTX A6000 GPU（显存 48 GB）。采用 Adam 优化器进行模型的参数更新，训练的批量大小设定为 4，训练历时 100 个 epoch。初始学习率设定为 1×10^{-5}，并设置 500 个 step 用于学习率的预热阶段，随后采用余弦退火策略逐步降低学习率，以提高模型收敛的稳定性。此外，在联合训练生成与检测模块时，关键参数 t_{thr} 设置为 100，以平衡检测与生成任务的性能。

2.2　不同暗光目标检测增强范式的效果比较

为了验证所提出方法在暗光环境下的行人检测性能，以及其相较于单一可见光模态的提升效果，将本文方法与同域增强和跨域增强的多种方法在 LLVIP 与 VTMOT 两个数据集上进行了实验，其中以常用检测指标评估的定量对比结果如表 1 所示，同时针对特定的任务场景设计了一套评估指标，其定量结果如表 2 所示，在 LLVIP 数据上的定性对比结果和生成红外模态辅助数据的可视化结果如图 3 和图 4 所示。

表 1 中的“单一可见光模态”范式采用的是 MMDetection 提供的 YOLOv3 模型，并加载 open-mmlab 提供的 darkent53 权重进行训练。“同域增强”与“跨域增强”范式旨在通过预先的数据增强改善暗光条件下的图像质量。这些对比方法都采用先增强后训练的方式，即先使用不同的图像增强手段对原始数据进行质量改善，在此基础上再对“单一可见光模态”范式的 YOLOv3 模型进行训练。而本文提出的“一体化”范式则将图像质量改善与目标检测模型的训练整合在一个端到端框架中。在这一框架中，红外图像生成模块可见光支路的 U-Net 使用 Stable Diffusion（Rombach 等，2022）预训练权重进行初始化，红外支路的 U-Net 须使用预训练的红外模态文生图权重进行初始化。该设计保证了本节实验中的下游检测模型的一致性，并能够严谨地验证本文方法在缓解图像增强与检测任务之间割裂问题上的有效性。

根据表 1 中的实验结果，本文方法在准确率（precision）、召回率（recall）、F1 得分（F1 score）和平均精度均值（mean average precision，mAP）这 4 项关键指标上均表现出明显优势。具体而言，在 LLVIP 数据集上，本文方法的准确率、召回率、F1 得分和

表1　两种传统范式与本文方法在常用检测指标上的比较

Table 1　Comparison of two traditional paradigms with the method of this paper on commonly used detection metrics

/%

方法	范式	LLVIP数据集				VTMOT数据集			
		准确率	召回率	F1 score	mAP	准确率	召回率	F1 score	mAP
YOLOv3	单一可见光模态	85.67	62.34	72.17	63.34	90.05	73.04	80.66	68.21
直方图均衡化	同域增强	90.26	74.89	81.86	71.99	94.43	71.97	81.69	67.06
SCI	同域增强	86.18	78.30	82.05	72.30	**95.10**	78.34	85.91	72.19
pix2pix	跨域增强	88.79	65.74	75.55	68.69	81.64	69.85	75.29	60.89
diffusion	跨域增强	89.67	70.21	78.76	70.21	92.54	71.13	80.43	67.03
本文	一体化	**94.15**	**78.72**	**85.75**	**75.35**	93.18	**87.05**	**90.01**	**73.44**

注：加粗字体表示各列最优结果。

表2　两种传统范式与本文方法在特定评估体系上的比较

Table 2　Comparison of the two traditional paradigms with the method of this paper on specific assessment systems

方法	范式	LLVIP数据集			VTMOT数据集		
		新增检测框数	正确数(比例/%)	错误数(比例/%)	新增检测框数	正确数(比例/%)	错误数(比例/%)
直方图均衡化	同域增强	100	77(77.00)	23(23.00)	63	25(39.68)	38(60.32)
SCI	同域增强	140	96(68.57)	44(31.43)	55	16(29.09)	39(70.91)
pix2pix	跨域增强	84	62(73.81)	22(26.19)	163	81(49.69)	82(50.31)
diffusion	跨域增强	114	79(69.30)	35(30.70)	86	22(25.58)	64(74.42)
本文	一体化	93	80(**86.02**)	13(**13.98**)	163	99(**60.74**)	64(**39.26**)

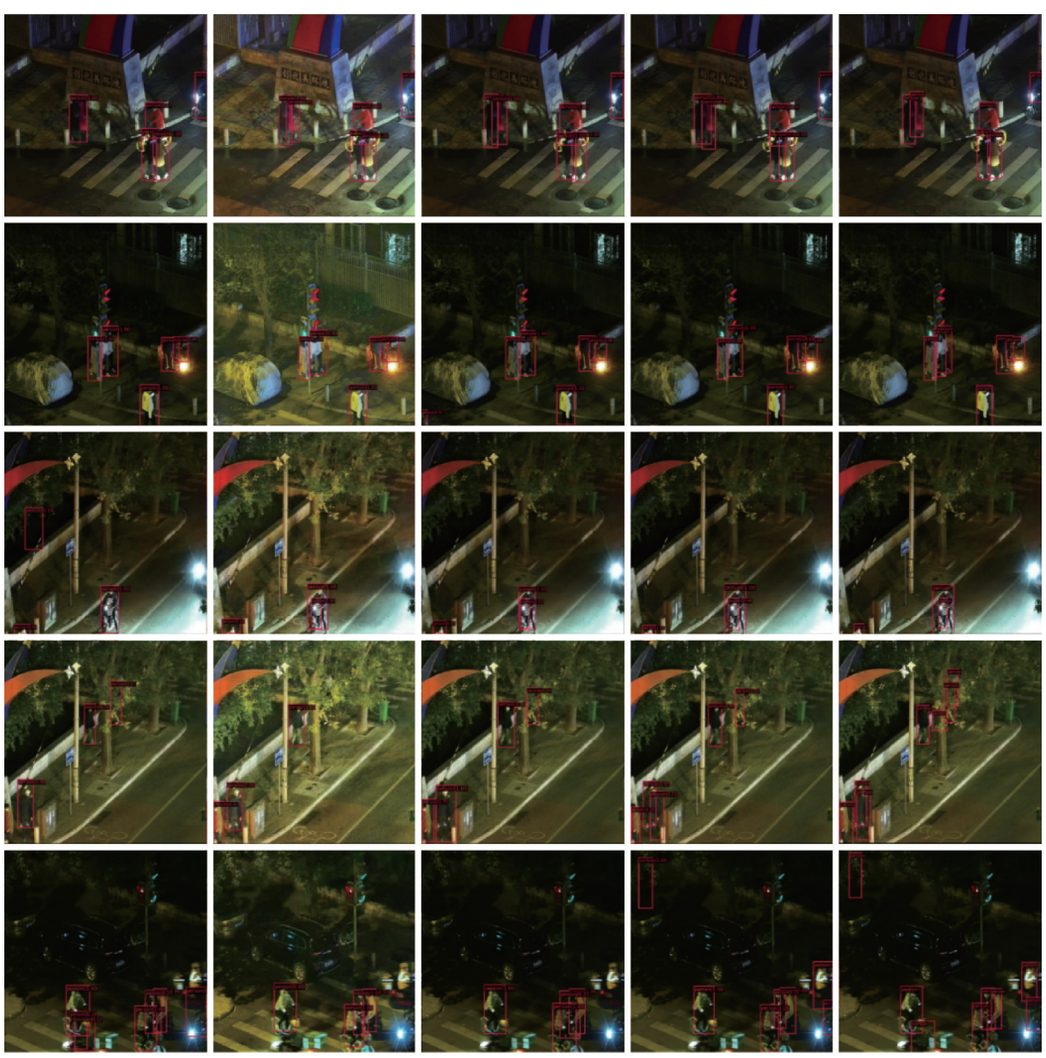
注：加粗字体表示各列最优结果。

mAP分别为94.15%、78.72%、85.75%和75.35%；在VTMOT数据集上，分别为93.18%、87.05%、90.01%和73.44%。相较于同域增强范式中的最优方法(SCI)，本文方法在LLVIP数据集上的mAP提升3.05%(从72.30%提升至75.35%)，F1分数提升3.70%(从82.05%提升至85.75%)。在VTMOT数据集上，本文方法的mAP提升1.25%(从72.19%提升至73.44%)，F1分数提升4.10%(从85.91%提升至90.01%)。这些提升表明本文方法成功通过可见光模态提取红外模态的信息，并充分利用辅助红外数据，表现出更高的检测平衡性和可靠性。此外，与跨域增强范式中的最优方法(diffusion)相比，本文方法在LLVIP数据集上的mAP提升5.14%(从70.21%提升至75.35%)，F1分数提升6.99%(从78.76%提升至85.75%)。在VTMOT数据集上，本文方法的mAP提升6.41%(从67.03%提升至73.44%)，F1分数提升9.58%(从80.43%提升至90.01%)。这些结果显示，本文提出的通过生成红外模态辅助数据并进行目标检测端到端联合优化的框架具有明显效果，不仅显著提升了暗光条件下的行人检测精度，还有效解决了生成过程中梯度无法经过VAE解码器回传的问题，验证了预训练扩散模型强大的图像生成能力和特征提取能力在行人检测任务中的优越性。

本文设计了一套针对特定场景的评估体系，包含3个关键指标：新增有效检测框数、正确数及比例、错误数及比例。这些指标的目的是用来衡量生成辅助模态后的检测能力提升程度，即在可见光模态的基础上，辅助模态能够带来多少额外检测到的目标。其中，新增有效检测框数表示各方法在单模态可见光检测结果的基础上，通过生成红外模态辅助信息所新增的有效检测框数。正确数及比例用来衡量新增有效检测框中有多少是真正有效的(即与真实标注有高度匹配)，并以此反映新增检测框的有效性和准确性。根据表2的实验结果显示，在LLVIP

数据集上，本文方法的新增有效检测框数为 93，其中正确数为 80（占比 86. 02%），错误数为 13（占比 13. 98%）；而在 VTMOT 数据集上，新增有效检测框数为 163，正确数为 99（占比 60. 74%），错误数为 64（占比 39. 26%）。相较于其他方法，本文方法在正确数比例上表现出显著优势，例如在 LLVIP 数据集中，SCI 方法的正确数比例为 68. 57%，diffusion 方法为 69. 30%，而本文方法达到了 86. 02%。尽管新增有效检测框数略低于某些方法，但本文方法生成的检测框更为可靠，大部分新增框都是正确的，有效减少了漏检的情况，体现了较高的检测质量。此外，本文方法在抑制误检方面表现优异，LLVIP 数据集上的错误数占比仅为 13. 98%，明显低于 SCI 方法的 31. 43% 和 diffusion 方法的 30. 70%。在 VTMOT 数据集上，本文方法的错误数占比为 39. 26%，也低于 SCI 方法的 70. 91% 和 diffusion 方法的 74. 42%。这些结果表明，本文方法在抑制误检方面表现优异，保证了检测结果的高可靠性。

图 3 展示了多种方法的检测结果在 LLVIP 测试集可见光图像上的定性对比结果。可以看出，本文方法在暗光环境下能够更好地融合生成红外模态和可见光模态的信息，从而有效地检测出行人目标。在 SCI 同域增强和 diffusion 跨域增强方法下，虽然对部分行人目标进行了检测，但是存在误检和漏检的情况。而本文方法通过生成和检测一体化的联合优化策略，有效地提高检测的全面性和准确性，尤其是



(a) 可见光单模态　(b) SCI 同域增强　(c) diffusion 跨域增强　(d) 本文方法　(e) 真实标签

图 3　两种传统范式与本文方法在 LLVIP 数据集上的可视化结果比较

Fig. 3　Comparison of visualization results on LLVIP dataset between two traditional paradigms and our method ((a) visible single mode; (b) SCI homology domain enhancement; (c) diffusion cross domain enhancement; (d) ours; (e) ground truth)

在复杂光照环境下能够更加稳定地检测行人。

图4展示了跨模态增强方法与本文方法生成红外数据的定性对比结果。首先需要说明的是，pix2pix方法与diffusion方法都是直接学习从可见光图像生成红外图像，在此基础上再对检测模型进行训练。而本文方法仅使用生成的红外模态多尺度特征 $\boldsymbol{F}_{\text{det}}=\{\boldsymbol{F}_{\text{inf}}^{1},\boldsymbol{F}_{\text{inf}}^{2},\boldsymbol{F}_{\text{inf}}^{3}\}$ 输入检测模块进行训练。可以看出，pix2pix生成的红外图像存在很多的噪点且目标边缘不清晰，给目标检测任务带来了更多挑战。diffusion和本文方法所生成的红外图像都更为接近真实红外图像，且diffusion所生成的结果在细节上更加丰富。但由于diffusion增强方法在图像增强和目标检测间存在割裂，导致其定量结果不如本文提出的生成检测一体化方法。

2.3 现有行人检测方法与本文方法的效果比较

为了验证所提方法相较于现有技术的优越性，在LLVIP数据集与VTMOT数据集上进行详尽的定量对比实验。具体而言，将所提方法与当前在“Papers with Code”平台上排名领先的开源框架Pedestron（Hasan等，2021）中提供的算法进行比较，评估指标涵盖了标准的检测性能度量。在对比过程中，Pedestron中的各个模型均按照其默认配置进行设置，并加载预训练权重以确保实验的一致性和可重复性。这样的设置保证了本文在公平且统一的基础上对不同方法的有效性进行评估。表3展示了基于常用检测指标的定量对比结果。

图4 跨模态增强与本文方法生成红外辅助数据在LLVIP数据集上的可视化结果比较

Fig. 4 Comparison of cross-modal enhancement with our method for generating infrared auxiliary data for visualization results at LLVIP dataset((a) visible images; (b) infrared images generated by pix2pix; (c) infrared images generated by diffusion; (d) infrared images generated by ours; (e) realistic infrared images)

由表3数据可知，本文方法在行人检测的准确率、召回率、F1分数和平均精度（mAP）等常用检测指标上均表现出明显优势。具体而言，在LLVIP数据集上，本文方法的准确率、召回率、F1分数和mAP分别为94.15%、78.72%、85.75%和75.35%；在VTMOT数据集上，分别为93.18%、87.05%、90.01%和73.44%。本文方法在总体指标F1分数和mAP上均高于其他对比方法。例如，在LLVIP数据集上，Faster_RCNN_hrnet方法的F1分数为82.91%，mAP为70.02%，而本文方法的F1分数为85.75%，mAP为75.35%；在VTMOT数据集上，Faster_RCNN_hrnet方法的F1分数为84.48%，mAP为71.16%，而本文方法的F1分数为90.01%，mAP为73.44%。虽然本文方法在两个数据集上的召回率均略低于Faster_RCNN_hrnet方法，但Faster_RCNN_hrnet方法存在较高的误检率，总体指标与本文方法存在一定的差距。这些结果表明，本文方法在行人检测任务中能够更有效地识别行人，同时保持较低的误检率，从而在实际应用中具有更高的价值。

总体来看，本文提出的生成检测一体化框架不仅在检测精度上取得了显著优势，其性能的突出还体现在特征提取、模态融合以及误差抑制等多个方面的深层次优化。在与通用行人检测方法对比下，本文方法展现出在暗光条件下进行图像增强和检测任务的可靠性和稳健性。通过这种生成和检测的联合策略，本文方法在处理复杂光照条件的行人检测任务中展现了更为优越的表现，为进一步研究暗光环境下的多模态行人检测提供了有力支持。

表3 行人检测先进方法与本文方法在常用检测指标上的比较

Table 3 Comparison of advanced methods for pedestrian detection with our method on commonly used detection metrics

/%

方法	LLVIP数据集				VTMOT数据集			
	准确率	召回率	F1 score	mAP	准确率	召回率	F1 score	mAP
retinanet_ResNeXt101	93.41	69.36	79.61	63.38	58.30	84.29	68.92	54.20
csp_r50	78.21	68.72	73.16	55.82	86.40	49.89	63.26	50.91
Faster_RCNN_hrnet	77.01	**89.79**	82.91	70.02	79.92	**89.60**	84.48	71.16
本文	**94.15**	78.72	**85.75**	**75.35**	**93.18**	87.05	**90.01**	**73.44**

注：加粗字体表示各列最优结果。

2.4 消融实验

本节深入探讨了辅助红外生成模块和生成检测联合优化策略在暗光条件下行人检测任务中的重要性和必要性。为了验证各个模块的具体贡献，在LLVIP数据上进行了消融实验，以单模态检测模型作为基线进行对照，设置了不同的模块组合，实验结果如表4所示。

实验1是基线方法，仅利用可见光数据进行检测，不包含辅助红外生成模块和生成检测联合优化策略。其检测性能在精准率、召回率、F1分数和mAP上分别为85.67%、62.34%、72.17%和63.34%。模型仅依赖可见光数据，由于暗光条件下的光照不足，行人与背景的对比度较低，导致检测模型对行人的识别能力受限，尤其是在召回率方面表

表4 引入不同模块的效果比较

Table 4 Comparison of the effects of introducing different modules

/%

实验	检测模块	辅助红外生成模块	生成检测联合优化策略	准确率	召回率	F1 score	mAP
实验1	√	×	×	85.67	62.34	72.17	63.34
实验2	√	√	×	89.67	70.21	78.76	70.21
实验2	√	√	√	**94.15**	**78.72**	**85.75**	**75.35**

注：加粗字体表示各列最优结果，“×”表示该实验中未包含对应模块，“√”表示该实验中加入对应模块。

现不佳，容易出现漏检情况。这表明，单纯依靠可见光信息的模型难以应对复杂的光照变化和暗光场景，模型在这种环境下缺乏对行人特征的有效提取能力。

实验2在基线方法的基础上引入辅助红外生成模块。该模块通过生成红外图像增强行人与背景之间的对比，使得行人的特征更加显著，改善了检测效果。实验结果显示，实验2的准确率提升至89.67%，召回率提升至70.21%，F1分数达到78.76%，mAP提升到70.21%。相比基线方法，辅助红外生成模块显著改善了召回率，这说明红外信息在暗光条件下有效补充了可见光的不足，使得模型在识别行人方面的表现更加鲁棒，从而有效减少了漏检的情况。尤其是在暗光环境下，红外图像的引入使行人更加容易被区分，从而显著提高检测的整体性能。

实验3进一步在实验2基础上加入生成检测联合优化策略，即端到端的联合训练方式。该方法在训练过程中同时优化辅助红外生成和行人检测模块，使得这两个模块能够更好地协同工作，学习到不同模态数据之间的关联性。实验结果显示，这一联合训练策略在各项性能指标上均达到最优表现：准确率达到94.15%，召回率增加至78.72%，F1分数提升到85.75%，mAP值达到75.35%。相比实验2，引入联合优化策略的实验3进一步提高了准确率和召回率。这表明通过联合优化，模型能够更好地整合来自不同模态的信息，提升了对行人特征的综合表征能力，使得模型不仅在行人识别的准确性上有所提高，还显著增强了模型在复杂光照条件下的敏感性和鲁棒性，从而在准确性和全面性之间取得了更好的平衡。

总体而言，消融实验的结果清楚地表明了辅助红外生成模块和联合优化策略在暗光行人检测任务中的必要性。辅助红外生成模块通过增加多模态的信息来增强行人与背景的对比，使得检测性能得以显著提升。而生成检测联合优化策略进一步提升了特征学习的质量，使得模型能够在暗光条件下依然保持较高的检测性能。这些模块的引入显著提升了检测性能，使得模型在暗光条件下依然能够稳定地检测到行人，体现了所提方法的有效性和稳健性。

3 结 论

在暗光条件下的行人检测任务中，先数据增强后训练检测模型的范式普遍存在增强、生成过程和检测过程割裂，限制了高质量生成图像对检测器学习效果的显著提升。针对这一问题，本文提出一种新颖的端到端架构，利用条件扩散模型从暗光图像生成辅助红外模态的对应图像，并利用其多尺度特征进行目标检测。该框架将生成模型与检测模型结合，通过生成模型为原始图像生成高质量的辅助图像，提升行人特征的清晰度和对比度，进而提高检测算法的精度。同时采用联合训练的方式，将扩散模型生成红外图像的任务与行人检测任务结合在一个流水线中，避免了梯度无法经过VAE解码器传递的问题。值得注意的是，提出的框架可以替换多种检测模型，只需要调整接收特征的维度即可轻松适配。最后，通过在LLVIP和VTMOT数据集上的实验定量和定性地证明了所提出框架的有效性，并探索了网络架构各模块的必要性。但目前依靠扩散模型的解决方案无法避免计算量大、推理耗时长的问题，因此后续可以通过加速采样策略和轻量化生成网络结构等方法提升系统的实时性和整体效率。

参考文献（References）

Cao Y, Bin J, Hamari J, Blasch E and Liu Z. 2023. Multimodal object detection by channel switching and spatial attention//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Vancouver, Canada: IEEE: 403-411 [DOI: 10.1109/CVPRW59228.2023.00046]

Chen C, Chen Q F, Xu J and Koltun V. 2018. Learning to see in the dark//Proceedings of 2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 3291-3300 [DOI: 10.1109/CVPR.2018.00347]

Chen F J, Zhu F, Wu Q X, Hao Y M, Wang E D and Cui Y G. 2021. A survey about image generation with generative adversarial nets. Chinese Journal of Computers, 44(2): 347-369 (陈佛计, 朱枫, 吴清潇, 郝颖明, 王恩德, 崔芸阁. 2021. 生成对抗网络及其在图像生成中的应用研究综述. 计算机学报, 44(2): 347-369) [DOI: 10.11897/SP.J.1016.2021.00347]

Fang R. 2023. Infrared and visible image fusion using generative adversarial network with feature complement block//Proceedings of the 6th International Conference on Electronics Technology (ICET).

Chengdu, China: IEEE: 133-138 [DOI: 10.1109/ICET58434.2023.10211883]

Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y. 2014. Generative adversarial nets//Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press: 2672-2680

Guo H F, Lu T and Wu Y R. 2021. Dynamic low-light image enhancement for object detection via end-to-end training//Proceedings of the 25th International Conference on Pattern Recognition (ICPR). Milan, Italy: IEEE: 5611-5618 [DOI: 10.1109/ICPR48806.2021.9412802]

Guo X J, Li Y and Ling H B. 2017. LIME: low-light image enhancement via illumination map estimation. IEEE Transactions on Image Processing, 26(2): 982-993 [DOI: 10.1109/TIP.2016.2639450]

Güzel S and Yavuz S. 2022. Infrared image generation from RGB images using CycleGAN//Proceedings of 2022 International Conference on Innovations in Intelligent Systems and Applications (INISTA). Biarritz, France: IEEE: 1-6 [DOI: 10.1109/INISTA55318.2022.9894231]

Hasan I, Liao S C, Li J P, Akram S U and Shao L. 2021. Generalizable pedestrian detection: the elephant in the room//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 11323-11332 [DOI: 10.1109/CVPR46437.2021.01117]

Ho J, Jain A and Abbeel P. 2020. Denoising diffusion probabilistic models//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 6840-6851

Hwang S, Park J, Kim N, Choi Y and Kweon I S. 2015. Multispectral pedestrian detection: benchmark dataset and baseline//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE: 1037-1045 [DOI: 10.1109/CVPR.2015.7298706]

Isola P, Zhu J Y, Zhou T H and Efros A A. 2017. Image-to-image translation with conditional adversarial networks//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE: 5967-5976 [DOI: 10.1109/CVPR.2017.632]

Jia X Y, Zhu C, Li M Z, Tang W Q and Zhou W L. 2021. LLVIP: a visible-infrared paired dataset for low-light vision//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). Montreal, Canada: IEEE: 3489-3497 [DOI: 10.1109/ICCVW54120.2021.00389]

Kruthiventi S S S, Sahay P and Biswal R. 2017. Low-light pedestrian detection from RGB images using multi-modal knowledge distillation//Proceedings of 2017 IEEE International Conference on Image Processing (ICIP). Beijing, China: IEEE: 4207-4211 [DOI: 10.1109/ICIP.2017.8297075]

Li C Y, Guo J C, Porikli F and Pang Y W. 2018. LightenNet: a convolutional neural network for weakly illuminated image enhancement. Pattern Recognition Letters, 104: 15-22 [DOI: 10.1016/j.patrec.2018.01.010]

Ma J Y, Xu H, Jiang J J, Mei X G and Zhang X P. 2020. DDcGAN: a dual-discriminator conditional generative adversarial network for multi-resolution image fusion. IEEE Transactions on Image Processing, 29: 4980-4995 [DOI: 10.1109/TIP.2020.2977573]

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 (CVPR). New Orleans, USA: IEEE: 5627-5636 [DOI: 10.1109/CVPR52688.2022.00555]

Pizer S M, Amburn E P, Austin J D, Cromartie R, Geselowitz A, Greer T, Ter Haar Romeny B, Zimmerman J B and Zuiderveld K. 1987. Adaptive histogram equalization and its variations. Computer Vision, Graphics, and Image Processing, 39(3): 355-368 [DOI: 10.1016/S0734-189X(87)80186-X]

Rahman Z, Jobson D J and Woodell G A. 1996. Multi-scale retinex for color image enhancement//Proceedings of the 3rd IEEE International Conference on Image Processing. Lausanne, Switzerland: IEEE: 1003-1006 [DOI: 10.1109/ICIP.1996.560995]

Redmon J and Farhadi A. 2018. YOLOv3: an incremental improvement [EB/OL]. [2024-12-14]. https://arxiv.org/pdf/1804.02767.pdf

Rombach R, Blattmann A, Lorenz D, Esser P and Ommer B. 2022. High-resolution image synthesis with latent diffusion models//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 10674-10685 [DOI: 10.1109/CVPR52688.2022.01042]

Song Y, Sohl-Dickstein J, Kingma D P, Kumar M, Ermon S and Poole B. 2021. Score-based generative modeling through stochastic differential equations [EB/OL]. [2024-12-14]. https://arxiv.org/pdf/2011.13456.pdf

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 Y G, Zhu Z P, Li J, Wang Z P, Chang Q and He B. 2023. Feature enhancement method for multimodal nighttime pedestrian detection//Proceedings of the 7th International Conference on Electrical, Mechanical and Computer Engineering (ICEMCE). Xi'an, China: IEEE: 1028-1032 [DOI: 10.1109/ICEMCE60359.2023.10491047]

Wei C, Wang W J, Yang W H and Liu J Y. 2018. Deep retinex decomposition for low-light enhancement [EB/OL]. [2024-12-14]. https://arxiv.org/pdf/1808.04560.pdf

Wu A C, Zheng W S, Yu H X, Gong S G and Lai J H. 2017. RGB-infrared cross-modality person re-identification//Proceedings of

2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE: 5390-5399 [DOI: 10.1109/ICCV.2017.575]

Yan Z H, Zhou C B and Li X C. 2024. Survey on generative diffusion model. Computer Science, 51(1): 273-283 (闫志浩, 周长兵, 李小翠. 2024. 生成扩散模型研究综述. 计算机科学, 51(1): 273-283) [DOI: 10.11896/jsjkx.230300057]

Ye M, Shen J B, Lin G J, Xiang T, Shao L and Hoi S C H. 2022. Deep learning for person re-identification: a survey and outlook. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(6): 2872-2893 [DOI: 10.1109/TPAMI.2021.3054775]

Yu P L, Shi Q and Wang H. 2021. Infrared-to-visible image translation based on parallel generator network. Journal of Image and Graphics, 26(10): 2346-2356 (余佩伦, 施佺, 王晗. 2021. 并行生成网络的红外—可见光图像转换. 中国图象图形学报, 26(10): 2346-2356) [DOI: 10.11834/jig.200113]

Zhang H and Yan J. 2024. Low-light image enhancement guided by semantic segmentation and HSV color space. Journal of Image and Graphics, 29(4): 966-977 (张航, 颜佳. 2024. 语义分割和HSV色彩空间引导的低光照图像增强. 中国图象图形学报, 29(4): 966-977) [DOI: 10.11834/jig.230182]

Zhang L M, Rao A Y and Agrawala M. 2023. Adding conditional control to text-to-image diffusion models//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE: 3813-3824 [DOI: 10.1109/ICCV51070.2023.00355]

Zhao W L, Rao Y M, Liu Z Y, Liu B L, Zhou J and Lu J W. 2023. Unleashing text-to-image diffusion models for visual perception//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE: 5706-5716 [DOI: 10.1109/ICCV51070.2023.00527]

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

Zhu J Y, Park T, Isola P and Efros A A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks//Proceedings of 2017 IEEE International Conference on Computer Vision (CVPR). Venice, Italy: IEEE: 2242-2251 [DOI: 10.1109/ICCV.2017.244]

Zhu Y B, Wang Q W, Li C L, Tang J and Huang Z X. 2024. Visible-thermal multiple object tracking: large-scale video dataset and progressive fusion approach [EB/OL]. [2024-12-14]. https://arxiv.org/pdf/2408.00969. pdf

作者简介

沈羽翔,男,硕士研究生,主要研究方向为多模态遥感影像目标检测。E-mail: shenyuxiang@csu.edu.cn

陶超,通信作者,男,教授,主要研究方向为遥感数据视觉表征、多模态遥感视觉基础模型。

E-mail: kingtaochao@csu.edu.cn

郭鑫,男,硕士研究生,主要研究方向为高分遥感影像智能解译。E-mail: guoxincsu@csu.edu.cn

王昊,男,博士研究生,主要研究方向为高分遥感影像智能解译。E-mail: haowang7cc@gmail.com

胡柯彦,男,硕士研究生,主要研究方向为多模态遥感影像语义分割。E-mail: phycheor@gmail.com