



SCAKD: a knowledge distillation framework based on spatial-corner attention for infrared and visible image fusion

Jiakun Zhao¹ · Yige Cai¹

Accepted: 1 March 2024 / Published online: 2 May 2024
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2024

Abstract

Infrared and visible image fusion is a kind of image enhancement technology, aiming at combining the characteristics of infrared image and visible image to synthesize them into an image that contains more information and is more suitable for human and computer perception and recognition. Most of the existing fusion methods focus on improving the fusion effect but ignore the model size and computing cost, which is not conducive to the actual deployment of such models on resource-constrained platforms such as mobile terminals. To solve this problem, we propose a new knowledge distillation framework for infrared and visible image fusion. Based on the spatial and corner attention features, the framework uses the middle-layer feature mapping to construct the distillation loss function and guides the lightweight student model to learn the fusion ability of the heavyweight teacher model. After distillation training, the student model can achieve a similar level of integration with the teacher model. A large number of qualitative and quantitative experiments show that our method has better fusion performance and lower model complexity than the existing methods.

Keywords Knowledge distillation · Image fusion · Infrared image · Attention mechanism.

1 Introduction

Infrared and visible image fusion is a process of synthesizing two images from the visible image sensor and the infrared sensor at the same time on the same target into one image. The infrared sensor reflects the temperature difference or radiation difference of the scene, which is not easy to be affected by complex conditions such as wind, sand, and smoke. However, the infrared image has some shortcomings such as not obvious details and poor imaging effect, so its visibility is not very ideal. Visible imaging sensor imaging according to the different reflectance of objects, only related to the reflection of the target scene, so visible image can capture rich texture details, to provide human visual perception of the real environment. But visible image is easy to be affected by the environment, which is not conducive to the detection of important targets. Therefore, the fusion of the two images can

play a complementary role, achieve a comprehensive expression of the scene, and provide necessary help for subsequent computer vision related tasks [1–4].

With the development of deep learning, many excellent fusion methods of infrared and visible have been proposed [5–10]. These fusion methods based on deep learning not only overcome the difficulty in designing fusion rules of traditional methods [11–14], but also achieve better fusion effect. However, most of the existing fusion methods focus on improving the fusion effect and ignore the model size and computing cost, which is not conducive to the actual deployment of such models on resource-constrained platforms such as mobile terminals. Therefore, it is essential to design lightweight fusion models with advanced fusion effects and low cost.

There are three main methods for lightweight models, including lightweight network design [15, 16], model pruning [17, 18], and knowledge distillation [19–22]. Lightweight network design and model pruning methods require elaborate design and can cause performance degradation. In contrast, knowledge distillation method can realize model compression more efficiently without loss of model performance.

The knowledge distillation method for infrared and visible image fusion has not been widely studied. In order to achieve

✉ Yige Cai
caiyige@stu.xjtu.edu.cn

Jiakun Zhao
zhaojk@xjtu.edu.cn

¹ School of Software Engineering, Xi'an Jiaotong University, Xi'an 710049, China

efficient infrared and visible image fusion, a new knowledge distillation framework for infrared and visible image fusion is proposed in this paper. The distillation framework is based on the spatial-corner attention features, named SCAKD. In this framework, we design a new complex fusion model based on generative adversarial network as the teacher model, and the student model is designed as a more lightweight network with the same basic architecture as the teacher model. The distillation module uses a combination of spatial attention and corner attention to transform the features of the middle layer corresponding to the teacher and student models from two directions: spatial domain containing more information and corner point reflecting texture details. Then, the distillation loss function is designed according to the obtained attention matrix, and the characteristic expression ability of the teacher model is taught to the student model through the loss function in the training process. Experiments show that the framework of knowledge distillation proposed in this paper effectively compresses the fusion model and makes the distillation student model achieve the best fusion performance.

The contribution of this work mainly includes the following aspects:

- In this paper, a knowledge distillation framework for infrared and visible image fusion is proposed. The framework uses the correlation in feature mapping to guide the training of student models and endows lightweight student models with excellent fusion ability.
- A novel infrared and visible image fusion model is proposed, which is based on spatially adaptive normalization (SPADE) [23] and efficient channel attention (ECA) [24], and the fusion results are superior to the existing fusion methods, so this model is used as the teacher model of the proposed distillation framework.
- A feature conversion method for spatial-corner attention (SCA) is proposed, and the effects of spatial attention and corner attention on distillation performance are discussed.
- Based on multiple experiments conducted on public datasets, it has been demonstrated that our proposed method exhibits the highest level of comprehensive performance for infrared and visible fusion tasks.

The organization of this paper is as follows. Section 2 presents a review of related works. Section 3 provides a detailed description of the proposed method. Section 4 analyzes the experimental results. Lastly, Sect. 5 concludes the paper.

2 Related work

Although traditional methods have achieved some research results, the fusion performance is limited. In this paper, we focus on the deep learning-based methods.

2.1 Infrared and visible image fusion

Traditional methods for the fusion of infrared and visible images mainly include sparse expression-based methods [25, 26], multi-scale transformation-based methods [27, 28], spatial transformation-based methods [14], and saliency detection-based methods [29]. However, traditional methods require the design of complex fusion rules, which inevitably leads to a decrease in fusion efficiency. In recent years, deep learning has been widely applied to the task of infrared and visible image fusion and has shown good performance. Based on deep learning, fusion methods can be divided into two categories according to theory: convolutional neural network (CNN)-based methods and generative adversarial network (GAN)-based methods. CNN has significant advantages in feature extraction, providing more information than traditional manual feature extraction methods [30–33]. Although CNN-based fusion methods can achieve good fusion results, CNN is more suitable for supervised learning, while the task of infrared and visible image fusion is unsupervised learning, making GAN more suitable for this type of task. In 2019, Ma et al. first proposed a GAN-based method for infrared and visible image fusion called FusionGAN [8]. This method first uses a generator to generate the fusion image and then sends both the generated fusion image and the original visible image to the discriminator for judgment. When the discriminator cannot distinguish, the generated fusion image is optimal. Zhang et al. proposed a fusion method called GAN-FM [9] based on full-size skip connections and dual Markov discriminators. The generator of this method can extract multi-scale, multi-level features, and the use of dual Markov discriminators can simultaneously estimate the probability distributions of infrared and visible images. In reference [10], Wang et al. proposed a feature interaction fusion module that can adaptively select features, avoiding feature smoothing caused by immature fusion rules and making the texture details of the fusion image more clear.

However, while the above-mentioned methods have achieved breakthroughs in fusion performance, they also suffer from the problem of high model complexity. In order to facilitate the practical deployment of fusion models, a knowledge distillation framework tailored to the infrared and visible image fusion task is needed to effectively compress the fusion model.

2.2 Knowledge distillation

The prototype of knowledge distillation stems from the study of Ba et al. [34], who proposed to train a small neural network under the supervision of a large neural network, and gradually enable the small network to perform the same functions as the large network. The large network, which plays a supervisory role in this process, is called the teacher network, and the small network, which is trained through knowledge learning, is called the student network. On the basis of summarizing previous works, Hinton et al. [35] formally proposed the concept of knowledge distillation, which aims to help the student model achieve competitive performance, or even outperform the teacher model, by imitating the teacher model. In knowledge distillation, we mainly focus on transfer knowledge, distillation strategies, and teacher–student architecture. Generally, transfer knowledge can be categorized into three types, namely response-based, feature-based, and relation-based knowledge. Response-based knowledge generally refers to neuron responses, i.e., the last layer’s logic output of the teacher model. The core idea is to let the student model imitate the output of the teacher model, which is the most classic, simple, and effective processing method. In deep convolutional neural networks, the learned knowledge is hierarchical, so the features of middle layers can also serve as carriers of knowledge. FitNets, proposed by Romero et al. [36], was the first to introduce intermediate layer representations and use the teacher model’s intermediate layer as a hint to the corresponding layer of the student model, thereby improving the performance of the student model by expecting it to directly imitate the feature activation values of the teacher model. Zagoruyko et al. [37] proposed to align the attention maps between the teacher network and student network in the feature domain, based on the idea of attention mechanism, thereby allowing the student network to learn the feature information of the teacher network. Relation-based knowledge further explores the relationships among different layers and different data samples. Yim et al. [19] proposed a method called flow of solution process (FSP), which defined the Gram matrix between two layers, thus summarizing the relationships between different feature maps.

In summary, the essence of knowledge distillation lies in designing the transfer knowledge in the teacher model, and transferring the extracted knowledge from the teacher model to the student model during training, enabling the lightweight student model to possess the performance of a heavyweight teacher model. In different task scenarios, designing appropriate transfer knowledge suitable for the task is also a worthwhile research problem. In the task of infrared and visible image fusion, the model output is a generated fusion image, but there is no standard fusion label. Although using the output of the teacher model as transfer knowledge is an effective method, it will limit the upper

limit of the fusion performance of the student model. Therefore, in the scenario of infrared and visible image fusion, it is more suitable to extract the transfer knowledge of the teacher model using middle-layer features that reflect feature expression ability, as the saying goes, “It is better to teach a man to fish than to give him a fish.”

3 Proposed method

3.1 Overall framework

The knowledge distillation framework we propose for infrared and visible image fusion is shown in Fig. 1. The framework consists of four main parts: (1) teacher model; (2) student model; (3) distillation module based on spatial-corner attention features; and (4) joint loss function. We designed a new fusion method based on generative adversarial networks and used the pretrained generator in this method as the teacher model. The student model was designed as a lighter network with a structure similar to that of the teacher model. The distillation strategy of the teacher–student network mainly adopted the proposed distillation module based on spatial-corner attention features. Based on this, a joint loss function was established to iteratively train the student model, which not only trained through adversarial training with the dual discriminators, but also obtained important knowledge transferred from the teacher model through the distillation module. The goal of the entire distillation framework is to obtain an infrared and visible fusion model with both effectiveness and real-time performance.

3.2 Teacher–student model

In order to improve the training effect of the student model, the teacher model needs to have sufficient complexity and high fusion accuracy to provide high-quality and rich knowledge to guide the training of the student model. In our framework, the teacher model is a pretrained model with high model complexity and excellent fusion performance, whose network structure is shown in the red dashed box in Fig. 1. The teacher model is an encoder–decoder structure network consisting of an encoder and a decoder. The encoder is a dual-stream network composed of two feature extraction branches, both of which are composed of four residual dense blocks (RDBs) [38] and one spatially adaptive normalization (SPADE) [23] module to extract infrared and visible features separately. Among them, RDB can fully extract features by using continuous memory mechanisms and local dense connections, while SPADE normalizes the feature maps differentially by treating the source image as an additional semantic input, which not only retains the normalization function, but also ensures that the output features

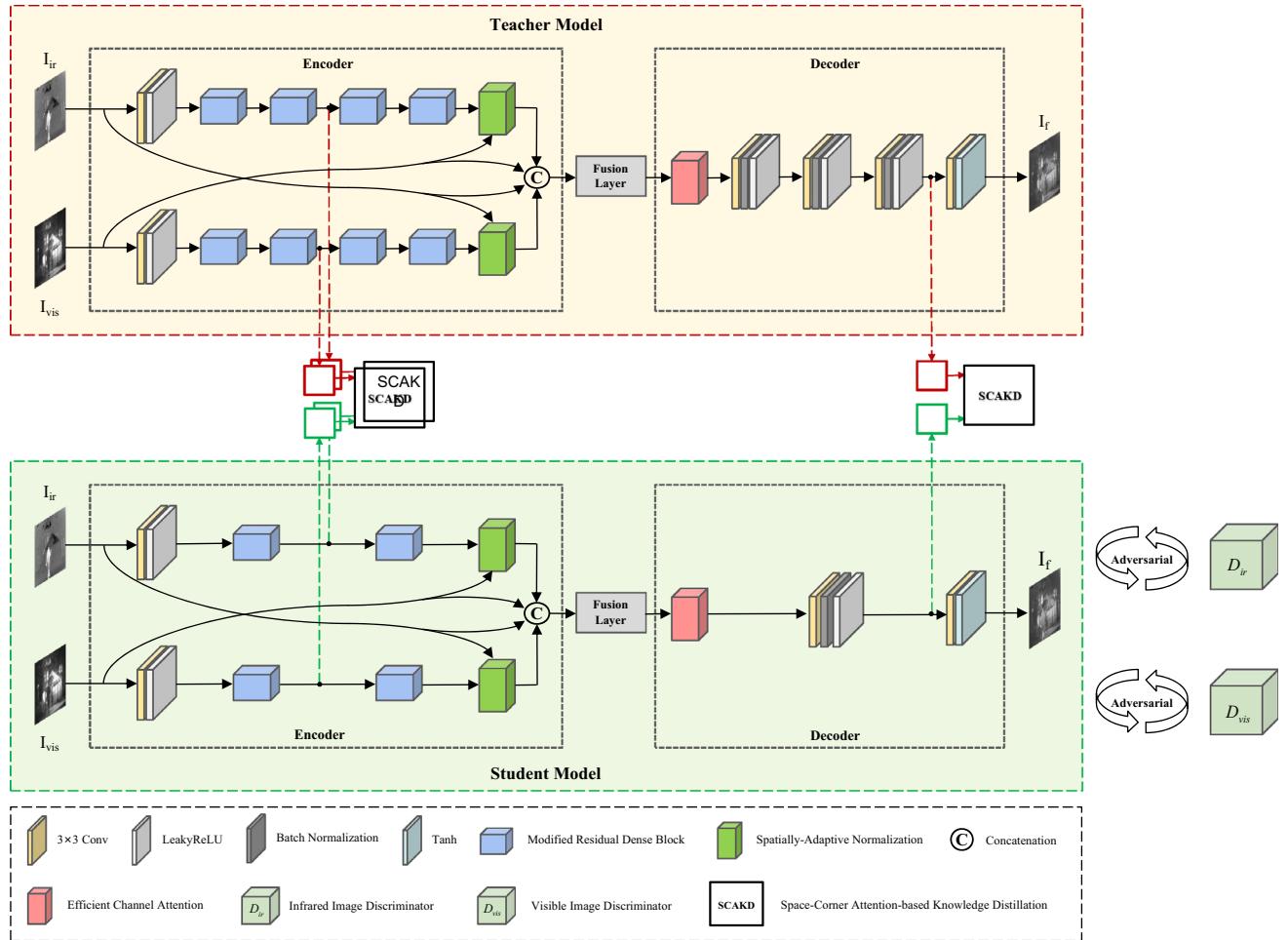


Fig. 1 Framework of the proposed SCAKD for infrared and visible image fusion

contain the initial semantic information of the input source image. The decoder consists of an efficient channel attention (ECA) [24] module and four convolution modules, where ECA is used to learn the channel attention weights of the fusion features, and the following four convolution modules are used to reconstruct the features and generate the final fusion image.

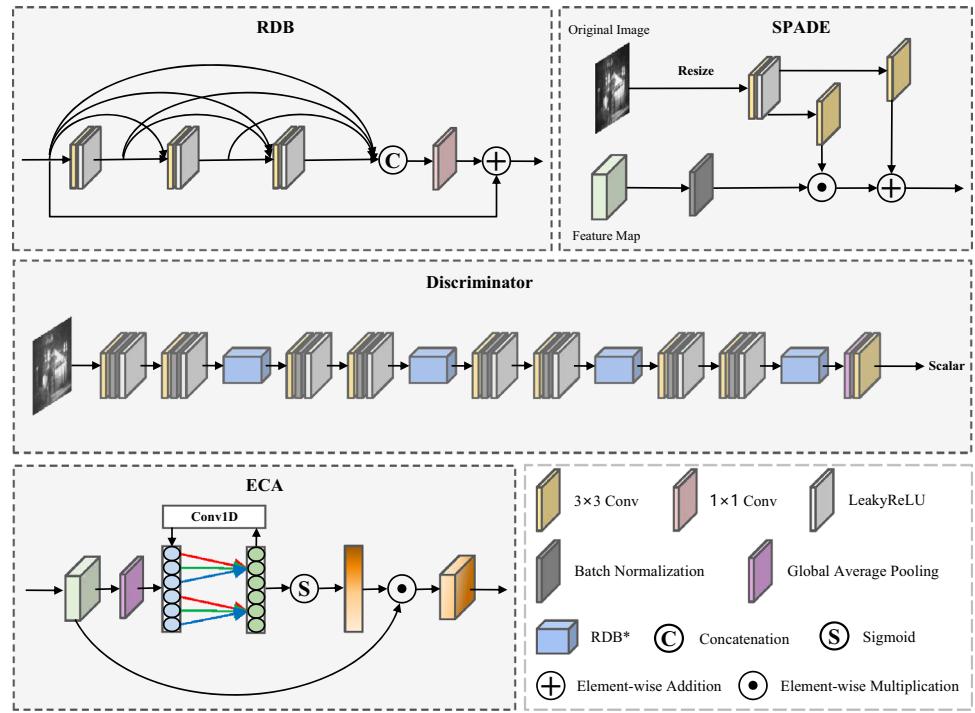
To ensure effective knowledge transfer, the structure of the student model should be similar to that of the teacher model. If the structural difference between the two is too large, it may result in the inability of the student model to effectively capture and use the knowledge, thereby reducing the effectiveness of knowledge distillation. Therefore, the student model in this chapter's method is designed to have a structure that is basically identical to that of the teacher model, as shown in the green dashed box in Fig. 1. Compared with the teacher model, in the encoder, the number of RDBs in the student model is reduced by half, and each branch only has two IRDBs, making the model more lightweight. In the decoder, the convolution modules of the student model are

simplified to two, further reducing the weight of the student model. In addition, we also designed dual discriminators, D_{ir} and D_{vis} , to perform adversarial training with the student model, which is used to further constrain the student model and guide the generated fusion image to have both significant contrast and rich texture information. The D_{ir} and D_{vis} have the same network structure but independent parameters.

3.3 Knowledge distillation based on spatial-corner attention features (SCAKD)

The key to knowledge distillation is to design an appropriate loss function to guide the student model to learn effective information extracted from the teacher model during the training process. However, most existing distillation methods [19, 36, 37] are designed for image classification and are not suitable for image generation tasks such as infrared and visible image fusion, because the feature representation space of regression problems like image generation is unbounded [39, 40]. Knowledge distillation for image fusion requires

Fig. 2 Network architecture of the main modules of the teacher–student network



limiting the solution space. Therefore, we propose an offline distillation method based on spatial-corner attention features to achieve efficient infrared and visible image fusion. This method extracts feature maps from corresponding positions of the teacher model and the student model, and then performs spatial-corner feature transformation, encouraging the student model to generate attention matrices similar to those of the teacher model.

For knowledge distillation, the intermediate layers of the model also carry a lot of knowledge, and the degree of knowledge abstraction represented by the feature map from shallow to deep also varies from low to high. This paper proposes a feature-based knowledge distillation approach from two aspects: spatial attention feature and corner attention feature. As shown in Fig. 1, we extract the feature maps F^T and F^S of the teacher–student model at the aligned positions of the encoder and decoder and define the spatial-corner attention feature mapping. Considering the feature differences between infrared and visible images, and to explore the consistency in the feature mapping, we propose to calculate the joint matrix F_{SCA} of spatial attention and corner attention. They are generated from two layers of feature mapping in both the infrared branch and visible branch of the encoder and high-layer feature mapping in the decoder. Each layer's feature map undergoes spatial attention transformation and corner attention transformation, and then weighted summation is performed to obtain the spatial-corner attention feature. The distillation loss based on spatial-corner attention can be expressed as:

$$L_{SCAKD} = \frac{1}{|F_{SCA}|} \sum_{l=1}^{l'} \|F_{SCAl}^T - F_{SCAl}^S\|_1 \quad (1)$$

where F_{SCAl}^T and F_{SCAl}^S are the teacher–student network's spatial-corner attention matrices extracted from the l -th layer feature map; l' is the total number of layers for feature extraction. $|F_{SCA}|$ represents the total number of elements in the spatial-corner attention matrix. The $\|\cdot\|_1$ denotes the L1 norm, which represents the sum of the absolute values of all elements in an n -dimensional vector. The F_{SCA} of each layer is composed of the spatial attention map F_{SA} and the corner attention map F_{CA} of that layer:

$$F_{SCA} = \alpha F_{SA} + (1 - \alpha) F_{CA} \quad (2)$$

where α is a balancing parameter. The extraction process of spatial attention features and corner attention features is shown in Fig. 1.

3.3.1 Spatial attention module

Pixels in an image are not independent of each other, but have certain correlations. Various objects in an image are represented by the interactions of a large number of pixels. Inspired by this, we explore the correlation between pixels from a spatial perspective. The pipeline is shown in the red dashed box in Fig. 3. Given a batch of feature maps $F \in \mathbb{R}^{B \times C \times H \times W}$, we first compress the features by an average pooling layer with a window size of 4×4 , resulting in $F \in \mathbb{R}^{B \times C \times H' \times W'}$.

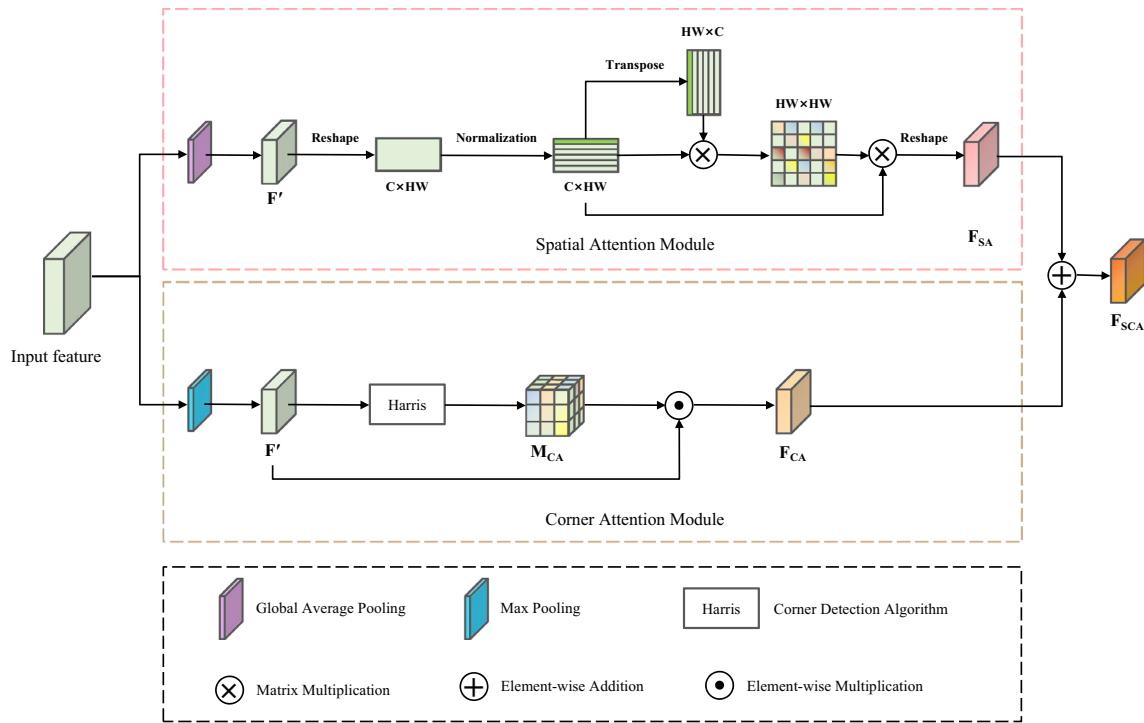


Fig. 3 Network architecture of spatial-corner attention distillation model

Then, $F \in \mathbb{R}^{B \times C \times H' \times W'}$ is reshaped into $F \in \mathbb{R}^{B \times C \times H'W'}$ and normalized along the channel dimension. Each pixel in the feature map can be regarded as a normalized vector in the channel dimension. Since cosine similarity is actually the normalization of inner product of vectors, after normalization, the transpose of the feature matrix is multiplied by the feature matrix to obtain the spatial attention matrix:

$$\tilde{F}_{[b',;h'w']} = \frac{F_{[b',;h'w']}}{\|F_{[b',;h'w']}\|_2} \quad (3)$$

where \tilde{F} represents the feature after normalization of channel dimension. The shape of the generated spatial attention matrix F_{SCA} is $B \times C \times H'W'$. The value of any element in the F_{SCA} represents the degree of spatial correlation between the corresponding two pixels.

3.3.2 Corner attention module

In an image, a point can be called an edge point if it has a significant gradient change in only one direction, while a point can be called a corner point if it has a significant gradient change in two directions. Corner points are important feature points in an image, and their surrounding areas are often accompanied by a large number of edge points, representing the texture details of the image. Therefore, corner points contain valuable knowledge. Based on this, we propose a

corner attention module. As shown in the yellow dashed box in Fig. 3, similar to the spatial attention module mentioned above, the input feature $F \in \mathbb{R}^{B \times C \times H' \times W'}$ is compressed by a pooling layer to obtain $F \in \mathbb{R}^{B \times C \times H' \times W'}$. Unlike the average pooling layer in the spatial attention module, since the corner attention module focuses more on the edges and textures of the image, the maximum pooling layer is used here. Then, after the Harris corner detection algorithm and normalization process, the corner attention matrix is obtained:

$$F_{CA}^{chw} = \begin{cases} 1, & \text{Harris}(F^{chv}) > \beta \text{Max}(\text{Harris}(F^c)) \\ 0, & \text{Harris}(F^{chv}) \leq \beta \text{Max}(\text{Harris}(F^c)) \end{cases} \quad (4)$$

where $c \in [1, C]$, $h \in [1, H']$, $w \in [1, W']$. If $F_{CA}^{chw} = 1$, then F is a corner point at (h, w) of channel c . F^c represents the feature diagram of F in channel c . $\text{Harris}(\cdot)$ is a corner detection function. The larger the value of the calculated point is, the closer the point is to the corner region. β is the corner attention factor.

3.4 Loss function

Distillation loss L_{SCKD} has been described above, as shown in Eq. (1). During the training process, in addition to knowledge distillation, the generator of the student model also plays a confrontation game with two discriminators, as shown in the knowledge distillation framework in Fig. 1. The adversarial training strategy of the student model and the teacher

model is the same. The loss function of the generator also includes content loss L_{content} and counter loss L_{adv} . The formula is as follows:

$$\begin{aligned} L_{\text{content}} &= \frac{1}{HW} \left(\left(\|I_f - I_{ir}\|_F^2 + \|I_f - I_{vis}\|_F^2 \right) \right. \\ &\quad \left. + \xi \left(\|\nabla I_f - \nabla I_{ir}\|_F^2 + \|\nabla I_f - \nabla I_{vis}\|_F^2 \right) \right) \end{aligned} \quad (5)$$

$$\begin{aligned} L_{\text{adv}} &= \frac{1}{N} \sum_{n=1}^N \left(D_{ir} \left(I_f^n \right) - a_1 \right)^2 \\ &\quad + \frac{1}{N} \sum_{n=1}^N \left(D_{vis} \left(I_f^n \right) - a_2 \right)^2 \end{aligned} \quad (6)$$

where I_{ir} represents infrared image, I_{vis} represents visible image, and I_f is fusion result. ξ is the weight coefficient, $\|\cdot\|_F$ is Frobenius norm, and ∇ stands for gradient operator. N represents the total number of fused images, and a_1 and a_2 represent the label values that the generator wants the discriminator D_{ir} and D_{vis} to trust.

In addition, the two discriminators of the student model also correspond to two loss functions, namely infrared discriminator loss function $L_{D_{ir}}$ and visible discriminator loss function $L_{D_{vis}}$:

$$L_{D_{ir}} = \frac{1}{N} \sum_{n=1}^N (D_{ir}(I_{ir}) - c_{ir})^2 + \frac{1}{N} \sum_{n=1}^N (D_{ir}(I_f) - c)^2 \quad (7)$$

$$\begin{aligned} L_{D_{vis}} &= \frac{1}{N} \sum_{n=1}^N (D_{vis}(I_{vis}) - c_{vis})^2 \\ &\quad + \frac{1}{N} \sum_{n=1}^N (D_{vis}(I_f) - c)^2 \end{aligned} \quad (8)$$

where c_{ir} , c_{vis} , and c , respectively, represent labels of infrared image I_{ir} , visible image I_{vis} , and fusion image I_f .

In summary, the joint loss function of the student model is:

$$L_S = \lambda_1 L_{\text{SCAKD}} + \lambda_2 L_{\text{content}} + L_{\text{adv}} \quad (9)$$

where λ_1 and λ_2 are the weight coefficients, which are used to control the importance of distillation loss and content loss in the joint loss function. By using this joint loss function, the student model is trained in confrontation and knowledge is distilled at the same time, and the knowledge transferred by the teacher model is used to achieve the best fusion performance of the student model.

4 Experimental results

4.1 Datasets and training settings

We selected 80 pairs of infrared and visible images from the TNO [41] and RoadScene [6] datasets as the training dataset for this experiment and expanded the dataset by cropping the images during training. Additionally, we randomly selected 20 pairs of images from each of the two datasets as the testing dataset for evaluating the model.

The knowledge distillation training process for the student model is described in Algorithm 1. The total number of epochs for training is $M = 200$, the batch size is $b = 32$, and it takes m steps to traverse the entire training set. The iteration ratio between the two discriminators and the generator is $k = 2$. The model is trained using the Adam optimizer with a learning rate set to $2e-4$. The values of λ_1 and λ_2 are both set to 100. The labels a_1 , a_2 , c_{ir} , and c_{vis} are randomly generated numbers between 0.7 and 1.2, and the label c is a random number between 0 and 0.3. In addition, the balance parameter α for the spatial-corner attention feature needs to be set according to the feature extraction location. In a large number of training experiments, the value of α is set to 0.8 for the infrared branch in the encoder, 0.2 for the visible branch in the encoder, and 0.5 in the decoder based on the training effectiveness.

The training experiments in this paper are conducted on three NVIDIA RTX-3090 GPUs, each with a memory size of 24GB, while inference is performed on a single GPU. The proposed model is built using the PyTorch 1.8 framework.

Algorithm 1 Training Procedure of SCAKD.

```

1: for M epochs do
2:   for m steps do
3:     for k times do
4:       Select b infrared patches  $\{I_{ir}^1, I_{ir}^2 \dots I_{ir}^b\}$ ;
5:       Select b visible patches  $\{I_{vis}^1, I_{vis}^2 \dots I_{vis}^b\}$ ;
6:       Select b fused patches  $\{I_f^1, I_f^2 \dots I_f^b\}$ ;
7:       Update the parameters of the  $D_{ir}$  and  $D_{vis}$  by Adam optimizer:  $\nabla D(L_D)$ ;
8:     end for
9:     Select b infrared patches  $\{I_{ir}^1, I_{ir}^2 \dots I_{ir}^b\}$ ;
10:    Select b visible patches  $\{I_{vis}^1, I_{vis}^2 \dots I_{vis}^b\}$ ;
11:    Select the intermediate feature maps of corresponding layers in the teacher-student model for spatial-corner attention feature conversion, and then construct the distillation loss  $L_{\text{SCAKD}}$ .
12:    Update the parameters of the student model by Adam optimizer:  $\nabla G(L_S)$ ;
13:  end for
14: end for

```

4.1.1 Performance metrics

This paper evaluates the performance of various methods in two aspects: subjective qualitative analysis and objective quantitative analysis. The subjective qualitative analysis mainly observes the fusion images of various methods through human visual perception, and obtains subjective evaluation. The objective quantitative analysis analyzes the fusion results by selecting six typical indicators [8]: entropy (EN), structural similarity index measurement (SSIM), standard deviation (SD), spatial frequency (SF), average gradient (AG), and correlation coefficient (CC). The higher the values of all objective evaluation indicators, the better the quality of the fusion results. In addition, we also conduct a model complexity evaluation of each method from the aspects of computation complexity and parameter quantity, to verify the efficiency of the proposed methods.

4.2 Ablation experiment

4.2.1 Ablation of distillation module

To verify the effectiveness of the proposed distillation method, we train the student model by controlling different components of the joint loss function, and then compare on the test set. The results of the ablation experiment are shown in Table 1. The first line in the result indicates that only data training (DT) is used, that is, knowledge distillation is not carried out in the student model in adversarial training. The second line shows that on the basis of DT, spatial attention module (SA) and corner attention module (CA) are added, respectively. As can be seen from the results in the table, both SA and CA provide performance gains. The last line indicates that both SA and CA are introduced on the basis of DT, and the two are combined weighted, namely, the spatial-corner attention distillation model (SCAKD) proposed in this paper. It can be seen that the average value of SCAKD on the six evaluation indexes is the largest. In addition, we also demonstrate the effectiveness of our distillation method from the perspective effect. As shown in Fig. 4, the fusion image using knowledge distillation has richer texture details and more prominent infrared targets than the fusion image without knowledge distillation. In the following experiments, this chapter uses a combination of DT, SA, and CA as the default experimental setting.

4.2.2 Ablation of balancing parameter α

α is a sensitive parameter, and its different values at different positions will affect the weight allocation of the spatial attention feature and the corner attention feature in transferring

knowledge. Therefore, we explore the effect of different α settings on the distillation effect. $[\alpha_{ir}^{en} \backslash \alpha_{vis}^{en} \backslash \alpha^{de}]$ represents a group of parameters for α at different feature alignment positions (corresponding to the two feature alignment positions of the infrared branch and visible branch in the encoder and one feature alignment position in the decoder). With other environmental variables fixed, different parameter groups are used to perform distillation training on the student model. We explore a large number of α parameter groups, and the fusion effects of six selected parameter groups are shown in Fig. 5. It can be seen that the balance between fusion in the infrared and visible modes is best achieved when $[\alpha_{ir}^{en} \backslash \alpha_{vis}^{en} \backslash \alpha^{de}]$ is $[0.8 \backslash 0.2 \backslash 0.5]$. Table 2 further shows the quantitative evaluation results of different parameter groups. The results show that when $[\alpha_{ir}^{en} \backslash \alpha_{vis}^{en} \backslash \alpha^{de}]$ is $[0.8 \backslash 0.2 \backslash 0.5]$, the fusion results have the largest average value in four of the six objective evaluation indicators, and the second-largest in the other two, achieving the best objective evaluation results consistent with subjective evaluation. Therefore, this paper uses this group of settings as the beta parameter of the final model.

4.2.3 Comparison with other distillation methods

To verify the advancement of the proposed distillation method, we replace the spatial-corner attention feature transformation method in the distillation method with three existing distillation methods. These three methods are attention transfer (AT) [37], inter-channel correlation (IC) [42], and focal and global (FG) [43]. As shown in Table 3, our spatial-corner attention method has the best performance in five metrics, which is superior to other distillation methods based on feature transformation. CC is the correlation coefficient, which is used to measure the correlation between images. Its poor performance may be due to the fact that our distillation method makes students learn more knowledge about the model. This rich knowledge makes the fusion image different from the source image, but this knowledge also makes other metrics perform better. On the one hand, the use of spatial attention transformation can compress the knowledge of the teacher model into a more informative spatial domain, enabling the student model to imitate the feature extraction ability of the teacher network from a spatial perspective. On the other hand, the use of corner attention transformation can make the student model have a more efficient ability to capture image corners. Therefore, the combination of the two makes the knowledge transfer more comprehensive and efficient compared to other methods. Figure 6 shows the fusion results of the student models trained with different distillation methods, and it can be seen that our knowledge distillation method produces higher quality fusion results.

Table 1 Objective evaluation results of distillation model ablation experiments on test datasets

Component	Metrics								
	DT	SA	CA	EN	SSIM	SD	SF	AG	CC
✓				6.7268	0.4641	30.2756	10.0214	5.1488	0.4273
✓	✓			7.2125	0.4660	43.0672	12.6994	6.4844	0.4907
✓		✓		7.2005	0.4618	42.6772	12.7671	6.4528	0.4954
✓	✓	✓		7.3241	<u>0.4811</u>	<u>45.8887</u>	13.7887	6.7746	0.5076

Fig. 4 Qualitative results of distillation model ablation experiment

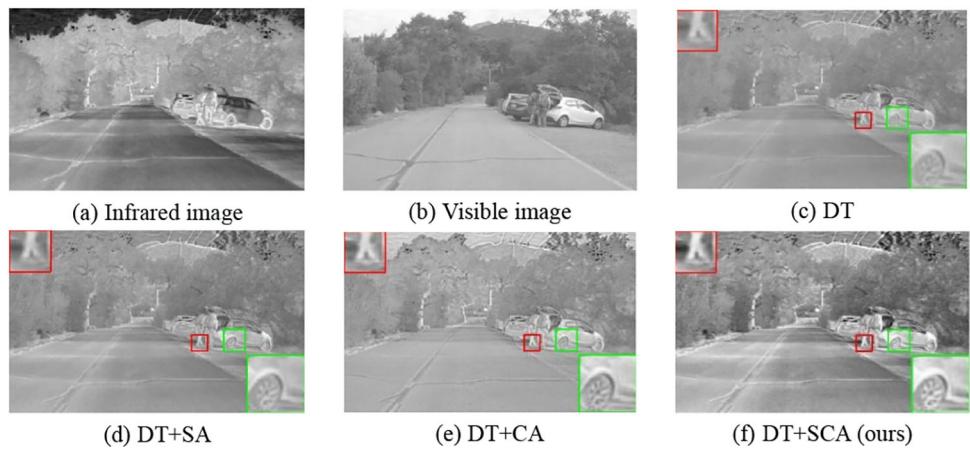


Fig. 5 Qualitative results of α ablation experiment

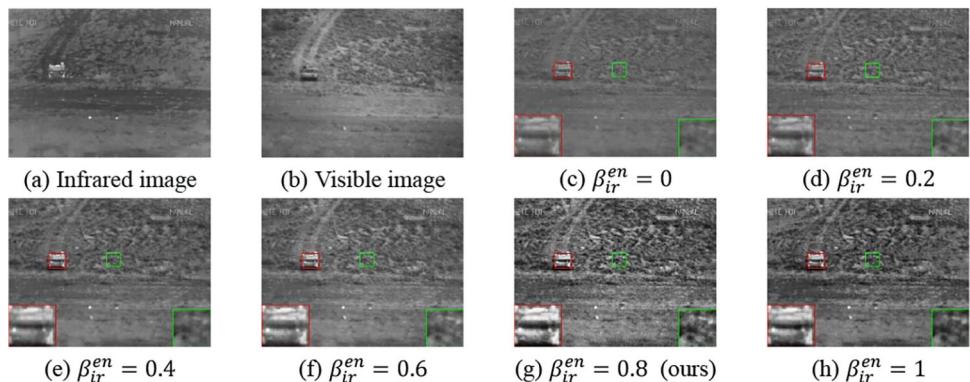


Table 2 Objective evaluation results of α ablation experiments on test datasets

$[\alpha_{ir}^{en} \setminus \alpha_{vis}^{en} \setminus \alpha^{de}]$	EN	SSIM	SD	SF	AG	CC
[0 \ 1 \ 0.5]	6.9803	0.3892	34.9481	9.9169	6.6099	0.4452
[0.2 \ 0.8 \ 0.5]	7.0189	0.4314	37.9688	10.0577	6.5561	0.4514
[0.4 \ 0.6 \ 0.5]	7.1184	0.4407	41.9615	11.3291	6.6561	0.4845
[0.6 \ 0.4 \ 0.5]	7.1955	0.4579	44.6184	<u>12.8558</u>	6.7704	<u>0.4931</u>
[0.8 \ 0.2 \ 0.5]	7.3241	<u>0.4811</u>	<u>45.8887</u>	13.7887	6.7746	0.5076
[1 \ 0 \ 0.5]	7.2564	0.4896	46.4266	12.7572	6.7188	0.4806

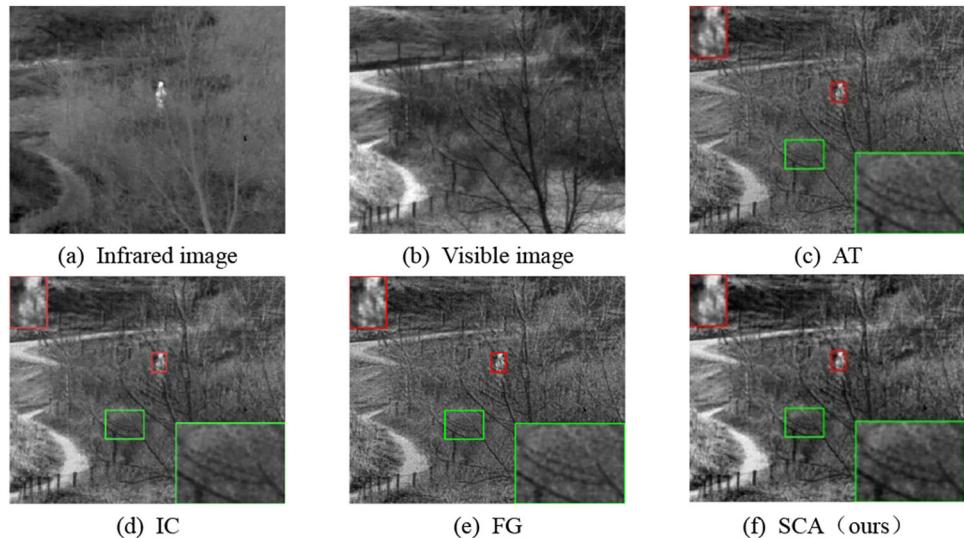
The maximum value in each metrics is represented in bold black, and the second highest value is represented with an underline

Table 3 Results of objective comparison with other distillation methods

Distillation methods	EN	SSIM	SD	SF	AG	CC
AT	7.2000	<u>0.4779</u>	43.9700	12.9300	<u>6.7433</u>	0.4998
IC	<u>7.2723</u>	0.4726	<u>44.0899</u>	12.7815	6.6364	0.5038
FG	7.2698	0.4773	44.0157	<u>13.1938</u>	6.7206	0.5191
SCA	7.3241	0.4811	45.8887	13.7887	6.7746	0.5076

The maximum value in each metrics is represented in bold black, and the second highest value is represented with an underline

Fig. 6 Qualitative comparison with other distillation methods



4.3 Comparative experiment of fusion methods

In order to demonstrate the superiority of the proposed method in the fusion of infrared and visible images, the student model after distillation is compared with FusionGAN [8], DenseFuse [7], GAN-FM [9], and UMF-CMGR [10] and the teacher model. All experiments are carried out in the same environment.

4.3.1 Qualitative comparison

As shown in Fig. 7, six pairs of representative image samples are selected for qualitative comparison of each method. The fusion results obtained by DenseFuse and UMF-CMGR methods have relatively rich texture details, but have lost a lot of thermal radiation information, as shown in the red boxes in row 4, column 2, and row 6, column 2, resulting in poor target detection performance due to darker brightness of the characters. Although the fusion images generated by FusionGAN and GAN-FM methods have a higher contrast between infrared targets and background, there are many artifacts, as shown in the red boxes in row 3, column 3, and row 5, column 3, where objects such as leaves and road signs are somewhat blurry and lack details. In comparison, the fusion results generated by the teacher model not only contain rich texture details, but also have significant contrast. It shows that

the proposed teacher model has more advanced fusion effect than the existing methods. As shown in the last two rows in Fig. 7, the fusion results of the student model have almost no visual difference compared to those of the teacher model. Based on these comparisons, it can be seen that the student model has significant advantages in visual effects compared to existing methods.

4.3.2 Quantitative comparison

Further quantitative comparison experiments are conducted on all parties. The average values of the six fusion evaluation metrics obtained by all methods are shown in Table 4. The quantitative results further show that the proposed teacher model has better comprehensive performance than the existing methods, and the results of the proposed student model are similar to the results of the teacher model, with three best average values and three second-best average values, indicating the superiority of our method in the task of fusing infrared and visible images, which is attributed to the excellent knowledge transfer ability of our knowledge distillation scheme. At the same time, we notice that the results of the student model are even slightly higher than those of the teacher model in some indicators, which indicates that our distillation method based on spatial-corner attention features extracts many dark knowledge [35] from the teacher model,



Fig. 7 Qualitative comparison of the student model with four existing methods and the teacher model. Note that we zoom in two regions (red box for high-contrast region and green box for visible texture) and put them in the two corners

which contains more valuable information than the teacher model itself. Therefore, our student model can achieve better fusion performance.

4.3.3 Complexity evaluation

In order to demonstrate the effectiveness of our proposed method, we conduct experiments comparing model complexity. As shown in Table 5, we calculate the model sizes and computational costs of each method, where model size is measured by the number of network parameters, and com-

putational cost is measured by the number of float point operations (FLOPs). It should be noted that for methods based on GAN, we only consider the generator's parameters and computational cost, because model inference only requires the well-trained generator, not the discriminator. The results show that the FLOPs of our student model are 14.08G, with 0.86M parameters. Compared to the teacher model, the FLOPs and parameters of the student model are reduced by approximately 45%, with the model complexity lower than other fusion methods except for FusionGAN. Although the complexity of our student model is not the lowest compared

Table 4 Quantitative comparison of the student model with four existing methods and the teacher model

Fusion methods	EN	SSIM	SD	SF	AG	CC
FusionGAN	6.8636	0.3753	35.3545	7.2782	3.5449	0.4117
DenseFuse	6.5674	0.4198	26.3270	7.3032	3.6111	0.5150
GAN-FM	7.3544	0.4438	45.6872	<u>13.7567</u>	6.7024	0.4616
UMF-CMGR	6.8225	0.4329	32.1160	9.3987	4.4593	0.4913
Teacher	7.3220	0.4849	<u>45.8790</u>	12.9300	<u>6.7433</u>	0.4998
Student	<u>7.3241</u>	<u>0.4811</u>	45.8887	13.7887	6.7746	<u>0.5076</u>

The maximum value in each metrics is represented in bold black, and the second highest value is represented with an underline

Table 5 Model complexity of different fusion methods

Fusion methods	FLOPs (G)	Params (M)
FusionGAN	1.16	0.07
DenseFuse	15.17	0.93
GAN-FM	167.95	10.21
UMF-CMGR	14.78	0.90
Teacher	25.94	1.58
Student	14.08	0.86

to FusionGAN, the fusion effect of FusionGAN has been qualitatively and quantitatively proven to be far inferior to our student model. In summary, our proposed SCAKD method has significant effects in reducing resource costs and improving efficiency.

5 Conclusion

This paper proposes a spatial-corner attention-based knowledge distillation (SCAKD) framework for infrared and visible image fusion tasks. In our SCAKD, we designed knowledge based on the characteristics of infrared and visible fusion tasks, considering the correlation between pixels in the feature map from the perspective of space and the distribution of corners in the feature map from the perspective of texture details. In order to facilitate effective transfer of knowledge from the teacher model to the student model, we extracted knowledge from different layers of the teacher model to guide the generation of corresponding features in the student model. Experimental results demonstrate that our distillation framework enables the student model, with lower number of parameters and computational cost, to achieve higher fusion performance.

Acknowledgements This work was supported by the National Key R&D Program of China (Grant No: 2020YFB2008403).

Data Availability Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

Declarations

Conflict of interest The authors declare no conflicts of interest.

References

1. Bajammal, M., Stocco, A., Mazinanian, D., Mesbah, A.: A survey on the use of computer vision to improve software engineering tasks. *IEEE Trans. Software Eng.* **48**(5), 1722–1742 (2020)
2. Ma, J., Ma, Y., Li, C.: Infrared and visible image fusion methods and applications: A survey. *Inf. Fusion* **45**, 153–178 (2019)
3. Peng, P., Geng, K., Li, S., Wang, Z., Qian, M., and Yin, G.: Sharpening mixture of experts fusion of infrared and visible images for night perception enhancement. In: 2021 5th CAA International Conference on Vehicular Control and Intelligence (CVCI), pp. 1–4. IEEE (2021)
4. Deng, L., Pan, M., Jin, R., and Xie, Z.: Night target detection approach based on near infrared image fusion on vehicles. In: 2022 5th international conference on pattern recognition and artificial intelligence (PRAI), pp. 755–759. IEEE (2022)
5. Zhang, H., Ma, J.: Sdnet: a versatile squeeze-and-decomposition network for real-time image fusion. *Int. J. Comput. Vis.* **129**, 2761–2785 (2021)
6. Xu, H., Ma, J., Jiang, J., Guo, X., Ling, H.: U2fusion: a unified unsupervised image fusion network. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**(1), 502–518 (2020)
7. Li, H., Wu, X.-J.: Densefuse: a fusion approach to infrared and visible images. *IEEE Trans. Image Process.* **28**(5), 2614–2623 (2018)
8. Ma, J., Yu, W., Liang, P., Li, C., Jiang, J.: Fusiongan: a generative adversarial network for infrared and visible image fusion. *Inf. Fusion* **48**, 11–26 (2019)
9. Zhang, H., Yuan, J., Tian, X., Ma, J.: Gan-fm: infrared and visible image fusion using gan with full-scale skip connection and dual Markovian discriminators. *IEEE Trans. Comput. Imaging* **7**, 1134–1147 (2021)
10. Wang, D., Liu, J., Fan, X., & Liu, R.: Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration. arXiv preprint [arXiv:2205.11876](https://arxiv.org/abs/2205.11876) (2022)
11. Ma, J., Chen, C., Li, C., Huang, J.: Infrared and visible image fusion via gradient transfer and total variation minimization. *Inf. Fusion* **31**, 100–109 (2016)
12. Zhao, J., Cui, G., Gong, X., Zang, Y., Tao, S., Wang, D.: Fusion of visible and infrared images using global entropy and gradient constrained regularization. *Infrared Phys. Technol.* **81**, 201–209 (2017)
13. Zhao, J., Chen, Y., Feng, H., Xu, Z., Li, Q.: Infrared image enhancement through saliency feature analysis based on multi-scale decomposition. *Infrared Phys. Technol.* **62**, 86–93 (2014)

14. Bavirisetti, D.P., Xiao, G., Liu, G.: Multi-sensor image fusion based on fourth order partial differential equations. In: 20th International Conference on Information Fusion (Fusion), pp. 1–9. IEEE (2017)
15. Li, G., Zhang, M., Li, J., Lv, F., Tong, G.: Efficient densely connected convolutional neural networks. *Pattern Recogn.* **109**, 107610 (2021)
16. Himeur, C.-E., Lejemble, T., Pellegrini, T., Paulin, M., Barthe, L., Mellado, N.: Pcednet: a lightweight neural network for fast and interactive edge detection in 3d point clouds. *ACM Trans. Graph. (TOG)* **41**(1), 1–21 (2021)
17. Jang, Y., Lee, S., Kim, J.: Compressing convolutional neural networks by pruning density peak filters. *IEEE Access* **9**, 8278–8285 (2021)
18. Ruan, X., Liu, Y., Li, B., Yuan, C., Hu, W.: Dpfps: dynamic and progressive filter pruning for compressing convolutional neural networks from scratch. *Proc. AAAI Conf. Artif. Intell.* **35**(3), 2495–2503 (2021)
19. Yim, J., Joo, D., Bae, J., and Kim, J.: A gift from knowledge distillation: fast optimization, network minimization and transfer learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4133–4141 (2017)
20. Gao, Q., Zhao, Y., Li, G., and Tong, T.: Image super-resolution using knowledge distillation. In: Computer Vision—ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part II, pp. 527–541. Springer (2019)
21. Gou, J., Yu, B., Maybank, S.J., Tao, D.: Knowledge distillation: a survey. *Int. J. Comput. Vis.* **129**, 1789–1819 (2021)
22. Tan, C., Liu, J.: Online knowledge distillation with elastic peer. *Inf. Sci.* **583**, 1–13 (2022)
23. Park, T., Liu, M.-Y., Wang, T.-C., and Zhu, J.-Y.: Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 2337–2346 (2019)
24. Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., and Hu, Q.: Eca-net: efficient channel attention for deep convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11534–11542 (2020)
25. Wang, J., Peng, J., Feng, X., He, G., Fan, J.: Fusion method for infrared and visible images by using non-negative sparse representation. *Infrared Phys. Technol.* **67**, 477–489 (2014)
26. Li, H., Wang, Y., Yang, Z., Wang, R., Li, X., Tao, D.: Discriminative dictionary learning-based multiple component decomposition for detail-preserving noisy image fusion. *IEEE Trans. Instrum. Meas.* **69**(4), 1082–1102 (2019)
27. Li, S., Kang, X., Fang, L., Hu, J., Yin, H.: Pixel-level image fusion: a survey of the state of the art. *Inf. Fusion* **33**, 100–112 (2017)
28. Zhu, Z., Zheng, M., Qi, G., Wang, D., Xiang, Y.: A phase congruency and local Laplacian energy based multi-modality medical image fusion method in NSCT domain. *IEEE Access* **7**, 20811–20824 (2019)
29. Zhang, X., Ma, Y., Fan, F., Zhang, Y., Huang, J.: Infrared and visible image fusion via saliency analysis and local edge-preserving multi-scale decomposition. *JOSA A* **34**(8), 1400–1410 (2017)
30. Yan, H., Yu, X., Zhang, Y., Zhang, S., Zhao, X., Zhang, L.: Single image depth estimation with normal guided scale invariant deep convolutional fields. *IEEE Trans. Circuits Syst. Video Technol.* **29**(1), 80–92 (2017)
31. Li, H., Wu, X.-J., Durrani, T.S.: Infrared and visible image fusion with resnet and zero-phase component analysis. *Infrared Phys. Technol.* **102**, 103039 (2019)
32. Zhang, Y., Liu, Y., Sun, P., Yan, H., Zhao, X., Zhang, L.: Ifcnn: a general image fusion framework based on convolutional neural network. *Inf. Fusion* **54**, 99–118 (2020)
33. Ma, J., Zhang, H., Shao, Z., Liang, P., Xu, H.: Ganmcc: a generative adversarial network with multiclassification constraints for infrared and visible image fusion. *IEEE Trans. Instrum. Meas.* **70**, 1–14 (2020)
34. Ba, L.J., & Caruana, R.: Do deep nets really need to be deep? In: Proceedings of the 27th International Conference on Neural Information Processing Systems, vol. 2, pp. 2654–2662 (2014)
35. Hinton, G., Vinyals, O., and Dean, J.: Distilling the knowledge in a neural network. arXiv preprint [arXiv:1503.02531](https://arxiv.org/abs/1503.02531) (2015)
36. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., & Bengio, Y.: Fitnets: hints for thin deep nets. arXiv preprint [arXiv:1412.6550](https://arxiv.org/abs/1412.6550) (2014)
37. Zagoruyko, S., & Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. arXiv preprint [arXiv:1612.03928](https://arxiv.org/abs/1612.03928) (2016)
38. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., & Fu, Y.: Residual dense network for image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2472–2481 (2018)
39. Zhang, F., Zhu, X., and Ye, M.: Fast human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3517–3526 (2019)
40. Saputra, M.R.U., De Gusmao, P.P., Almaliooglu, Y., Markham, A., & Trigoni, N.: Distilling knowledge from a deep pose regressor network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 263–272 (2019)
41. Toet, A.: The TNO multiband image data collection. *Data Brief* **15**, 249–251 (2017)
42. Liu, L., Huang, Q., Lin, S., Xie, H., Wang, B., Chang, X., and Liang, X.: Exploring inter-channel correlation for diversity-preserved knowledge distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8271–8280 (2021)
43. Yang, Z., Li, Z., Jiang, X., Gong, Y., Yuan, Z., Zhao, D., and Yuan, C.: Focal and global knowledge distillation for detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4643–4652 (2022)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.



Jiakun Zhao received the B.S. degree in Mathematics from Xuzhou Normal University, Xuzhou, China, in 1995, M.S. in Applied Mathematics from Xi'an Jiaotong University in 2002 and Ph.D. in Computational Mathematics from Xi'an Jiaotong University in 2010. He is currently a Full Professor of software engineering at Xi'an Jiaotong University. His research interests focus on analysis and applications of big data, automated machine learning and deep learning.



Yige Cai is currently working toward the M.S. degree in Software Engineering at Xi'an Jiaotong University, China. He received the B.S. degree in Communication Engineering from Qingdao University of Science and Technology, China, in 2015. His research interests focus on image processing techniques, computer vision and knowledge distillation.