Multi-Modal Feature Enhanced Pedestrian Detection in Nighttime

Yi Wei
Tongji University
Shanghai, China
ve159334759111@gmail.com

Yanmin Zhou*
Tongji University
Shanghai, China
yanmin.zhou@tongji.edu.cn

Zhongpan Zhu*
Tongji University
Shanghai, China
521bergsteiger@tongji.edu.cn

***Abstract*—Recently, methods for moving pedestrian detection have attracted considerable attention. But it has great challenges in feature extraction from low-quality visible-light images in nighttime. Infrared image, with its excellent sensitivity to the thermal information of objects, provides a more effective complementary. Thus, multimodal pedestrian detection methods often yield superior results. This paper proposes a novel pedestrian detection method utilizing adaptive feature enhancement. Our approach is adopted in the UNet backbone, incorporating a cross-modal attention-guided learning module with an adaptive structure extraction module at the shallow layers and a channel exchange module based on genetic algorithms at the deep layers. This feature enhancement module is fully embedded within the detection network. The effectiveness and robustness of our method are tested on the LLVIP dataset, which includes paired infrared and visible-light pedestrian images with annotations. The results show that our method significantly improves pedestrian detection accuracy and notably enhances the quality of pedestrian images in low-light road scenes.**

Keywords—pedestrian detection, cross-modality, genetic algorithm, feature enhance

I. Introduction

The automatic detection and localization of pedestrians are becoming increasingly crucial in research on computer vision and smart wearable devices. Pedestrian detection is widely applied in autonomous vehicles [1]and surveillance systems [2]. Pedestrians detection in real-world scenarios can be challenging due to various factors, including different lighting conditions, diversity in appearances and poses of pedestrian, occlusions, camouflage, and cluttered backgrounds. This paper mainly focuses on low-light conditions.

Generally, low-light scene images suffers from degraded image quality and blur objects. Many researchers use image enhancement methods to improve the visibility of degraded images and satisfy human visual perception. The first approach is to train an image enhancement network, and then the enhanced images are used as input to train the detection network. The second way is to cascade the enhance and detection networks in an end-to-end approach. This paper proposed a cross-modal fusion enhancement method based on the cascade method. The contributions of this paper are summarized as follows:

- We proposed a cross-modal attention guided enhance module that can efficiently fuse different modal attention weight and guide the enhance process of visible light image.
- We used an adaptive structure extract fusion module to enhance the most focused structural features in visible light image and filter out redundant structural features.
- We introduced a novel genetic algorithm based channel switch method to fuse channel features from different modals. It contains two stages: the gene encoding stage and the chromosome crossover mutation stage.

II. Related Works

Image enhancement and multi-modal fusion are two key factors that improve the performance of pedestrian detection in low-light scenes. In this section, we first present the work related to image enhancement, then dive into multi-modal fusion.

A. Low Light Image Enhancement

Low-light image enhancement methods can be divided into handcrafted features-based and learning-based methods. Among methods based on handcrafted features, Ueda and Suetake et al. [3] used histogram normalization technology to transform colors. This method combines coefficients in vector space while preserving the color properties of the original image. Long Ma and Tengyu Ma et al. [4] proposed a self-calibrated supervised light learning enhancement method based on Retinex [5] theory. As for the learning-based method, Park et al. [6] designed a deep reinforcement learning scheme based on the Markov decision process. Their agent network only uses high-quality standard images for distortion and restoration during the training stage.

B. Infra and Visble Light Image Fusion

Infrared and visible fusion methods can be divided into three categories: CNN-based, GAN-based, and Transformer-based. Zhang et al. proposed IFCNN [7], a general CNN-based image fusion framework. This method relies on two convolution blocks for feature extraction and reconstruction, and it adopts a simple fusion strategy, including element-level average, max pooling, and min pooling. In GAN-based fusion, Li et al. proposed AttentionFGAN [8]. The multi-scale attention mechanism is introduced into the fusion model, which enhances the model's ability to extract distinctive and important information. Among transformer-based methods, Jun Huang et al. proposed the PTET [9], which transfers the features of the source image through a token exchange strategy, removes redundant information, and gradually enhances the features on the fusion branch through cascade layers.

Compared with various enhancement methods and fusion methods, the model proposed in this paper applied both image enhancement and multi-modal fusion methods.

III. Methods

The multi-modal feature enhanced pedestrian detection model proposed in this paper is adapted based on UNet [10] and Yolov5s [11]. The framework of our model consists of two parts: the feature enhancement module and the pedestrian detection module. The detailed structure of the proposed module is shown in Figure 1.

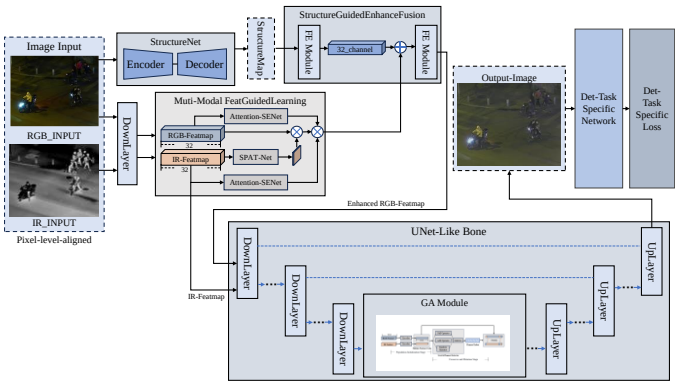

Fig. 1. The overview of our proposed model. ⊗ represents element-wise multiplication and ⊕ means summation operation.

A. Framework Overview

UNet is adopted as our feature enhancement backbone, as it contains a unique U-shaped network that can maintain spatial and contextual information when processing images and obtain improved performance in gradient backpropagation and feature localization. The model takes RGB and IR images as input; it obtains shallow features of them through the first DownLayer, then we use the cross-modal attention guide enhance module (CMAGE), calculate channel attention and spatial attention of the infrared mode as the guidance to enhance the visible light features. Then, we use an adaptive structure extractor (ASE)to adaptively extract visible light structural features and fuse them with visible light features through a structure-guided enhance fusion module (SGEF). In the bottleneck part, we get deep featmaps of two modals through Downlayers; the genetic algorithm [12] based channel switch module (GACS) takes them and gets the crossover mutation fusion results. Finally, we get the enhanced visible light image through Uplayers. The total enhance process can be described as:

$$I_e = U(GACS(Fd_{ir}, ASEF(I_{rgb}, CMAGE(Fs_{ir}, Fs_{rgb})))) \quad (1)$$

Where I_e denotes the enhanced image, U denotes UNet process, Fs_{ir} and Fs_{rgb} denote the shallow feature map of infrared and visible light modal, I_{rgb} and Fd_{ir} represent the visible light image and deep feature map of infrared.

B. Units

Cross-Modal Attention Guided Enhance Module

The CMAGE proposed in this paper is shown in Figure 2. Infrared sources and visible light sources have complementary characteristics. The infrared modal can provide temperature information, while the visible light modal can provide color and texture information. When pixels are aligned, feature maps in different modalities display different information about the same targets. Therefore, the GMAGE designed in this paper aims to achieve the cross-modal complementary adaptive attention guidance in channel and space dimensions.

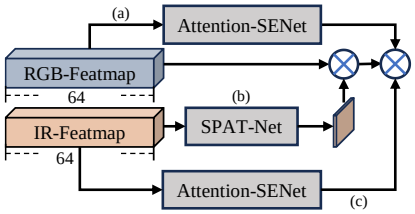

Fig. 2. The structure of CMAGE module

CMAGE module process can be described as below, where Wch_{rgb} and Wch_{ir} denote the channel attention of visible light and infrared. Wsp_{ir} represents the spatial attention of infrared feature. ⊗ denotes element-wise multiplication.

$$F_{out} = (Wsp_{ir} \otimes F_{rgb}) \otimes Wch_{rgb}) \otimes Wch_{ir} \quad (2)$$

This module includes three branches. In branch (a), we use the channel attention mechanism to perform a nonlinear transformation on the global semantic information in visible light modality. The attention of color and texture information is stacked on the visible light featmap. In branch (b), we apply the spatial attention mechanism to the infrared featmap and obtain the spatial attention which represents the thermal information attention in infrared space. Subsequently, we perform element-by-element multiplication of this weight with the visible light feature map, using the spatial feature attention of infrared thermal information to guide the feature extraction process of visible light at the corresponding location. In branch (c), we use the channel attention mechanism with the same structure as branch (a) to extract the global feature weight of infrared in the channel dimension and guide the learning process of visible light features.

Adaptive Structure Extract Fusion Module

ASEF module consists of two parts: the adaptive structure extraction module (ASE) and the structural information guided enhanced fusion (SGEF) module. The process of ASEF can be described as below, where I_{rgb} denotes visible light image, F_{rgb} denotes the visible light feature map after CMAGE.

$$F_{out} = SGEF(ASE(I_{rgb}), F_{rgb}) \quad (3)$$

The ASE module is shown in Figure 3. The objective of this module, on the one hand, is to balance the possible excessive attenuation of color and structural information under the cross-modal attention guidance of infrared data. It ensures the integrity of critical structural information and optimizes the overall performance. On the other hand, the adaptive structure extraction module has dynamic attention capabilities that can independently focus on the structural features required for the detection task and enhance them through detection stage's loss calculation, gradient backpropagation, and parameter update. At

the same time, this module can effectively erase the noise of redundant structures and color features from visible light source input.

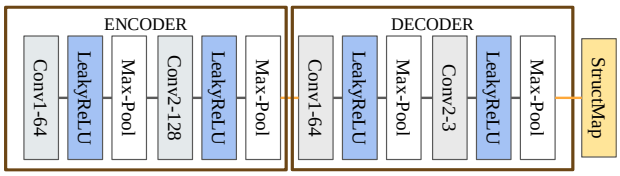

Fig. 3. The ASE module structure.

The SGEF proposed in this paper is shown in Figure 4. We take the structure map as input and employ channel embedding method, then concatenate it with the visible light featmap processed by CMAGE module. Finally, the channel dimension structural feature guided fusion is deployed.

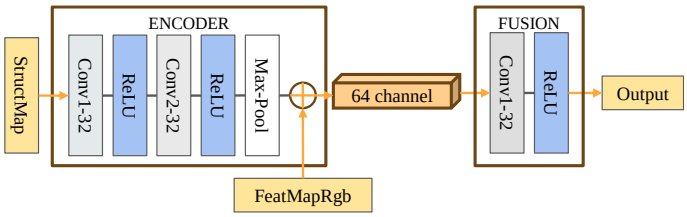

Fig. 4. The SGEF module structure.

Genetic Algorithm based Channel Switch

In the bottleneck part of UNet, this paper proposed a channel exchange module based on genetic algorithms, as shown in Figure 5. This module contains two stages: the gene encoding stage and the chromosome crossover mutation stage; Our GACS process can be described as:

$$Ind_{Select} = Selector(Code_{ir}, Code_{rgb}) \quad (4)$$

$$F_{out} = Ex(Ind_{Select}, Code_{ir} \otimes F_{ir}, Code_{rgb} \otimes F_{rgb}) \quad (5)$$

Where Ex denotes the channel exchange method; Ind_{Select} represents the selected channel index to switch. $\otimes$ denotes element-wise multiplication. $Code_{ir}$ and $Code_{rgb}$ represent the infrared and visible light genetic code after the calculation of genetic encoder.

The feature maps of the two modalities encode different feature information in different channels. The inputs from different modals have feature isolation in the same channel. Therefore, in the genetic chromosome encoding stage, this paper simulates the chromosome encoding in the natural genetic process. This way, the feature information in different channels is effectively compressed and compiled. To update the encoder parameters during the training process, we use the element-wise method to multiply visible light code with visible light featmap and infra code with infrared featmap. This process not only increases the feature independence of two modalities, but the difference between them. It allows us to adjust the encoder parameters at a more fine-grained level.

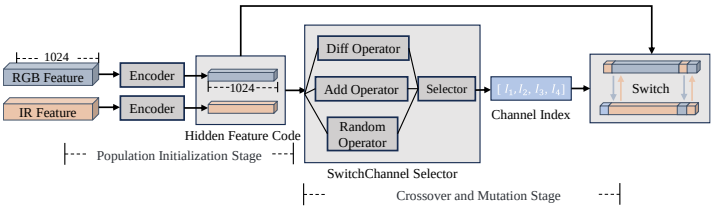

Fig. 5. The structure of the GACS module.

In the chromosome crossover mutation stage, this study employed addition (add) and difference (diff) operations to calculate the add-diff vector, which reflects the commonality and difference of each channel feature in different modes. Next, this vector is fed into the channel selector. The channel selector uses a random operator to determine the exchange channel index. Then it switches and fuses features from two modals in channel dimension, allowing the model to dynamically reconfigure the feature map's information flow. Finally, after channel exchange, the visible light feature map is passed through the Upsampling operation of UNet and the final feature-enhanced visible light image is produced.

IV. Experiment

Experiments are conducted on the LLVIP [13] dataset to compare the overall performance of the model proposed in this paper with the baseline model in terms of object detection precision and enhancement effect.

A. Experiment setup

We use the recently released multi-modal pedestrian detection dataset LLVIP for training and evaluation. LLVIP includes low-light IR-RGB pixel aligned image pairs captured at 26 different positions from a camera perspective, of which 12025 pairs are used for training and 3463 pairs for testing. This dataset only contains pedestrians in one class.

Our model is adapted based on UNet and Yolov5s. It is trained on a single NVIDIA Tesla V100 GPU. The Yolov5s and UNet used in the experiment have not been pretrained with any data other than LLVIP. We use a resolution of 640*512 as input. The optimizer employed is SGD with a learning rate of lr0= 0.01, lrf= 0.01. The model is trained for 100 epochs with batch_size=4.

We employ a variety of multi-modal target detection models as baselines, including DIVFusion [14], SDNet [15], GAFF [16], Halfway Fusion [17], and CSSA [18]. For a fair comparison, the experimental setting for all the baselines is identical to ours. Besides, we also compared the enhancement effect of our method with SCI method.

B. Result

The detection performance of the proposed model and the baselines mentioned above are evaluated on LLVIP datasets. The results are shown in Figure 6 and Table 1. Compared with the CSAA method, our proposed method outperforms by 1.8% on AP50 and 4.4% on mAP. Our method outperforms Halfway Fusion by 4.7% on AP50 and 8.5% on mAP. Both our method and CSAA have the idea of featmap exchange. Still, this method is different from the CENet method because it uses an adaptive selection method based on the gene encoding weight instead of

using the weight of the BN layer and setting a fixed threshold for exchange.

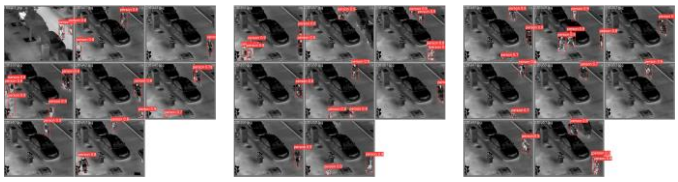

Fig. 6. Example of a figure caption. (figure caption)

TABLE I. PEDESTRIAN DETECTION RESULT COMPARISON

Method	AP50	mAP
CSAA	94.3	59.2
SDNet	86.6	50.8
DIVFusion	89.8	52.0
GAFF	94.0	55.8
Halfway Fusion	91.4	55.1
Ours	**96.1**	**63.6**

We also compared the effect of image enhancement with other low-light enhance methods. In the inference stage, we take the enhanced image after UNet upsampling for visualization and compare it with the inference result of the current low-light enhancement SOTA method SCI; The enhanced images of our method are shown in Figure 6. The compare result is shown in Table 2. We can see that our method has achieved great enhancement results in the LLVIP evaluation set, and PNSR SSIM indicates that our method is efficient and improves low-light image quality.

TABLE II. ENHANCE EFFECT COMPARISON

Metrics	**SSIM↑**	**PSNR↑**	**EME↑**	**LOE↓**
SCI	27.775	0.4086	0.3307	2.8952
Ours	**28.383**	**0.7416**	**0.5874**	**2.9612**

Fig. 7. The visualization of our enhance method. (The line above is the original image, The line below is the enhanced image)

V. CONCLUSION

We proposed a channel exchange feature enhancement method based on the genetic a-lgorithm, it integrates infrared thermal information and visible light structural texture information by deploying a cross-modal attention mechanism, an adaptive structural feature extraction, and a channel switch fusion method. By simulating the genetic encoding and cross-mutation operations in genetic algorithms, we evaluate the importance of different channel features of the two modalities and perform feature cross-fusion. Training and testing were conducted on the LLVIP dataset. The method proposed in this paper not only showed excellent accuracy and performance in the detection task but also achieved good results in low-light image enhancement.

REFERENCES

[1] Parekh, Darsh, et al. "A review on autonomous vehicles: Progress, methods and challenges." Electronics 11.14 (2022): 2162.

[2] Valera, Maria, and Sergio A. Velastin. "Intelligent distributed surveillance systems: a review." IEE Proceedings-Vision, Image and Signal Processing 152.2 (2005): 192-204.

[3] Ueda, Yoshiaki, and Noriaki Suetake. "Hue-preserving color image enhancement on a vector space of convex combination coefficients." 2019 IEEE International Conference on Image Processing (ICIP). IEEE, 2019.

[4] Ma, Long, et al. "Toward fast, flexible, and robust low-light image enhancement." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

[5] Land, Edwin H., and John J. McCann. "Lightness and retinex theory." Josa 61.1 (1971): 1-11.

[6] Park, Jongchan, et al. "Distort-and-recover: Color enhancement using deep reinforcement learning." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[7] Zhang, Yu, et al. "IFCNN: A general image fusion framework based on convolutional neural network." Information Fusion 54 (2020): 99-118.

[8] Li, Jing, et al. "AttentionFGAN: Infrared and visible image fusion using attention-based generative adversarial networks." IEEE Transactions on Multimedia 23 (2020): 1383-1396.

[9] Huang, Jun, et al. "PTET: A progressive token exchanging transformer for infrared and visible image fusion." Image and Vision Computing (2024): 104957.

[10] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. Springer International Publishing, 2015.

[11] Redmon, Joseph, et al. "You only look once: Unified, real-time object detection." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[12] Harvey, Inman. "The microbial genetic algorithm." Advances in Artificial Life. Darwin Meets von Neumann: 10th European Conference, ECAL 2009, Budapest, Hungary, September 13-16, 2009, Revised Selected Papers, Part II 10. Springer Berlin Heidelberg, 2011.

[13] Jia, Xinyu, et al. "LLVIP: A visible-infrared paired dataset for low-light vision." Proceedings of the IEEE/CVF international conference on computer vision. 2021.

[14] Tang, Linfeng, et al. "DIVFusion: Darkness-free infrared and visible image fusion." Information Fusion 91 (2023): 477-493.

[15] Zhang, Hao, and Jiayi Ma. "SDNet: A versatile squeeze-and-decomposition network for real-time image fusion." International Journal of Computer Vision 129.10 (2021): 2761-2785.

[16] Zhang, Heng, et al. "Guided attentive feature fusion for multispectral pedestrian detection." Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2021.

[17] Liu, Jingjing, et al. "Multispectral deep neural networks for pedestrian detection." arXiv preprint arXiv:1611.02644 (2016).

[18] Cao, Yue, et al. "Multimodal object detection by channel switching and spatial attention." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.