Jiang *et al*. *Intell. Robot.* **2025**, *5*(1), 34-49
DOI: 10.20517/ir.2025.03

Intelligence & Robotics

Research Article

Open Access

CMMF-Net: a generative network based on CLIP-guided multi-modal feature fusion for thermal infrared image colorization

Qian Jiang[1,2], Tao Zhou[1,2], Youwei He[1,2], Wenjun Ma[1,2], Jingyu Hou[3], Ahmad Shahrizan Abdul Ghani[4], Shengfa Miao[1,2], Xin Jin[1,2]

[1]School of Software, Yunnan University, Kunming 650000, Yunnan, China.
[2]Engineering Research Center of Cyberspace, Yunnan University, Kunming 650000, Yunnan, China.
[3]School of Information Technology, Deakin College, Vic 3125, Australia.
[4]Manufacturing & Mechatronic Engineering Technology, Universiti Malaysia Pahang, Pekan, Malaysia.

Correspondence to: Prof. Shengfa Miao, School of Software, Yunnan University, East Outer Ring South Road, Chenggong District, Kunming 650000, Yunnan, China. E-mail: miaosf@ynu.edu.cn

How to cite this article: Jiang, Q., Zhou, T., He, Y., Ma, W., Hou, J., Abdul Ghan, A.S., Miao, S., Jin, X. CMMF-Net: a generative network based on CLIP-guided multi-modal feature fusion for thermal infrared image colorization. *Intell. Robot.* **2025**, *5*(1), 34-49. http://dx.doi.org/10.20517/ir.2025.03

Received: 12 Sep 2024 **First Decision:** 30 Nov 2024 **Revised:** 20 Dec 2024 **Accepted:** 31 Dec 2024 **Published:** 16 Jan 2025

Academic Editor: Simon Yang, Guang Chen **Copy Editor:** Pei-Yun Wang **Production Editor:** Pei-Yun Wang

Abstract

Thermal infrared (TIR) images remain unaffected by variations in light and atmospheric conditions, which makes them extensively utilized in diverse nocturnal traffic scenarios. However, challenges pertaining to low contrast and absence of chromatic information persist. The technique of image colorization emerges as a pivotal solution aimed at ameliorating the fidelity of TIR images. This enhancement is conducive to facilitating human interpretation and downstream analytical tasks. Because of the blurred and intricate features of TIR images, extracting and processing their feature information accurately through image-based approaches alone becomes challenging for networks. Hence, we propose a multi-modal model that integrates text features from TIR images with image features to jointly perform TIR image colorization. A vision transformer (ViT) model will be employed to extract features from the original TIR images. Concurrently, we manually observe and summarize the textual descriptions of the images, and then input these descriptions into a pretrained contrastive language-image pretraining (CLIP) model to capture text-based features. These two sets of features will then be fed into a cross-modal interaction (CI) module to establish the relationship between text and image. Subsequently, the text-enhanced image features will be processed through a U-Net network to generate the final colorized images. Additionally, we utilize a comprehensive loss function to ensure the network's ability to generate high-quality colorized images. The effectiveness of the methodology put forward in this study is evaluated using the KAIST datasets. The experimental results vividly showcase the superior performance of our CMMF-Net method in comparison to other methodologies for the task of TIR image colorization.

Keywords: Thermal infrared image colorization, vision and language, transformer

1. INTRODUCTION

Infrared imaging differentiates various objects from their surroundings based on emitted radiation. This technique remains effective even under challenging weather conditions, including precipitation and fog. Consequently, thermal infrared (TIR) imaging finds extensive utility across military operations, surveillance systems, vehicle imaging, nocturnal traffic management, and various other applications[1–3]. Nonetheless, TIR images exhibit a deficiency in color depth and intricate details inherent in RGB images, posing challenges for downstream tasks such as object recognition[4,5], semantic segmentation[6], obstacle avoidance for mobile robots in dynamic environments[7], structural health monitoring of bridges[8], the detection of cracks and damage[9] and related domains[10–12]. By applying colorization to infrared images of building facades, roofs, and windows, personnel can more clearly identify heat leaks, insulation issues, and thermal bridges within the structures. Furthermore, in the realm of security inspections, infrared image colorization significantly enhances the effectiveness of surveillance systems. Thus, the process of colorizing TIR images is increasingly vital to enhance their usability and overall quality.

Image colorization stands as a pivotal method for enhancing TIR imagery, effectively improving its overall quality. It aligns TIR images more closely with human visual expectations while also restoring their grayscale intricacies. Current grayscale-image-based colorization techniques have yielded remarkable outcomes[13–16]. However, grayscale image-based colorization typically involves estimating only colors, whereas TIR image colorization demands the reconstruction of both texture and color information. Additionally, the substantial disparity between the two types of images complicates the implementation of colorization algorithms. Consequently, a key challenge is to design mechanisms that enable our network to efficiently capture nuanced feature details from the input images and, under the guidance of the loss function, generate high-quality, color-rich images.

Over the past few years, The primary focus of traditional methods is to restore the color of infrared images. However, the resulting colors often exhibit significant discrepancies from real-life counterparts[14,17]. In recent times, deep learning has significantly advanced various computer vision tasks. Notably, the effective utilization of convolutional neural networks (CNNs) and generative adversarial networks (GANs) has propelled TIR image colorization techniques to considerable success. Nonetheless, owing to the inherent limitations of TIR imagery, the outcomes still exhibit disparities compared to real RGB images.

This paper proposes CMMF-Net (a generative network based on CLIP-guided multi-modal feature fusion for TIR image colorization) for the colorization of TIR images, which is capable of efficiently extracting and processing image feature information by integrating textual descriptions. The introduction of text helps to mitigate the inherent low quality and blurry features of TIR images, leading to a noticeable improvement in the quality of the final colorized image results. We utilize a multi-modal network framework to address the issue of conventional colorization methods based solely on single TIR images failing to understand and process the complex feature information inherent in TIR images effectively and accurately.The network is composed of three parts. Text_Encoder is responsible for encoding textual information to obtain textual tokens, while Image_Encoder is responsible for extracting image information to obtain image tokens. Next, the obtained textual and image tokens are inputted into the cross-modal interaction (CI) module to obtain image features enhanced with textual information. Finally, combining with the U-Net network to decode this feature, the final colorized image is obtained. Our loss function includes content loss, perceptual loss, total variation loss and structural similarity index (SSIM) loss. Overall, we summarize the contributions below:

- A novel multi-modal feature fusion network is carefully proposed for the task of visible light image restoration from TIR image.
- Text_Encoder parameters of contrastive language-image pretraining (CLIP) model are migrated in this model, and the text and image feature information are processed by combining vision transformer (ViT) encoder module and CI module.
- The experimental results demonstrate that our method achieves leading performance in TIR image colorization.

2. RELATED WORK

2.1. Visible image colorization

Image colorization involves transforming a monochromatic TIR image into a colorized image with three channels. Traditional grayscale image colorization techniques often necessitate user participation. For instance, the Scribble-based method relies on prior color annotations provided by users, which are then extended to the entire grayscale image[18]. Another approach, the Example-based method, utilizes color data from reference images to assist in the colorization process[19].

In recent years, the field of image colorization has witnessed a significant shift towards deep learning-based approaches, which have demonstrated superior performance compared to traditional methods. Larsson *et al.* introduced a fully automated technique utilizing CNNs[20]. Building upon this, Zhao *et al.* integrated image segmentation information to further enhance colorization accuracy[21]. The integration of generative adversarial network (GAN) models[22] has also become prevalent in numerous colorization tasks[14,23]. Moreover, recent advancements in denoising diffusion probabilistic models (DDPM)[24,25] have introduced a novel colorization approach based on visible image diffusion models. This method utilizes visible images as conditions for diffusion models, resulting in color images through denoising processes. Impressively, this technique has shown promising results in the realm of visible image colorization[26]. Collectively, these advancements underscore the transformative potential of deep learning techniques in revolutionizing image colorization processes.

2.2. TIR image colorization

With the continuous development of deep learning, various popular network frameworks have been applied in TIR image colorization tasks[27–31]. However, due to the substantial differences in data distribution between TIR and visible images, alongside the abundance of TIR datasets, CNN architectures often struggle to adapt effectively to TIR image colorization tasks. TIR stands out as the pioneering CNN-based colorization network, leveraging the U-Net architecture for TIR image colorization[27]. While TIR boasts lightweight design, its colorization performance remains less than satisfactory.

In contrast, the inherent capability of GAN-based architectures in image generation has led to the predominance of GAN-based networks in image colorization. Leveraging the advantages of GAN structures, these networks demonstrate enhanced performance and effectiveness in generating realistic colorized images. Advancements in TIR image colorization have predominantly been driven by GAN-based approaches, notably exemplified by Pix2Pix[23], which, while enhancing image details and colors to some extent, still exhibits a degree of unnaturalness compared to true images. Subsequent refinements, such as those introduced by Kuang *et al*., have significantly bolstered colorization quality through the integration of composite loss functions[28]. Innovations such as the ToDayGAN-based network proposed by Lou *et al.* have further pushed the boundaries by effectively translating nighttime TIR images into visually satisfying daytime colorized counterparts[32]. Additionally, efforts by Liao *et al.* with the mixed-skipping (MS) U-Net architecture have notably improved colorization quality through advanced attention mechanisms[30]. He *et al.* proposed a large kernel

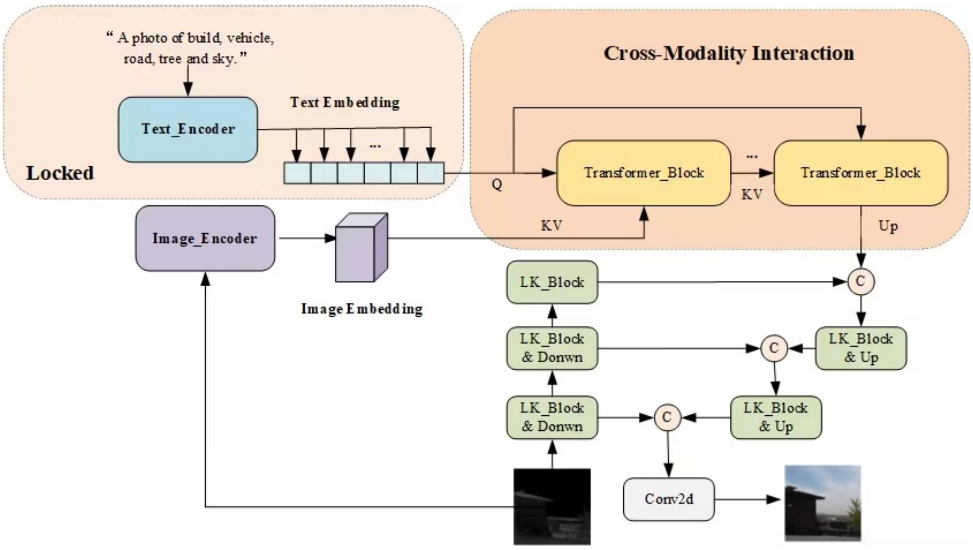

Figure 1. The overall framework. Including a ViT Image_Encoder module, a CLIP Text_Encoder module, a cross-modality alignment module and a U-net module. CMMF-Net takes image-sentence pairs as input, and outputs the colorized image. ViT: Vision transformer; CLIP: contrastive language-image pretraining; CMMF-Net: a generative network based on clip-guided multi-modal feature fusion for thermal infrared image colorization.

network and a two-branch structure network incorporating a Transformer module to fully process TIR image features[31]. However, due to inherent defects in the image, the color effect remains unsatisfactory. Recently, several networks have been proposed for the colorization of TIR videos, including Recycle-GAN[33] and unpaired infrared-to-visible video translation (I2VGAN)[34]. These approaches primarily focus on enhancing the realism of video content.

2.3. CLIP net

In tasks involving vision and language, a profound grasp of semantics is pivotal for achieving enhanced collaborative representation. In the realm of comprehending visual content, diverse backbone models[35,36] have emerged, aimed at pure visual comprehension. These models have convincingly demonstrated their efficacy across substantial datasets[37]. In recent cross-modality research, numerous methods[38–40] have been introduced. These approaches primarily emphasize understanding the intricate connection between visual and textual elements across diverse modalities. CLIP[41] is a type of neural network trained on various pairs of images and texts. It can be directed using natural language to predict the most relevant textual snippet given a specific image, without the need for direct optimization for the task, similar to the zero-shot capability of GPT-2 and GPT-3. CLIP consists of two models: text encoder and image encoder. The text encoder is utilized to extract features from text and can employ commonly used text transformer models from NLP. On the other hand, the image encoder is employed to extract features from images, and it can use conventional CNN models or ViTs. In this paper, we will utilize its pretrained text-encoder module to extract text information.

3. PROPOSED METHOD

3.1. Overview

Figure 1 presents the architecture of the model. The proposed method integrates both image feature information and textual information, enabling a more comprehensive and accurate understanding and processing of the relevant features of TIR images. We utilize the pretrained CLIP model to capture and process text-based feature information. In this approach, the Text_Encoder module refers to the one used in the original CLIP method. Meanwhile, we utilize the Encoder module of the ViT model[42] to extract image feature information from TIR images. The ViT model is an image processing model based on the Transformer architecture,

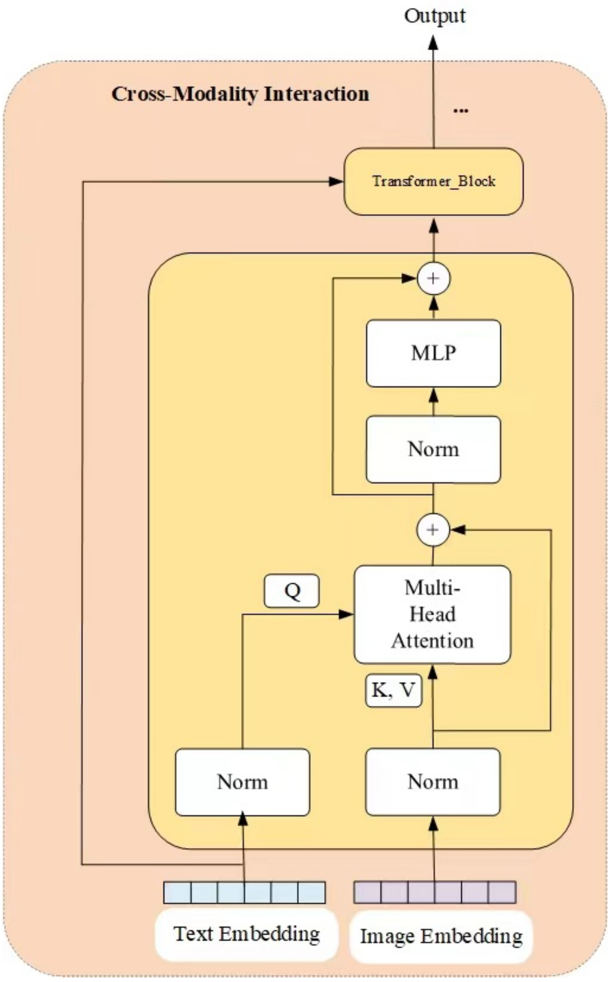

Figure 2. The CI module. The text feature is Q, and the image feature or the current layer output feature is K and V for multi-head attention operation, and the corresponding relationship is established. CI: Cross-modal interaction.

which has achieved significant success in the field of image processing. It can effectively extract feature information from images and encode it into vector representations. Next, the text and image feature information obtained through the CI module is processed. Through a multi-head attention mechanism, relationships between text and images are established, ultimately resulting in enhanced image feature representations based on text. The cross-modality design presented in this paper consists of two transformer blocks. These blocks are implemented based on the ViT. Unlike the conventional ViT, we integrate the generation of the attention mechanism's Key (K), Query (Q), and Value (V) with the modality interaction between images and text. Specifically, we generate K from the text and Q and V from image embeddings. Additionally, we employ a multi-head attention mechanism to enable the model to capture more global information. The detailed processing flow is illustrated in Figure 2. Next, drawing from the successful practice of the LK_U-Net module in the first large kernel U-Net and attention_U-Net-transformer (ViT-Based)-based generative adversarial network (LKAT-GAN) method[31] for TIR image colorization task, we employ this U-Net module to process and decode the text-image features obtained from the previous stage to generate the final colorized image. To achieve higher-quality results, The total loss comprises content loss, perceptual loss, total variation loss, and SSIM loss. Details will be provided in the following sections.

3.2. Text_Encoder and Image_Encoder

In the context of the large-scale dataset WebImageText[41] collected and curated by the OpenAI team, the emergence of the CLIP model signifies a significant breakthrough in the field of text-image prediction matching. The key innovation of the CLIP method lies in its contrastive learning framework, which matches natural language descriptions with images, prompting the model to learn a shared embedding space. In this space, relevant language descriptions and images are mapped to nearby points, while irrelevant language descriptions and images are pushed apart, enabling the model to better understand the semantic relationships between images and language.

Through pretraining on large-scale image-text pairs, the CLIP model has learned a universal cross-modal representation. This representation not only captures rich relationships between images and language but also exhibits strong generalization capabilities. In practice, this means that when faced with new tasks or categories, the CLIP model possesses powerful zero-shot learning abilities. Even in unseen scenarios, this model can perform inference and generalization because the universal representation it has learned demonstrates strong generalization properties.

In this paper, we designed a general module that introduces text guidance for image colorization. The text input for our method is based on the observation of typical object information in the dataset, and the text features are extracted using the CLIP Text Encoder model. The Text Encoder module of CLIP is essentially a basic Bert, fundamentally composed of Transformer modules, so the structures of Text Encoder and Image Encoder are quite similar. In CLIP, the Text_Encoder consists of 12 layers of Transformer Encoder. Due to its pretraining phase, where it has learned rich semantic information representations through large-scale datasets, by transferring the parameters of this module and combining them with the Image_Encoder module of the ViT model and the CI module, we extend these advantages to the TIR image colorization task. In our approach, the depth of the Image_Encoder module is set to nine layers. The Text_Encoder module, pretrained on massive datasets, provides a strong foundation, enabling our approach to accurately extract relevant features. These features capture the rich associations between text descriptions and image features, providing robust support for the colorization task.

3.3. CI

To better adapt the text features obtained by the transferred Text_Encoder module and the image features obtained by the Image_Encoder module to TIR image colorization, we designed the CI module to process the TIR image feature information and corresponding text information. The specific structure is depicted in Figure 2.

The overall structure is similar to the Transformer module, where we use the token sequence representation obtained by the Text_Encoder module as Query, and the image feature information obtained by the Image_Encoder module as Key and Value, which are jointly inputted into the CI module for multi-head attention operations. Through attention mechanisms, the relationships between text and images are established, and then the text-enhanced image representation is obtained through the multi-layer perceptron (MLP) layer output. The output of the current layer is then fed along with the text token into the next layer to repeat the above operation. Finally, the OutputProj layer is used to adjust the feature size and dimensions, outputting the image features with text feature information. Since the input image features and text features already contain positional information, no Embedding operation is performed in the CI module; instead, the features are directly inputted into our Transformer_Block.

For the Text_Encoder module with fixed parameters, the input text is mapped to a predefined feature space, which mainly contains categorical classification information of text tokens. In the CI module, the fixed text feature information can effectively highlight and emphasize image features, enhancing the model's understanding. As for the Image_Encoder module, with the involvement of text features, through network training and optimization, it can better retrieve and process image features. The enhanced image features ultimately ameliorate the problem of inaccuracies in directly extracting image features from TIR images, resulting in significant enhancements in both image details and color information in the final colorized images.

3.4. U-Net

We introduced a U-Net network module as a critical component in the colorization task. This module plays a vital role in the network, effectively integrating text and image information, and is crucial in generating colorized images. The design inspiration for the U-Net module comes from the innovation in the LKAT-GAN[31] work, particularly the U-Net module with large kernel convolutions. This structure better handles feature extraction and fusion in colorization tasks. Specifically, during the downsampling process, we employ a network with large kernel convolutions to effectively extract important feature information from the original TIR images while ensuring that these features maintain the same size as the features with text information. Next, the downsampled features and the comprehensive output features of text-image are concatenated for final upsampling. This approach fully utilizes the information from both types of features while maintaining coherence and consistency between them. Similar to most U-Net networks, skip connections and concatenation are applied between the downsampling and upsampling modules, preserving more detailed information and enhancing the performance of network and robustness. Finally, through a convolutional layer with a 3 × 3 kernel, the number of channels is adjusted to 3, outputting the final colorized result. This step ensures that the generated color images have good visual effects and quality. In summary, the U-Net network module in our approach is trainable, meaning it has the capability to adjust and optimize parameters, allowing for flexible tuning based on tasks and data to meet the demands of different scenarios, thereby enhancing the quality and accuracy of colorization results. Additionally, by combining the downsampling and upsampling processes, this module can effectively fuse text and image information without losing feature details. During the downsampling stage, the network with large kernel convolutions can better capture global features and structural information of the image. In the upsampling stage, skip connections and concatenation operations preserve detailed information and seamlessly integrate text and image information. This efficient fusion mechanism provides reliable support for the colorization task, achieving more accurate and clearer colorization effects while maintaining image quality.

3.5. Loss function

The loss function of our model comprises four components: Content loss $L_{content}$, Perceptual loss $L_{perceptual}$, Total Variant loss L_{tv}, and SSIM loss L_{ssim}.

Content loss can ensure the minimization of differences in chrominance and luminance between the generated colorized image and ground truth image. $L_{content}$ is defined as follows:

$$L_{content} = \mathbb{E}_{x,y}[||y - G(x,z)||_1], \tag{1}$$

where y represents he actual colorized image, x indicates the input infrared image, z denotes the text vector, and G(x,z) stands for our colorization model.

Perceptual loss can regulate the perceptual details between the colorized image and the ground truth image. Based on the experience of existing research, the VGG-19[43] network pretrained on the Image-Net dataset is utilized as the perceptual feature extractor. $L_{perceptual}$ is defined as follows:

$$L_{perceptua} = \mathbb{E}_{x,y}[||(\Phi_i(G(x,z) - \Phi_i(y)||_1], \tag{2}$$

where Φ_i represents the i-th layer features extracted by the pretrained network.

Total variant loss can significantly enhance the spatial smoothness of the images generated by the model while simultaneously reducing image noise. First, the pixel differences of the generated image along the height and width dimensions are calculated. Then, the total variation value along the width direction is obtained. Finally, the total variation values are summed and normalized by dividing by the total number of pixels in the image, yielding the average total variation. L_{tv} is formulated as follows:

$$L_{tv} = \frac{1}{WH} \sum |\nabla_x G(\widetilde{y})| + |\nabla_y G(\widetilde{x})|, \tag{3}$$

where W and H denote the height and width of the image, respectively.

SSIM loss can minimize the difference between generated images and ground truth, we introduce SSIM loss which can measure the similarity between two images. L_{ssim} is defined as follows:

$$L_{\text{ssim}} = 1 - \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \tag{4}$$

where μ_x and μ_y represent the mean luminance values of image x and image y, C is a small constant introduced to prevent numerical instability, and σ_x and σ_y represent the contrast values of image x and image y, σ_{xy} represents the structural correlation between the two images.

Thus, the overall loss function for our network is formulated as follows:

$$\begin{aligned} L_{total} = \lambda_{content} L_{content} + \lambda_{perceptua} L_{perceptua} \\ + \lambda_{tv} L_{tv} + \lambda_{ssim} L_{ssim}, \end{aligned} \tag{5}$$

where $\lambda_{content}$, λ_{adv}, $\lambda_{perceptua}$, and λ_{tv} are parameters that control the weight of each loss function.

4. EXPERIMENTS

4.1. Dataset and implementation details

4.1.1. KAIST dataset

In image colorization tasks, accurate pixel-to-pixel correspondence between infrared and visible images is essential. Consequently, this paper utilizes the KAIST multi-spectral pedestrian dataset[44] for our experimentation. The training set comprised 26,387 pairs of thermal and RGB images captured during the daytime, with an additional 23,925 daytime images used for evaluation. In the experiments, each image was randomly cropped to a resolution of 256 × 256.

4.1.2. IRVI dataset

The IRVI (dataset for infrared-to-visible video translation) dataset[34] includes two distinct scenarios: traffic and monitoring scenarios. Due to the relatively uniform content of the images in this dataset, we utilized it as an auxiliary dataset to further validate the model's performance.

4.1.3. Implementation details

We implemented the model using PyTorch, with the training conducted on an NVIDIA 3090 GPU environment utilizing the PyTorch framework. The parameters of Text_Encoder module are derived from the original paper[41]. The weight coefficient of the loss function $\lambda_{content}$, $\lambda_{perceptua}$, λ_{tv} and λ_{ssim} are set to 1, 1, 1 and 1. For the comparative methods, we employed two evaluation metrics: SSIM and peak signal-to-noise ratio (PSNR).

4.2. Experiments on KAIST dataset

We select six comparative TIR image colorization methods: Pix2pix[23], thermal infrared image colorization using mixed-skipping UNet and generative adversarial network (MUGAN)[30], LKAT-GAN[31], PealGAN[32],

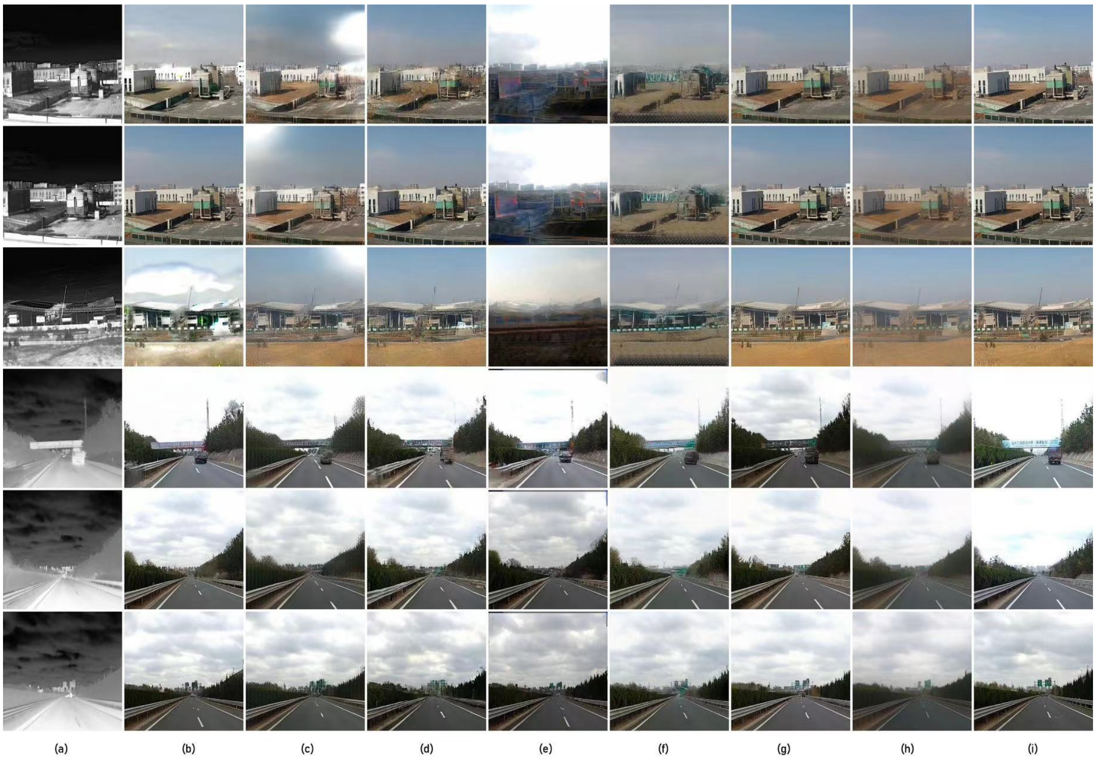

Figure 3. Colorized images using different image colorization methods on IRVI. (A) TIR images; (B) I2VGAN [34]; (C) TICC-GAN [28]; (D) SCGAN [29]; (E) PealGAN [32]; (F) MUGAN [30]; (G) LKAT-GAN [31]; (H) CMMF-Net; (I) True RGB images. The original infrared images (A) and RGB images (J) were obtained from https://github.com/BIT-DA/I2V-GAN, while the other images were reproduced using methods from other papers and generated with our laboratory's equipment. The specific methods can be found in the corresponding references listed in the bibliography. IRVI: Dataset for infrared-to-visible video translation; TIR: thermal infrared; I2VGAN: unpaired infrared-to-visible video translation; TICC-GAN: thermal infrared colorization via conditional generative adversarial network; SCGAN: saliency map-guided colorization with generative adversarial network; MUGAN: thermal infrared image colorization using xixed-skipping UNet and generative adversarial network; LKAT-GAN: a GAN for thermal infrared image colorization based on large kernel and attentionUNet-transformer; CMMF-Net: a generative network based on clip-guided multi-modal feature fusion for thermal infrared image colorization.

CycleGAN [33], saliency map-guided colorization with generative adversarial network (SCGAN) [29], and thermal infrared colorization via conditional generative adversarial network (TIC-CGAN) [28]. The best performance metrics of the aforementioned methods are shown in Table 1. To illustrate the objective evaluation of each model, line graphs showing the scores from 16 to 20 epochs are provided in Figure 4. Across all epochs, the two metrics of our model consistently surpass those of other methods, clearly demonstrating the potential and performance of our method. Moreover, the outstanding SSIM values provide additional evidence that the results produced by our method are highly consistent with the ground truth. In quantitative comparisons, CMMF Net exhibits the best overall performance, excelling in both detail reconstruction and color. The qualitative evaluation metrics of the aforementioned methods are illustrated in Figure 5. Overall, SCGAN exhibits the weakest performance, with inadequate color and detail reconstruction. Similarly, TIC-CGAN demonstrates shortcomings in reconstructing fine image details. While PealGAN, as an unsupervised model, excels in handling unpaired datasets, its performance on paired datasets is inferior to that of supervised models. MUGAN, although yielding better results than TIC-CGAN and SCGAN, still exhibits unsatisfactory detail reconstruction, with minor noise present in certain cases. LKAT-GAN produces more realistic colors and richer details compared to the other methods, but the object recognition is not clear and accurate enough, and the image noise is significant. In terms of image object details, our proposed method is significantly better than other comparison methods, and the colorized images generated by our model are smoother and less noisy. However, it is worth noting that the results generated by our method, as a whole, do not appear as bright compared to other methods.

Table 1. Average results on the KAIST test dataset

Method	SSIM	PSNR
Pix2pix [23]	0.50	15.10
MUGAN [30]	0.54	15.55
LKAT-GAN [31]	0.54	15.79
PealGAN [32]	0.43	14.83
CycleGAN [33]	0.30	12.17
SCGAN [29]	0.53	15.48
TIC-CGAN [28]	0.54	15.52
CMMF-Net	**0.58**	**16.22**

KAIST: A multispectral pedestrian dataset, proposed by the Korea Advanced Institute of Science and Technology; SSIM: structural sim-ilarity index; PSNR: peak signal-to-noise ratio; MUGAN: thermal infrared image colorization using xixed-skipping UNet and generative ad-versarial network; LKAT-GAN: a GAN for ther-mal infrared image colorization based on large kernel and attentionUNet-transformer; SC-GAN: saliency map-guided colorization with generative adversarial network; TIC-CGAN: thermal infrared colorization via conditional generative adversarial network; CMMF-Net: a generative network based on clip-guided multi-modal feature fusion for thermal in-frared image colorization.

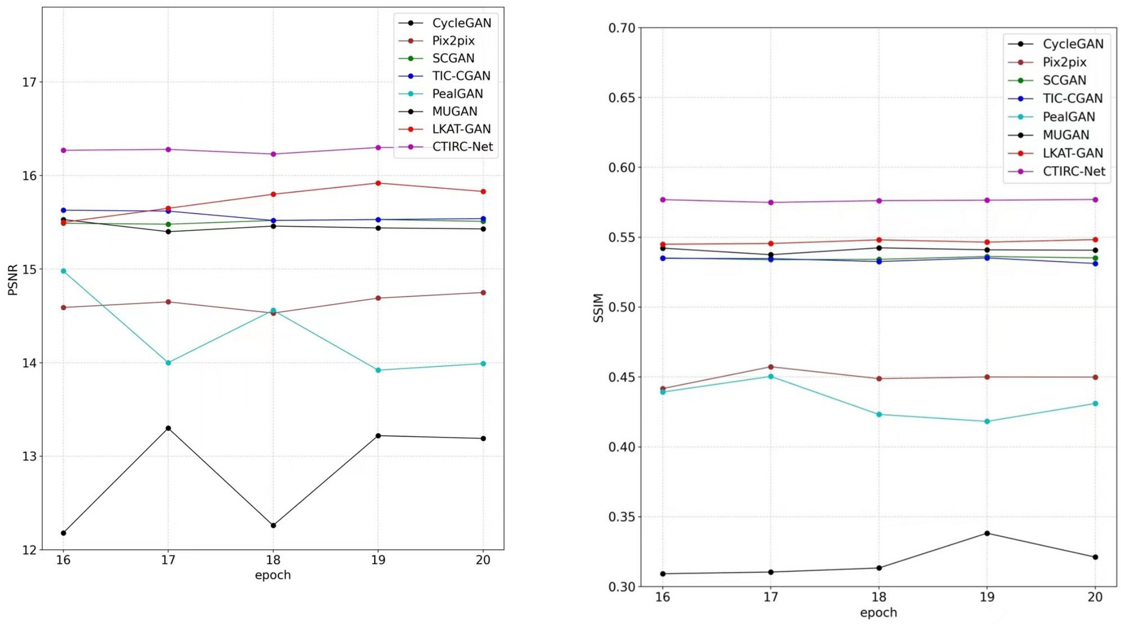

Figure 4. Results of different image colorization methods on KAIST. X-axis is the epoch and Y-axis is the index. KAIST: A multispectral pedestrian dataset, proposed by the Korea Advanced Institute of Science and Technology.

4.3. Experiments on IRVI dataset

On the IRVI dataset, we conducted complementary experiments with I2VGAN [34], SCGAN [29], TICC-GAN [28], PealGAN [32], MUGAN [30] and LKAT-GAN [31]. Due to the limited size of the dataset, we combined the data

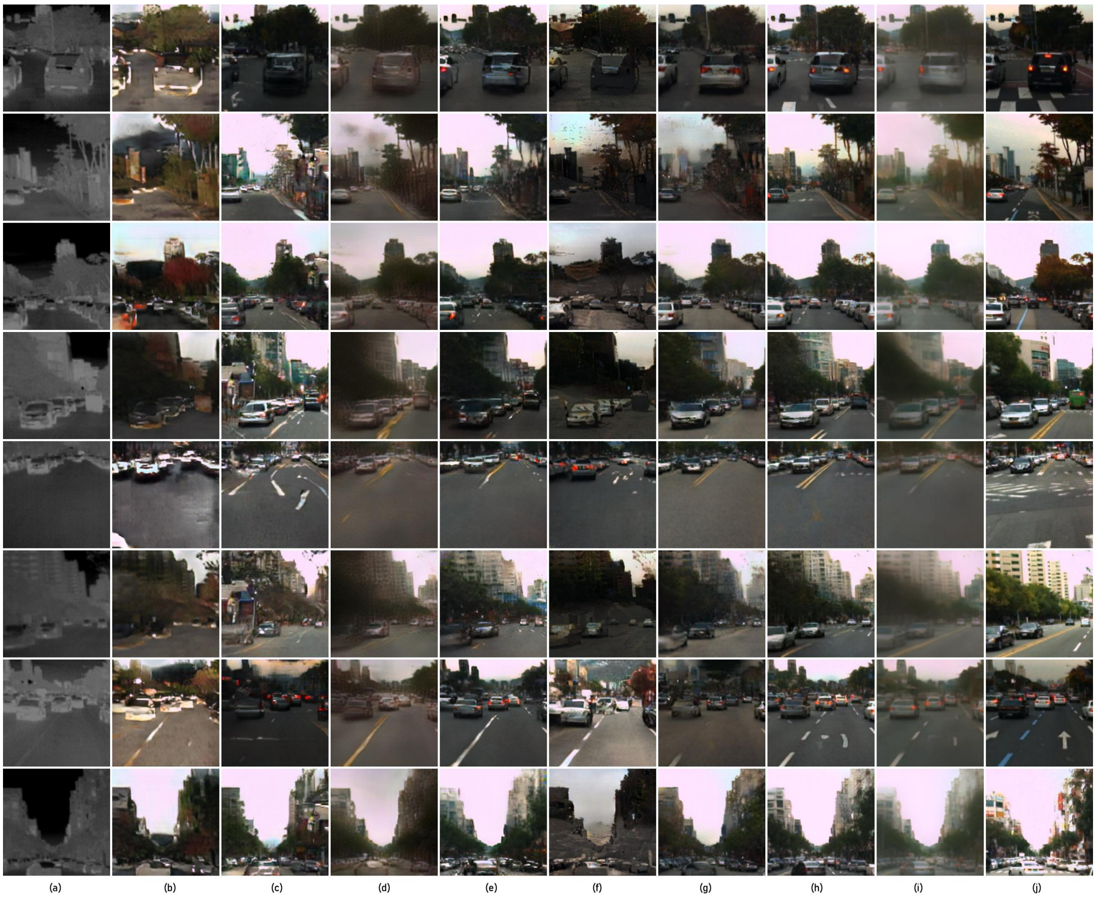

Figure 5. Colorized images using different image colorization methods on KAIST. (A) TIR images; (B) CycleGAN [33]; (C) Pix2pix [23]; (D) SCGAN [29]; (E) TIC-CGAN [28]; (F) PealGAN [32]; (G) MUGAN [30]; (H) LKAT-GAN [31]; (I) CMMF-Net; (J) True RGB images. The original infrared images (A) and RGB images (J) were obtained from https://soonminhwang.github.io/rgbt-ped-detection/data/, while the other images were reproduced using methods from other papers and generated with our laboratory's equipment. The specific methods can be found in the corresponding references listed in the bibliography. KAIST: A multispectral pedestrian dataset, proposed by the Korea Advanced Institute of Science and Technology; TIR: thermal infrared; SCGAN: saliency map-guided colorization with generative adversarial network; TIC-CGAN: thermal infrared colorization via conditional generative adversarial network; MUGAN: thermal infrared image colorization using xixed-skipping UNet and generative adversarial network; LKAT-GAN: a GAN for thermal infrared image colorization based on large kernel and attentionUNet-transformer; CMMF-Net: a generative network based on clip-guided multi-modal feature fusion for thermal infrared image colorization.

from two scenarios to form the training set, while selecting data from other scenarios for the test set. The IRVI dataset encompasses a wide range of scenes and diverse image content, including low-light, haze, and rain/snow weather conditions. The image content includes vehicles, pedestrians, buildings, and natural landscapes. The results show that our method consistently achieves the best overall performance across various lighting conditions and diverse image content, further confirming the superior generalization capability of our model. From the perspective of objective indicators, our method achieves the best performance compared with other comparison methods in traffic scenarios, and also performs well in monitoring scenarios, second only to LKAT-GAN method, as shown in Table 2. From the perspective of objective indicators, our method achieves optimal performance compared with other comparison methods in traffic scenes. Moreover, as shown in Figure 3, compared with other comparison methods, our proposed method has better performance in image detail and less noise, but the overall image is darker. In general, in the IRVI dataset, our method achieves good results in the objective indicators, but there are some shortcomings in subjective evaluation. For example, the detailed texture is better than that of other methods, but we lose the edges. The edges of other methods show better visual results, but with more pixel-level loss compared with our method, and the metrics reveal this

Table 2. Scores of different methods on IRVI. The best results are in bold

Method	Traffic		Monitoring	
	SSIM	PSNR	SSIM	PSNR
I2VGAN [34]	0.60	17.02	0.46	17.30
SCGAN [29]	0.65	17.72	0.50	16.55
TIC-CGAN [28]	0.62	16.94	0.52	17.65
PealGAN [32]	0.60	16.34	0.48	13.84
MUGAN [30]	0.66	18.42	0.46	14.72
LKAT-GAN [31]	0.65	17.43	**0.58**	**18.51**
CMMF-Net	**0.68**	**18.84**	0.56	18.39

IRVI: Dataset for infrared-to-visible video translation; SSIM: structural similarity index; PSNR: peak signal-to-noise ratio; SCGAN: saliency map-guided coloriza-tion with generative adversarial network; TIC-CGAN: thermal infrared colorization via conditional genera-tive adversarial network; MUGAN: thermal infrared image colorization using xixed-skipping UNet and generative adversarial network; LKAT-GAN: a GAN for thermal infrared image colorization based on large kernel and attentionUNet-transformer; CMMF-Net: a generative network based on clip-guided multi-modal feature fusion for thermal infrared image colorization.

phenomenon.

4.4. Ablation study

To evaluate the contribution of each module in CMMF-Net, we conducted an ablation study. Two experimental settings were designed to verify the effect of the module by excluding some parts of the original structure. The first experiment involved removing the Text_Encoder module experiment, which excluded text information in the model. The second experiment removed the CI module, which did not perform the text and image feature interaction processing but directly added the two features for output. Through these experiments, we were able to analyze the impact of each module on the colorization task and determine which modules were most critical for the final performance improvement.

This study will demonstrate the impact of each innovative module on the colorization results through ablation experiments. The colorization results of each ablation method are shown in Figure 6. For the subjective evaluation, the color results generated by CMMF-Net have been slightly improved in terms of detail and overall, and no obvious difference can be seen in the color results of each ablation method, which also proves that the performance of our overall method is excellent. The objective evaluation results are shown in Table 3. Finally, the overall CMMF-Net model had the highest SSIM and PSNR indicators among all ablation methods. Through the ablation experiment without adding text information, it is found that the introduction of text features can effectively assist the model to better extract and process image feature information, and improve the network color performance. This proves the superiority of our proposed multi-modal model. In the ablation experiment without CI module, it is found that the effect of the model is improved when text information is directly added to the network without text and image related feature processing, which once again proves the auxiliary role of text information to the model. At the same time, the performance of the model is further improved after the addition of CI module, which also proves the importance of this module for multi-modal models in colorization tasks. In addition, the results of two ablation experiments show that our overall network model is significantly better than other comparison models.

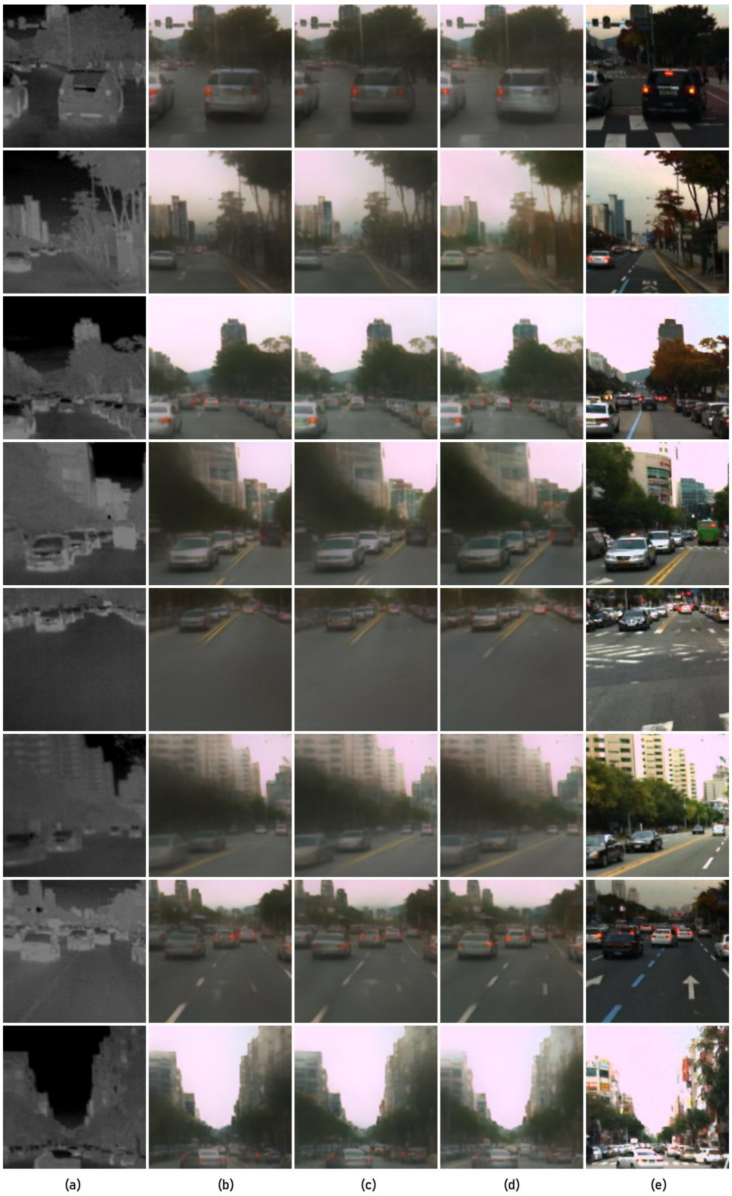

Figure 6. Colorization images of ablation studies on KAIST. (A) TIR images; (B) Without Text_Encoder block; (C) Without CI block; (D) CMMF-Net; (E) True RGB images. The original infrared images (A) were obtained from https://soonminhwang.github.io/rgbt-ped-detection/data/, while the other images were generated through our own experiments. KAIST: A multispectral pedestrian dataset, proposed by the Korea Advanced Institute of Science and Technology; TIR: thermal infrared; CI: cross-modal interaction; CMMF-Net: a generative network based on clip-guided multi-modal feature fusion for thermal infrared image colorization.

Table 3. Quantitative comparisons of ablation studies of CMMF-Net on the KAIST dataset

Index	Without Text_Encoder block	Without CI block	CMMF-Net
SSIM	0.57	0.57	**0.58**
PSNR	16.10	16.13	**16.22**

CMMF-Net: A generative network based on clip-guided multi-modal feature fusion for thermal infrared image colorization; KAIST: a multispectral pedestrian dataset, proposed by the Korea Advanced Institute of Science and Technology; CI: cross-modal interaction; SSIM: structural similarity index; PSNR: peak signal-to-noise ratio.

5. CONCLUSION

In this paper, we introduce an innovative method known as CMMF-Net for TIR image colorization. Leveraging the extensive handling of image-language features through our multi-modal model, CMMF-Net adeptly captures and processes the feature-rich information within TIR images, enabling accurate comprehension of the semantic content they contain. Furthermore, our multi-modal model is designed to process both image and language features. Supplementing text information assists the network in better extracting and processing feature information of the TIR image. To enhance the capacity of the network to represent data, a U-Net architecture is implemented, decoding features from diverse perspectives. Additionally, we implement a composite loss function to ensure the fidelity of the generated images to their real counterparts. The experimental outcomes obtained from the KAIST datasets underscore the superiority of CMMF-Net when contrasted with alternative methodologies for TIR image colorization tasks. Looking ahead, our future research will focus on two primary directions: colorization of TIR images captured at night and application of the proposed method in real-world applications. Infrared image colorization holds significant potential for various practical applications, including structural analysis of thermal imaging in the architectural domain, detection of temperature variations or pollution sources in environmental monitoring, and identification of fire sources or hazardous materials in security inspections. Furthermore, infrared image colorization and fusion techniques may be further integrated[45,46], for instance, by leveraging colorized infrared images in combination with visible light images in multi-modal tasks, thereby achieving superior performance in real-world scenarios.

DECLARATIONS

Authors' contributions

Made substantial contributions to the research process and wrote the original draft: Jiang, Q., Zhou, T., Jin, X.
Performed data acquisition: He, Y., Ma, W.
Provided guidance and support: Hou, J., Abdul Ghan, A.S., Miao, S.

Availability of data and materials

The original infrared images in Figures 4-6 were sourced from publicly accessible datasets, which are available for open access and use. The other images used in the comparative experiments were generated using our laboratory's experimental equipment. The usage of these images complies with the usage guidelines of the platforms and datasets. Additionally, these two datasets are widely utilized by other researchers in the same field, and the method we employed to generate the comparative experimental images is consistent with common practices in the field. Specifically, the datasets used are: the KAIST Dataset and the IRVI Dataset.

Financial support and sponsorship

The research was supported by Yunnan Fundamental Research Projects (202301AW070007, 202301AU070210, 202201AT070112), Yunnan Province Expert Workstations (202305AF150078), Major Scientific and Technological Project of Yunnan Province (202202AD080002, 202302AD080006), National Key Research and Development Program of China under Grant No. 2024YF03014300 and National Natural Science Foundation of China (62261060).

Conflicts of interest

Jin, X. is a Junior Editorial Board Member of the journal *Intelligence & Robotics*. Together with Jiang, Q., he serves as a Guest Editor for the Special Issue titled "Applications of Generative Adversarial Networks in Computer Vision and Image Processing". They are not involved in any steps of the editorial process, including reviewer selection, manuscript handling, or decision-making. The other authors declared that there are no conflicts of interest.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Copyright

REFERENCES

1. Erden F, Çetin AE. Hand gesture based remote control system using infrared sensors and a camera. *IEEE Trans Consum Electron* 2014;60:675–80. DOI
2. Gade R, Moeslund TB. Thermal cameras and applications: a survey. *Mach Vis Appl* 2014;25:245–62. DOI
3. Wijnhoven RGJ, de With PHN. Identity verification using computer vision for automatic garage door opening. *IEEE Trans Consum Electron* 2011;57:906–14. DOI
4. Kang J, Anderson DV, Hayes MH. Face recognition for vehicle personalization with near infrared frame differencing. *IEEE Trans Consum Electron* 2016;62:316–24. DOI
5. Chen C, Xu Y, Yang X. User tailored colorization using automatic scribbles and hierarchical features. *Digit Signal Process* 2019;87:155–65. DOI
6. Tang Y, Zhu M, Chen Z, et al. Seismic performance evaluation of recycled aggregate concrete-filled steel tubular columns with field strain detected via a novel mark-free vision method. *Structures* 2022;37:426–41. DOI
7. Tang Y, Qi S, Zhu L, Zhuo X, Zhang Y, Meng F. Obstacle avoidance motion in mobile robotics. *J Syst Simul* 2024;36:1-26. DOI
8. Wan S, Guan S, Tang Y. Advancing bridge structural health monitoring: insights into knowledge-driven and data-driven approaches. *J Data Sci Intell Syst* 2024;2:129–40. DOI
9. Hu K, Chen Z, Kang H, Tang Y. 3D vision technologies for a self-developed structural external crack damage recognition robot. *Autom Constr* 2024;159:105262. DOI
10. Zou C, Mo H, Gao C, Du R, Fu H. Language-based colorization of scene sketches. *ACM Trans Graph* 2019;38:1-16. DOI
11. Shin YG, Choi KA, Kim ST, Ko SJ. A novel single IR light based gaze estimation method using virtual glints. *IEEE Trans Consum Electron* 2015;61:254–60. DOI
12. Dong X, Li W, Wang X, Wang Y. Learning a deep convolutional network for colorization in monochrome-color dual-lens system. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2019. pp. 8255–62. DOI
13. Gupta RK, Chia AYS, Rajan D, Ng ES, Zhiyong H. Image colorization using similar images. In: Proceedings of the 20th ACM International Conference on Multimedia. MM '12. New York, NY, USA: Association for Computing Machinery; 2012. pp. 369–78. DOI
14. Iizuka S, Simo-Serra E, Ishikawa H. Let there be color! Joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. *ACM Trans Graph* 2016;35:1-11. DOI
15. Royer A, Kolesnikov A, Lampert CH. Probabilistic image colorization. *arXiv* 2017;arXiv:1705.04258. Available from: https://doi.org/10.48550/arXiv.1705.04258. [accessed 8 Jan 2025].
16. Deshpande A, Lu J, Yeh MC, Jin Chong M, Forsyth D. Learning diverse image colorization. *arXiv* 2016;arXiv:1612.01958. Available from: https://doi.org/10.48550/arXiv.1612.01958. [accessed 8 Jan 2025].
17. Larsson G, Maire M, Shakhnarovich G. Learning representations for automatic colorization. In: European conference on computer vision. Springer; 2016. pp. 577–93. DOI
18. Levin A, Lischinski D, Weiss Y. Colorization using optimization. In: ACM SIGGRAPH 2004 Papers. SIGGRAPH '04. New York, NY, USA: Association for Computing Machinery; 2004. pp. 689–94. DOI
19. Reinhard E, Adhikhmin M, Gooch B, Shirley P. Color transfer between images. *IEEE Comput Graph Appl* 2001;21:34–41. DOI
20. Larsson G, Maire M, Shakhnarovich G. Learning representations for automatic colorization. *arXiv* 2016;arXiv:1603.06668. Available from: https://doi.org/10.48550/arXiv.1603.06668. [accessed 8 Jan 2025].
21. Zhao J, Han J, Shao L, Snoek CGM. Pixelated semantic colorization. *Int J Comput Vis* 2020;128:818–34. DOI
22. Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks. *Commun ACM* 2020;63:139–44. DOI
23. Isola P, Zhu JY, Zhou T, Efros AA. Image-to-image translation with conditional adversarial networks. *arXiv* 2016;arXiv:1611.07004. Available from: https://doi.org/10.48550/arXiv.1611.07004. [accessed 8 Jan 2025].
24. Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. *arXiv* 2020;arXiv:2006.11239. Available from: https://doi.org/10.48550/arXiv.2006.11239. [accessed 8 Jan 2025].
25. Song J, Meng C, Ermon S. Denoising diffusion implicit models. *arXiv* 2020;arXiv:2010.02502. Available from: https://doi.org/10.48550/arXiv.2010.02502. [accessed 8 Jan 2025].
26. Saharia C, Chan W, Chang H, et al. Palette: image-to-image diffusion models. *arXiv* 2021;arXiv:2111.05826. Available from: https://doi.org/10.48550/arXiv.2111.05826. [accessed 8 Jan 2025].
27. Berg A, Ahlberg J, Felsberg M. Generating visible spectrum images from thermal infrared. In: 2018 IEEE/CVF Conference on Computer

Vision and Pattern Recognition Workshops (CVPRW); 2018 Jun 18-22; Salt Lake City, USA. IEEE; 2018. DOI

28. Kuang X, Zhu J, Sui X, et al. Thermal infrared colorization via conditional generative adversarial network. *Infrared Phys Technol* 2020;107:103338. DOI
29. Zhao Y, Po LM, Cheung KW, Yu WY, Rehman YAU. SCGAN: saliency map-guided colorization with generative adversarial network. *IEEE Trans Circuits Syst Video Technol* 2021;31:3062–77. DOI
30. Liao H, Jiang Q, Jin X, et al. MUGAN: thermal infrared image colorization using mixed-skipping UNet and generative adversarial network. *IEEE Trans Intell Vehicles* 2023;8:2954-69. DOI
31. He Y, Jin X, Jiang Q, et al. LKAT-GAN: a GAN for thermal infrared image colorization based on large kernel and attentionUNet-transformer. *IEEE Trans Consum Electron* 2023;69:478-89. DOI
32. Luo F, Li Y, Zeng G, Peng P, Wang G, Li Y. Thermal infrared image colorization for nighttime driving scenes with top-down guided attention. *IEEE Trans Intell Transp Syst* 2022;23:15808-23. DOI
33. Bansal A, Ma S, Ramanan D, Sheikh Y. Recycle-GAN: unsupervised video retargeting. *arXiv* 2018;arXiv:1808.05174. Available from: https://doi.org/10.48550/arXiv.1808.05174. [accessed 8 Jan 2025].
34. Li S, Han B, Yu Z, Liu CH, Chen K, Wang S. I2V-GAN: unpaired infrared-to-visible video translation. *arXiv* 2021;arXiv:2108.00913. Available from: https://doi.org/10.48550/arXiv.2108.00913. [accessed 8 Jan 2025].
35. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27-30; Las Vegas, USA. IEEE; 2016. pp. 770-8. DOI
36. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2015. pp. 1–9. Available from: https://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Szegedy_Going_Deeper_With_2015_CVPR_paper.html. [Last accessed on 8 Jan 2025]
37. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z. Swin transformer: hierarchical vision transformer using shifted windows. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10-17; Montreal, Canada. IEEE; 2021. pp. 9992-10002. DOI
38. Li G, Duan N, Fang Y, Gong M, Jiang D. Unicoder-Vl: a universal encoder for vision and language by cross-modal pre-training. *Proc AAAI Conf Artif Intell* 2020;34:11336–44. DOI
39. Lu J, Batra D, Parikh D, Lee S. Vilbert: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv* 2019;arXiv:1908.02265. Available from: https://doi.org/10.48550/arXiv.1908.02265. [accessed 8 Jan 2025].
40. Tan H, Bansal M. Lxmert: learning cross-modality encoder representations from transformers. *arXiv* 2019;arXiv:1908.07490. Available from: https://doi.org/10.48550/arXiv.1908.07490. [accessed 8 Jan 2025].
41. Radford A, Kim JW, Hallacy C, et al. Learning transferable visual models from natural language supervision. *arXiv* 2021;arXiv:2103.00020. Available from: https://doi.org/10.48550/arXiv.2103.00020. [accessed 8 Jan 2025].
42. Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L. Segment anything. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV); 2023 Oct 01-06; Paris, France. IEEE; 2023. pp. 3992-4003. DOI
43. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv* 2014;arXiv:1409.1556. Available from: https://doi.org/10.48550/arXiv.1409.1556. [accessed 8 Jan 2025].
44. Hwang S, Park J, Kim N, Choi Y, So Kweon I. Multispectral pedestrian detection: benchmark dataset and baseline. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2015 Jun 07-12; Boston, USA. IEEE; 2015. pp. 1037-45. DOI
45. Luo FY, Liu SL, Cao YJ, Yang KF, Xie CY, Liu Y. Nighttime thermal infrared image colorization with feedback-based object appearance learning. *IEEE Trans Circuits Syst Video Technol* 2024;34:4745–61. DOI
46. Tan MJ, Gao SB, Xu WZ, Han SC. Visible-infrared image fusion based on early visual information processing mechanisms. *IEEE Trans Circuits Syst Video Technol* 2021;31:4357–69. DOI