Light-weight Fusion Network for UAV Visible-light and Infrared Images based on Real-time Brain-like Intelligent Computing

Kun Hu[1,*], Jiayi Qu[1], Wenhao Zheng[1], Haoyang Zhou[1], Tianyu Cai[2], Qingle Zhang[1], Shichao Wang[3], Bo Li[3]

[1]Institute of Artificial Intelligence, Beihang University, 100191, Beijing, China - (kunhu, 21373177, 21371374, 21371169, qlzhang)@buaa.edu.cn
[2]The University of Sydney, Camperdown NSW 2050, Sydney, Australia - tcai7097@uni.sydney.edu.au
[3]Geovis Technology Co.,Ltd, 101399, Beijing, China - (wangsc01, lib)@geovis.com.cn

Keywords: Unmanned Aerial Vehicle (UAV) , Multi-modal image fusion, Light-weight algorithm, Real-time computing.

Abstract

Multiple sensors equipped on the Unmanned Aerial Vehicle (UAV) enables the acquisition of multi-modal and multi-source remote sensing data. UAV remote sensing usually faces with real-time or near-real-time tasks in complex and highly dynamic environments, such as disaster monitoring, traffic management, border patrol and so on. Under these conditions, the image fusion algorithm needs to be high efficiency, precision and reliability. In this paper, we proposed an intelligent real-time fusion network for UAV multi-source remote sensing data based on AI brain-like chips, and deployed the algorithm on the UAV platform to achieve online high-efficiency computing. Firstly, we have developed a novel image fusion algorithm named SFNet for infrared and visible image fusion based on ShuffleNetv2. Then, we use ZCA and ℓ_1-norm to process the remodeled deep feature. The weight maps are generated by bi-cubic interpolation and soft-max operation. Finally, the fused image is reconstructed by weighted-average operation. The proposed SFNet is deployed on the Lynxi KA200 brain computing chip, and a comprehensive inference test is carried out with UAV remote sensing data. Several State-Of-The-Art (SOTA) data fusion algorithms are deployed on the same chip for experimental comparison. The proposed SFNet is proved to have faster inference speed and better feature extraction results on brain-like chips. It is more suitable for real-time UAV remote sensing image fusion tasks.

1. Introduction

Image fusion, particularly Infrared and Visible Image Fusion (IVIF), is the process of integrating information from multiple images of the same scene captured by different sensors or imaging modalities into a single composite image. The objective of image fusion is to enhance the overall visibility, improve image quality, and extract more comprehensive information than can be achieved with individual images alone. In the context of visible and infrared image fusion, it involves combining the complementary information provided by both modalities to generate a fused image that reveals details not readily visible in either the visible or infrared images alone. As an essential component of computer vision, IVIF solutions are extensively researched and categorized based on adopted theories into five categories: multi-scale transformation-based methods, sparse representation-based methods, subspace decomposition-based methods, hybrid-based methods, and optimization model-based methods.

Convolutional Neural Networks (CNNs) have demonstrated remarkable success in the field of IVIF. CNN-based approaches offer numerous advantages for image fusion tasks, such as automatic feature learning, end-to-end optimization, and the ability to capture intricate spatial dependencies within images. Through training CNNs on extensive datasets containing paired visible and infrared images, these models can effectively learn to extract and fuse relevant features from both modalities, resulting in fused images with enhanced visibility, improved image quality, and better preservation of important details.

CNN-based methods for IVIF have proven highly effective in a wide range of applications, such as target detection, surveillance, remote sensing, and medical imaging. These methods have excelled in addressing challenges, such as modality alignment, variations in illumination conditions, and differences in scene content. Additionally, CNN-based fusion methods can be customized to meet specific application needs by modifying network architectures, loss functions, and training strategies. This adaptability further enhances their performance across diverse scenarios.

IVIF plays a crucial role in unmanned aerial vehicle (UAV) remote sensing as it offers complementary information about the observed environment. Visible imagery captures surface details such as color, texture, and shape, while infrared imagery captures thermal radiation emitted or reflected by objects, providing insights into temperature distribution and material composition. By fusing these modalities, UAV remote sensing systems can overcome limitations associated with each individual modality and gain a more comprehensive understanding of the observed scene.

In the context of UAV applications, such as environmental monitoring, agricultural assessment, and surveillance, the fusion of visible and infrared images enables more comprehensive analysis and interpretation of the captured data. For example, in agriculture, the fusion of visible and infrared imagery can facilitate crop health monitoring, detection of irrigation issues, and assessment of soil moisture levels. In environmental monitoring, it can aid in the identification of land cover types, detection of vegetation stress, and mapping of ecological parameters. Additionally, in surveillance and security applications, fused imagery can improve target detection, enhance situational awareness, and support decision-making processes.

Overall, the fusion of visible and infrared images in UAV remote sensing tasks enhances the effectiveness and efficiency of data collection, analysis, and interpretation, leading to better-informed decision-making and improved outcomes in various fields.

[*] Corresponding author

https://doi.org/10.5194/isprs-archives-XLVIII-1-2024-225-2024 | © Author(s) 2024. CC BY 4.0 License.

Therefore,we propose a light-weight image fusion network. Our method realizes the simplification of the model by transferring some model parameters of ShuffleNetv2. Meanwhile,our SFNet can efficiently extract infrared and visible image features, and then obtain high quality fusion results.

In this paper, experiments are carried out on public image data sets and compared with other 7 methods. The experimental results show that the proposed method can effectively improve the quality and efficiency of image fusion.

2. Related Works

2.1 Traditional Image Fusion Methods

In the recent years, extensive traditional infrared and visible image fusion methods are proposed and applied well. Many splendid theories In all these traditional methods,multi-scale transform(MST)-based methods are used widely.

MST methods consider that objects in the physical world are typically composed of components of various scales, and the multi-scale transform is consistent with the human visual system. Therefore, the fused images obtained by MST have pleasing visual effect. MST achieved outstanding fusion performance thanks to the design of diverse transformation tools such as wavelet transform(Petrovic et al,2004),non-subsampled contourlet trans-form(Bhatnagar G et al,2013), edge-preserving filter based transform, and Retinex theory-based transform, to extract features at different scales. After this, we decompose the source image into a set of base images. Then apply Singular Value Decomposition(SVD) to each base image to extract its singular values and corresponding singular vectors.Next, select the most informative singular values from each base image.

Now focus on subspace-based(SR based) image fusion. In image fusion, the idea is to decompose the source images into different subspaces and then fuse these subspaces to create a composite image that preserves the most relevant information from each source.Kim M et al(2016) proposed a fusion method based on patch clustering. The authors cluster patches from different sources with their structural similarities.This learning method is called a clustering-based dictionary learning. Sparse coefficients are estimated by a simultaneous orthogonal matching pursuit. And their proposed method requires lower processing time with better fusion quality.To sum up,SR based methods target to construct an over-complete dictionary from high-quality natural images.

The core idea of model-based methods is to choose suitable models. The authors proposed a new fusion method based on gradient transfer and TV minimization. It can keep both the thermal radiation and the appearance information in the source images.More recently,Liu et al(2021) propose a generic image fusion method with a bilevel optimization paradigm, targeting on multi-modality image fusion tasks. Corresponding alternation optimization is conducted on certain components decoupled from source images. Via adaptive integration weight maps, we are able to get the flexible fusion strategy across multi-modality images.

Although the aforementioned IVIF methods have a lot of success,they have some common drawbacks, such as the loss of information, limited adaptability, artifacts and distortions.

2.2 Deep Learning-Based Fusion Methods

With strong non-linear fitting and feature learning abilities of the neural network, deep learning technique has achieved significant advances in image fusion.At first, deep learning is only employed in feature extraction or weight-map generation. An typical example is adopting two pretrained CNN models to get two weight maps, so they can be used to merge the base and detail layer. It is explicit that the overall process still has limitations.

Recently, A new architecture--auto encoder model has emerged. Feature extraction and feature reconstruction are realized by this architecture, in which the fusion rules are designed manually. In the encoder part, a dense block is integrated, so the feature can be extracted comprehensively. Then use addition and ℓ_1-norm rule in the fusion layer to generate fused results. Considering that vital information often degenerates from the network, the author employed different reception dilated convolutions to extract feature from a multi-scale prospective, and then employ edge attention mechanism to refine the details.

Li and Wu(2018) present a novel deep learning architecture for infrared and visible images fusion problems. Their encoding network is combined with convolutional layers, a fusion layer, and dense block in which the output of each layer is connected to every other layer. Two fusion layers (fusion strategies) are designed to fuse these features. The proposed fusion method achieves the state-of-the-art performance in objective and subjective assessment.

Moreover, extensive generative adversarial network(GAN)-based fusion methods has superior unsupervised distribution estimation ability, which well suits to IVIF.An adversarial game between the visible image and fused results, aiming to enhance the textural details. As an attempt, Li et al (2021) introduced an end-to-end GAN model that integrates multi-classification constraints.More recently,Li et al(2023)first apply a multi-scale extractor to achieve shallow features, which are employed as the necessary input to build graph structures. Then construct the extracted intermediate features of the infrared/visible branch into graph structures. Besides, the proposed leader nodes can improve information propagation in the same modality. Finally, we merge all graph features to get the fusion result.

As a hot topic in computer vision, Vision Transformer(VIT) is also designed to fuse the images. The patch embeddings along with positional encodings are then passed through transformer encoder layers. V.Vs et al(2021) proposed a method that follows a two-stage training approach. Firstly train an auto-encoder to extract deep features at multiple scales. Secondly, multi-scale features are fused using a Spatio-Transformer (ST) fusion strategy. The ST fusion blocks capture local and long-range features, respectively.Extensive experiments show that the proposed method performs better than many competitive fusion algorithms.

Moreover,Liu Z et al(2023) established the hierarchical dual tasks-driven deep model to bridge these tasks. They construct an image fusion module to fuse complementary characteristics and cascade dual task-related modules. They provide a bi-level perspective to formulate image fusion and follow-up downstream tasks. An efficient first-order approximation is developed to compute corresponding gradients and present dynamic weighted aggregation to balance the gradients for fusion learning.

The fusion task is formulated as a conditional generation problem under the proposed sampling framework by Z. Zhao et al.(2023), which is further divided into an unconditional generation sub-problem and a maximum likelihood sub-problem. The latter is modeled in a hierarchical Bayesian manner with latent variables and inferred by the expectation-maximization (EM) algorithm. So it is worth noting that diffusion models can also be utilized for IVIF tasks. And we are actively looking for new perspectives in the direction of image fusion.

3. The Proposed Method

3.1 Motivation

With the development of remote sensing technology and the continuous maturity of computer vision field,tasks about real-time image fusion are booming.Deep learning techniques offer new ideas to MMIF.There are so many brilliant network frameworks that we can use.However,sizeable networks only use direct features,which means feature extractions are not participated,leading to the fusion performance degradation.But we are glad to see there are Image Fusion with network and zero-phase component analysis.In that article,authors chose ResNet50.Although it is a nice light-weight network, its not fit for the research of real-time infrared and visible image fusion on UAV. Thinking about the later work of AI chip deployment,we decide to follow the practical guidelines,and use ShuffleNetv2 to achieve real-time image fusion with light-weight network.

3.2 Network Architecture

We have been aware that infrared images and visible light are fed separately into the ShuffleNetv2,and with the process of feature extraction,images fusion and weighted fusion,ultimately obtain the fused images.The architecture is shown in Fig.1.

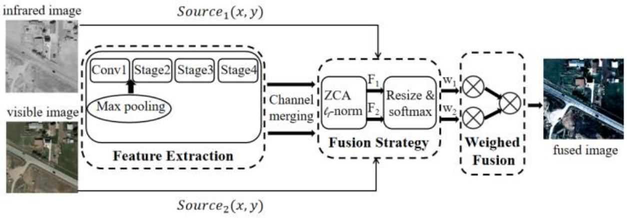


Figure 1: The whole architecture of our network SFNet

3.3 The Proposed SFNet

The infrared and visible images are called $Scource_1$ and $Scource_2$ respectively.ShuffleNetv2 typically includes several 1x1 convolutional layers, 3x3 convolutional layers, as well as depthwise separable convolution layers and standard convolutional layers.In our SFNet,we use 5 convolution blocks(from conv1 to conv5).The output of the i-th block is shown by the deep features $S_k^{i,1:C}$,which contain C channels. Then all the channels are combined into a tensor,so we have less time consumption. We use ZCA and ℓ_r-norm to process the remodeled deep feature. By bicubic interpolation and soft-max operation, the weight maps w_k is ready. Finally by using weighted-average operation, the fused image is reconstructed, which has the same dimension to the origin one.

3.3.1 Model Migration:ShuffleNetv2

When we evaluate a network structure, accuracy is always the most important.But besides this,computation complexity is also worth noting. As AI technology has made huge progress, they are closer to the real world tasks, which often aim at obtaining better accuracy under a limited time and calculation environment. And that's why we design a light-weight network.We chose ShuffleNetV2, which starts from practice and is guided by the actual inference speed.

In ShuffleNetv1's module, 1x1 group convolution is used extensively, and v1 uses a bottleneck layer similar to ResNet's, with different input and output channels. In order to improve v1's shortcomings, v2 introduced a new operation: channel split. Specifically, at the beginning, the input feature graph is divided into two branches in the channel dimension, and the number of channels is half of the original. The left branch is equally mapped, and the right branch contains three continuous convolution, and the input and output channels are the same. The other two branches have been divided into two groups, their output is no longer Add elements, but concat together, followed by a channel shuffle of the concat results of the two branches to ensure that the two branches communicate information.

ShuffleNetv2 summarizes 5 design essentials of lightweight network, and proposes ShuffleNetV2 according to the essentials, which gives a good balance to accuracy and speed. Among them, the channel split operation is very bright, and the input features are divided into two parts.

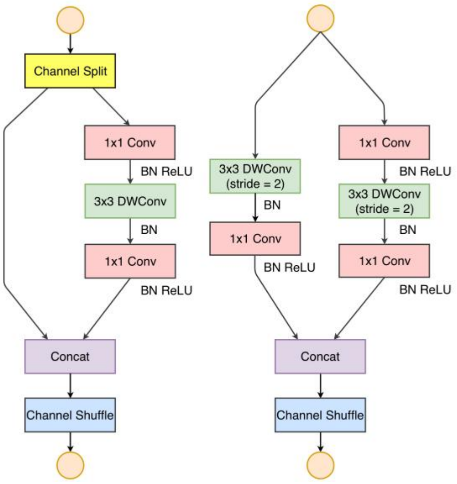


Figure 2:Units in ShuffleNetv2

Table 1:Overall architecture of ShuffleNetv2

Layer	Output size	KSize	Stride	Repeat	Output channel
Image	224×224				3
Conv1 MaxPool	112×112 56×56	3×3 3×3	2 2	1	24
Stage2	28×28 28×28		2 1	1 3	116
Stage3	14×14 14×14		2 1	1 7	232
Stage4	7×7 7×7		2 1	1 3	464
Conv5	7×7	1×1	1	1	1024

Above guidelines and empirical studies,authors(Ma et al.,2018) conclude that an efficient network architecture should :
1) use convolutions that hold equal channel width;
2) be aware of the cost of using group convolution;
3) reduce the degree of fragmentation;
4) reduce element-wise operations.

These desirable properties depend on platform characteristics (such as memory manipulation and code optimization) that are beyond theoretical FLOPs. They should be taken into account for practical network design.
As our tasks based on UAV real-time image fusion, speed is a direct metric to be considered.Choosing ShuffleNetV2 could better meet the needs.We also make some progress to adapt to remote sensing multi-model images.We will introduce the proposed network in the next section.

3.3.2 Zero-phase Component Analysis(ZCA):The whitening and decorrelation by ZCA operation is analyzed by Kessy et al. ZCA is a preprocessing technique commonly used in machine learning and signal processing to decorrelate and normalize data. Its primary goal is to transform the data into a new space where the covariance matrix is the identity matrix, making the features statistically independent and better suited for learning algorithms. So let us briefly introduce the concrete operation.

The d-dimensional random vector is shown by Eq.1.

$$\mathrm{X}=(x_1, x_2, \ldots, x_d)^T \quad (1)$$

And the mean values are shown by Eq.2.

$$\mathrm{u}=(u_1, u_2, \ldots, u_d)^T \quad (2)$$

The covariance matrix Co is calculated bt Eq.3.

$$\mathrm{Co} = (\mathrm{X}-\mathrm{u}) \times (\mathrm{X}-\mathrm{u})^{\mathrm{T}} \quad (3)$$

Then utilize the Singular Value Decomposition(SVD) in Eq.4.

$$\mathrm{Co}([U, \Lambda, V]=\mathrm{SVD}(\mathrm{Co})) \quad (4)$$

so Co is calculated by Eq.5.

$$\mathrm{Co}= (U\Lambda V)^{\mathrm{T}} \quad (5)$$

Finally,get the new random vector $\widehat{X} = \mathrm{U}(\Lambda + \epsilon I)^{-\frac{1}{2}}\mathrm{U}^{\mathrm{T}} \times \mathrm{X}$, where ϵ is a small value to avoid bad matrix inversion,and I means the identity matrix.

3.3.3 ZCA and l_1-norm Operations:ZCA can project the raw features into the same space,and these features benefit more to later work.In our SFNet,we choose ShuffleNet v2 as our backbone.The feature maps produced by stage2,stage3 and stage4 are used to fuse the images.

After generating the deep features,we process these deep features F_k^i by ZCA operation,i =1,2,3,4,5,i indicates the i-th convolutional block.So we get the processed deep features $\widehat{F_k^i}$.

Let's see the calculation in ZCA,which is shown by Eq.6.

$$Co_k^i = F_k^i \times \left(F_k^i\right)^{\mathrm{T}},$$
$$Co_k^i= (U\Lambda V)^{\mathrm{T}} \quad (6)$$

So as mentioned,$\widehat{F_k^i}$=$s_k^i \times F_k^i$,and s_k^i is shown in Eq.7.

$$\mathrm{s_k^i} = \mathrm{U}(\Lambda + \epsilon I)^{-\frac{1}{2}}\mathrm{U}^{\mathrm{T}} \quad (7)$$

Finally,we utilize local l1-norm and average operation,so the original weight maps S_k^i is calculated by Eq.8.

$$S_k^i = \frac{\sum_{p=x-s}^{x+s}\sum_{q=y-s}^{y+s}\left\|\widehat{F_k^i}(p,q)\right\|}{(2s+1)\times(2s+1)} \quad (8)$$

3.3.4 Reconstruction:We have get the infrared and visible images' original weight maps F_1^i and F_2^i,we utilize upsampling and soft-max operations to obtain the final weight maps w_1^i and w_2^i,as is shown in Fig.3.

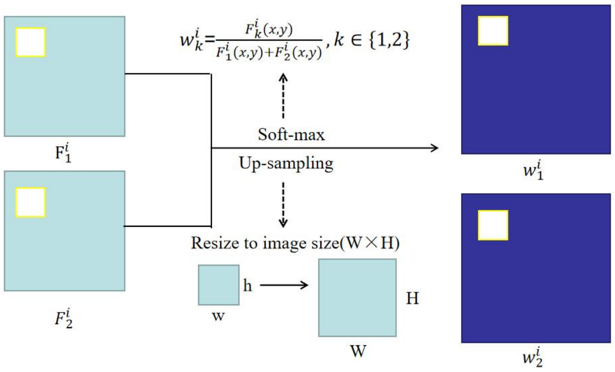

w_k^i=$\frac{F_k^i(x,y)}{F_1^i(x,y)+F_2^i(x,y)}$, $k \in \{1,2\}$
Soft-max
Up-sampling
Resize to image size(W×H)
h
w
H
W
F_1^i
F_2^i
w_1^i
w_2^i

Figure 3: Resize and soft-max operation and the formula

So we obtain the fused images in Eq.9.

$$\mathrm{Fused}(\mathrm{x},\mathrm{y}) = \sum_{k=1}^{2} w_k^i(x,y)Source_k(x,y) \quad (9)$$

4. Experiments

4.1 Experiments settings

4.1.1 Datasets:We choose VEDAI datasets.It is a widely used benchmark datasets in the field of computer vision and remote sensing,specifically designed for vehicle detection in aerial imagery. VEDAI contains high-resolution aerial images captured from different sensors and platforms, covering various geographic locations and environmental conditions.

VEDAI comprises images captured from different altitudes, angles, and lighting conditions, providing a diverse set of visual data for training and testing vehicle detection algorithms.And each image in the VEDAI datasets is annotated with bounding boxes around vehicles, along with corresponding class labels, such as car, truck, van, etc. These annotations facilitate the evaluation and benchmarking of object detection models.The images also have high spatial resolution, enabling the detection of vehicles with fine details even from aerial viewpoints. VEDAI is commonly used for developing and evaluating vehicle detection algorithms, especially in scenarios where aerial surveillance or monitoring is required, such as traffic management, urban planning, and environmental monitoring.

4.1.2 Evaluation Metrics:We choose four common metrics ,which play crucial roles in assessing the quality and performance of image fusion algorithms, helping researchers and practitioners make informed decisions regarding algorithm design and parameter tuning.

Average gradient:It is often used as an objective metric to evaluate the performance of various image fusion algorithms.It provides a quantitative measure of the fusion quality, helping researchers compare different fusion techniques and optimize parameters.A larger average gradient can be considered indicative of better image clarity and fusion quality.

Spatial frequency:In the field of image fusion, spatial frequency is a crucial evaluation metric. It refers to the rate of change of intensity or color information in an image across space. High spatial frequency indicates rapid changes and fine details, while low spatial frequency suggests slower changes and broader features. This metric helps assess the level of detail and information preservation in fused images.

Standard Deviation: It measures the amount of variation or dispersion of pixel values within an image. In image fusion evaluation, standard deviation is often utilized to assess the level of contrast and detail preservation in the fused image. Higher standard deviation values indicate greater variability in pixel intensities, which may suggest better preservation of image details.

Correlation Coefficient: It quantifies the degree of linear relationship between pixel values of two images. In image fusion evaluation, correlation coefficient is employed to determine the similarity or consistency between the fused image and the source images. A high correlation coefficient indicates strong similarity, implying that the fused image effectively retains information from the source images.

4.2 Ablation experiment

ShuffleNetv2 consists of 3 Stage models, which is Stage2, Stage3 and Stage4, and some pooling layers and convolution layers. However, when carrying out image fusion, only a certain layer in the ShuffleNetv2 structure needs to be selected as the output of the whole feature layer. The choice of which of the three stages has a crucial impact on the accuracy and effect of the network. Therefore, we carried out image fusion for each output feature map of the three output layers from stage2 to stage4, and obtained the average image fusion results of the different three output layers in Tab2. For the highest data of each indicator, we made the font bold, and the second highest data was underlined.

Table 2:Fusion quality comparison of different layer outputs

Output layers	AG	SF	STD	CC
Stage2	21.1286	**48.3601**	72.0311	0.7876
Stage3	21.1863	48.3146	74.3925	0.7935
Stage4	**21.2448**	48.2539	**74.6924**	**0.8028**

From the data in the above table, we know that Stage4 has the highest image fusion quality,meaning that our work is effective. Stage4 has the highest value of AG, indicating the best image clarity. The three stages have similar performance in SF, showing that the processing in these stages preserves as much details of the image as possible.From Stage2 to Stage4, the value of STD and CC becomes higher and higher, indicating that the level of contrast and detail preservation and similarity between the infrared and visible images in the fused image is getting better,therefore the image fusion strategy of our network is undoubtedly effective.

4.3 Fusion Performance Evaluation

Among the above four indicators we selected, our network SFNet performed well.We ran several superb baselines.Because we want to build a light-weight image fusion network based on deep learning, we do not choose a network using traditional fusion methods, but directly compare it with well-known deep learning-based image fusion networks.Let us introduce the seven methods we compared with.

CoConet (A Collaborative Convolutional Network,Liu,2022) is a convolutional neural network designed for multi-spectral and panch-romatic image fusion.It operates in a coarse-to-fine manner, gradually refining the fused image at multiple scales. The network architecture incorporates skip connections to facilitate information flow between different layers.

Reconet (Residual Convolutional Neural Network,Gao,2018) leverages residual learning to ease the training process and improve convergence.The network architecture emphasizes feature reuse and propagation to enhance fusion performance.

CBF (Convolutional Block Fusion,Faye,2024) utilizes convolutional blocks to capture multi-scale features from input images.The network architecture focuses on adaptability and scalability across various fusion scenarios.

SeAFusion (Selective Attention Fusion Network, Linfeng Tang,2022) selectively attends to informative regions in the input images, enhancing fusion quality.It incorporates attention mechanisms at different stages of the fusion process to highlight relevant features.

YDTR (Infrared and Visible Image Fusion via Y-Shape Dynamic Transformer,Tang,2022)design a network structure with two branches, two encoders and one decoder,which is called Y-shape dynamic transformer.The infrared image and visible image are fed into two Y branches,respectively. Each branch consists of an decoder and special transformer structure. Afterwards, these two branches are added and fed into the main path, which involves a module for feature integration and a decoder for dimensionality reduction.

DIDFuse(Deep Image Decomposition for Infrared and Visible Image Fusion,Zhao,2020)is an image fusion network based on auto-encoder (AE), and the network structure is based on UNet. The core idea is that the encoder decomposes an image into background and detail feature maps with low- and high-frequency information, respectively, and that the decoder recovers the original image. To this end, the loss function makes the background/detail feature maps of source images similar/dissimilar. In the test phase, background and detail feature maps are respectively merged via a fusion module, and the fused image is recovered by the decoder.

AUIF (Attention-based Unimodal and Intermodal Fusion,Zhao, 2020) is a fusion network that leverages attention mechanisms to fuse information from both unimodal and intermodal sources. It employs attention mechanisms to dynamically weight the contribution of each modality during fusion, focusing on relevant information. AUIF enhances the fusion process by selectively attending to informative regions, improving performance in tasks such as image classification and segmentation. The comparison data are shown in tab.3.

Table 3:Fusion quality comparison of different fusion methods

Methods	AG	SF	STD	CC
Our SFNet	**21.24**	**48.25**	**74.69**	0.81
CoConet	18.97	39.57	61.70	0.77
Reconet	4.54	10.3	32.34	0.76
SeAFusion	10.55	21.92	35.10	**0.83**
CBF	13.66	25.20	30.22	0.73
YDTR	7.78	17.48	28.94	0.76
DIDFuse	11.42	24.73	44.39	0.75
AUIF	13.20	28.04	46.56	0.77

For the highest data of each indicator, we made the font bold, and the second highest data was underlined. It is obvious that our SFNet has a splendid performance in AG, SF and STD, although the index of correlation coefficient is slightly lower than that of network SeAFusion, but it is still at a high level.

It is worth noting that although the performance of the first three indicators of these eight methods is quite different, the performance of the correlation coefficient is very good.This is because the image fusion method based on deep learning has stronger adaptability and nonlinear fitting ability, and can better capture the features of remote sensing images, making the fusion accuracy high. Moreover, the deep learning algorithm can automatically learn features without manual design, and can extract features and fuse them more accurately for various types of remote sensing images.

4.4 Model Complexity Analysis

The Shuffle Unit is the building block of ShuffleNetV2. It consists of a grouped convolution operation followed by channel shuffling and point-wise convolution. This architecture enables efficient information exchange across channels while maintaining computational efficiency.So we compare the fusion efficiency of different fusion methods.

Table 4:Comparison of fusion efficiency of different methods

Methods	Model size/MB	Mean running time/s
Our SFNet	**1.37**	**0.1033**
CoConet	42.81	0.2333
Reconet	20.79	0.5821
SeAFusion	51.92	0.7817
CBF	72.56	0.3461
YDTR	32.98	0.2332
DIDFuse	49.62	0.1206
AUIF	39.45	0.1379

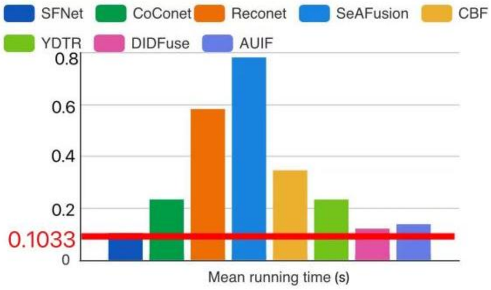
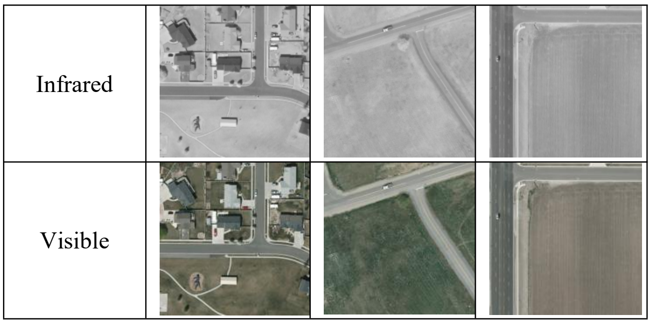


Figure 4: mean running time of different fusion methods

The mean running time of different image fusion methods can be observed visually from fig.4.And it is clear that our SFNet has the least amount of the time,while the slowest SeAFusion consumes more than six times of ours.While DIDFuse and AUIF perform well in running time, the AG and STD is much less than our SFNet. So the proposed SFNet achieves the balance of image fusion quality and time efficiency.

4.5 Subjective Evaluation



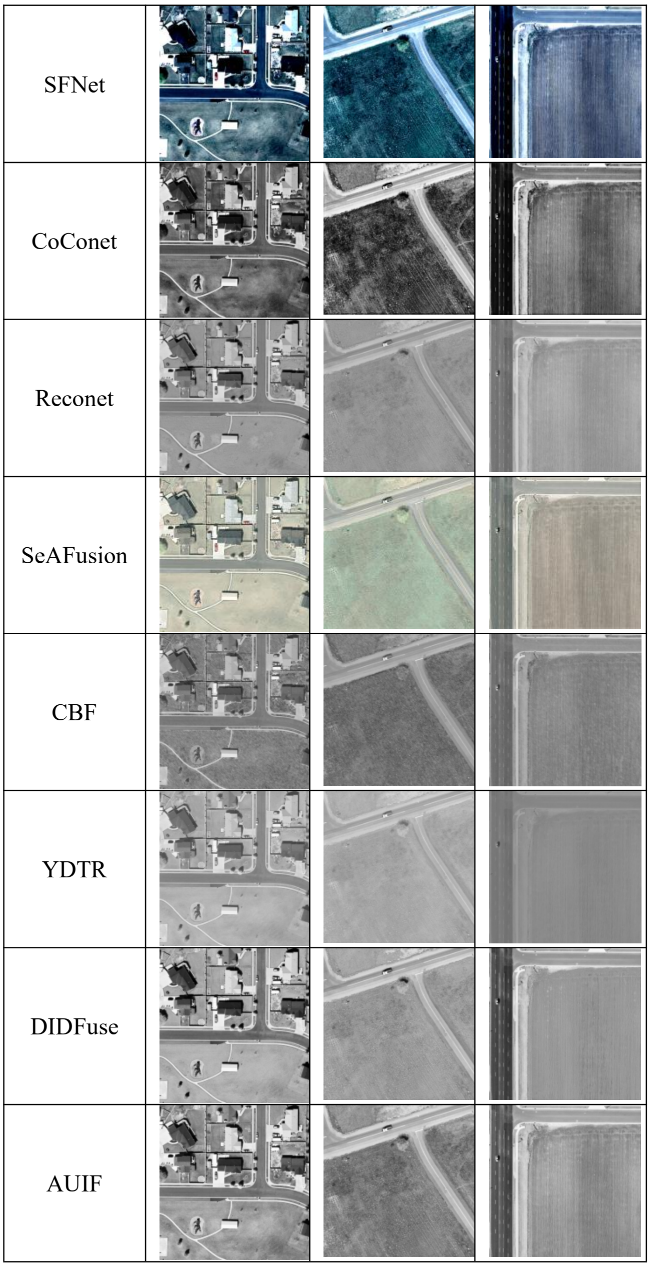


Figure 5:Fusion results of different image fusion methods

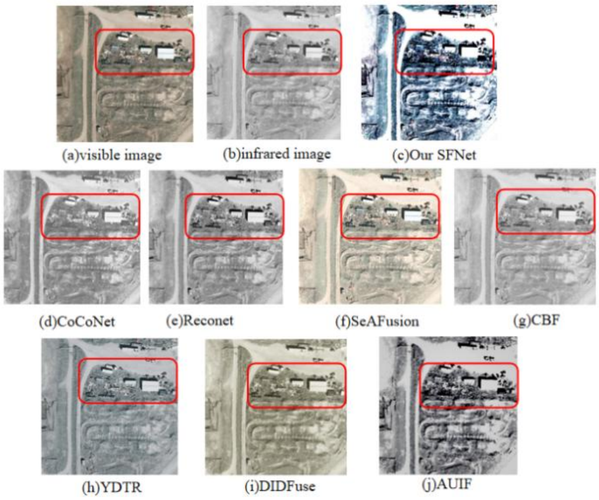


Figure 6:Fusion details of different image fusion methods

Fig.5 shows different methods’ performance to the same images. Our SFNet has the best visual effect.

And as you can see in fig.6, our network is clearest for the small, closely connected houses in the red box.This is due to the highest AG value,and it indicates the best image clarity. As for texture information,our SFNet performs just as well as the Coconet. Because Coconet has a lot of attention mechanisms, refining the fused image at multiple scales, while our SFNet obtain the fused images with ZAC and other light-weight algorithms.

Most image fusion methods based on neural networks have a good fusion effect, because deep learning takes into account both semantic information and detailed information. However, different neural networks have different performance in the clarity and rationality of fused images.Under such a premise, we hope to do a more lightweight network is meaningful, how to balance efficiency and quality, is we will continue to study.

4.6 SFNet Deployed On Brain-like Chips

The proposed multi-model data fusion network is deployed on the Lynxi KA200 brain computing chip, and a comprehensive inference test is carried out with UAV remote sensing data.

We implemented all the experiments using PyTorch and deployed SFNet on the Lynchip KA200 brain-inspired computing chip to conduct infrared and visible image fusion of UAV remote sensing cameras. That displays the high efficiency as well as high fusion quality of our network.

We almost achieved the same fine image fusion effect on the chip as on the end.

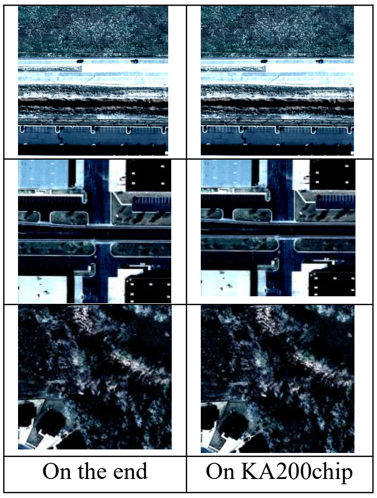


Figure 7:Fusion results of different devices

And more particular data are shown in tab.5. You can see the tiny difference between the computer and the KA200 chip.

Table 5: Fusion quality comparison of different devices

Devices	AG	SF	STD	CC
computer	**21.245**	48.254	**74.692**	**0.803**
KA200 chip	21.242	48.249	74.419	0.797

Our SFNet continues to perform well on brain-like chips, although all four indicators are slightly lower than the results on the computer, the overall integration is excellent. This indicates that our network has good performance and high scalability. The average time per image is 1.2 seconds, it is lower than our running time on computer. However, our network is also capable of performing real-time fusion work on the drone platform, we will also continue to improve the network to improve efficiency.

5. Conclusions

Image feature extraction and fusion strategy design are the key to infrared and visible image fusion. The existing deep convolutional feature extraction network has many parameters, deep structure and time-consuming calculation, and is not suitable for mobile and embedded devices.

In this paper, an image fusion method based on lightweight ShuffleNetv2 is proposed, and ShuffleNetv2 is used as an image feature extraction network to improve the shortcomings of the existing network. Experiments show that the proposed method can not only compress the network scale, but also greatly improve the speed and efficiency of fusion, and can adapt to mobile and embedded devices well.

We use ZCA and ℓ_1-norm to process the remodeled deep feature.By bicubic interpolation and soft-max operation,the weight maps w_k is ready.Finally by using weighted-average operation,the fused image is reconstructed.This fusion strategy works well on our selected remote sensing datasets. And we deployed our network on KA200 brain-like chip, which is based on a new integrated storage and computing, multi-core parallel, heterogeneous fusion architecture, and can efficiently support deep learning neural networks, biological neural networks and large-scale brain simulation. Our SFNet performed well on the brain-like chip. This gives us a lot more confidence in model porting and really working in real time on the drone platform. But it is worth noting that the design of fusion strategy is still a challenging task in the field of image fusion. It is of great significance to select appropriate network structure and feature extraction method according to specific needs, which will improve the quality of fusion image.

References

Bhatnagar, G., Wu, Q.M.J., Liu, Z., 2013. Directive contrast based multimodal medical image fu- sion in nsct domain. IEEE Transactions on Multimedia 15, 1014–1024. doi:10.1109/TMM.2013.2244870.

Faye, B., Azzag, H., Lebbah, M., Bouchaffra, D., 2024. Context-based multimodal fusion.URL:https://api.semanticscholar.org/CorpusID:268264421.

Gao, C., Gu, D., Zhang, F., Yu, Y., 2018. Reconet: Real-time coherent video style trans- fer network, in: Asian Conference on Computer Vision.

Kim, M., Han, D.K., Ko, H., 2016. Joint patch clustering-based dictionary learning for mul- timodal image fusion. Information Fusion 27, 198–214. doi:https://doi.org/10.1016/j.inffus.2015.03.003.

Li, H., Wu, X., 2018. Infrared and visible image fusion with resnet and zero-phase component analysis. ArXivabx/1806.07119.

Li, H., Wu, X.J., 2019. Densefuse: A fusion approach to infrared and visible images. IEEE Transactions on Image Processing 28, 2614–2623. doi:10.1109/TIP.2018.2887342.

Li, J., Chen, J., Liu, J., Ma, H., 2023. Learning a graph neural network with cross modality interaction for image fusion. Proceedings of the 31st ACM International Conference on Multi- media.

Li, Q., Han, G., Liu, P., Yang, H., Wu, J., Liu, D., 2021. An infrared and visible image fusion method guided by saliency and gradient information. IEEE Access 9, 108942– 108958. doi:10.1109/ACCESS.2021.3101639.

Liu, J., Lin, R., Wu, G., Liu, R., Luo, Z., Fan, X., 2022. Coconet: Coupled con- trastive learning network with multi-level feature ensemble for multi-modality image fusion. ArXiv abs/2211.10960.

Liu, R., Liu, J., Jiang, Z., Fan, X., Luo, Z., 2021. A bilevel integrated model with data-driven layer en- semble for multi-modality image fusion. IEEE Transactions on Image Processing 30, 1261–1274. doi:10.1109/TIP.2020.3043125.

Liu, Z., Liu, J., Wu, G., Ma, L., Fan, X.Y., Liu, R., 2023. Bi-level dynamic learn- ing for jointly multi-modality image fusion and beyond, in: In- ternational Joint Conference on Artificial Intelligence. URL: https://api.semanticscholar.org/CorpusID:258615243.

Ma, J., Chen, C., Li, C., Huang, J., 2016a. Infrared and visible image fusion via gradient transfer and total variation minimization. Information Fusion 31, 100–109. doi:https://doi.org/10.1016/j.inffus.2016.02.001.

Ma, J., Ma, Y., Li, C., 2019a. Infrared and visible image fusion methods and applica- tions: A survey. Information Fusion 45, 153–178.

Ma, J., Yu, W., Liang, P., Li, C., Jiang, J., 2019b. Fusiongan: A generative adversarial network for infrared and visible image fusion. Inf. Fusion 48, 11–26. URL: https://api.semanticscholar.org/CorpusID:71142966.

Ma, N., Zhang, X., Zheng, H., Sun, J., 2018. Shufflenet v2: Practical guidelines for ef- ficient cnn architecture design. ArXiv abs/1807.11164. URL: https://api.semanticscholar.org/CorpusID:51880435.

Ma, Y., Chen, J., Chen, C., Fan, F., Ma, J., 2016b. Infrared and visible image fusion using total variation model. Neurocomputing 202, 12–19.

Meng, F., Song, M., long Guo, B., Shi, R., Shan, D., 2017. Image fu- sion based on object region detection and non-subsampled con- tourlet transform. Comput. Electr. Eng. 62, 375–383.

Perona, P., Malik, J., 1990. Scale-space and edge de- tection using anisotropic diffusion. IEEE Transactions on Pattern Analysis and Machine Intelligence 12, 629–639. doi:10.1109/34.56205.

Petrovic, V., Xydeas, C., 2004. Gradient-based multiresolution image fusion. IEEE Transactions on Image Processing 13, 228–237. doi:10.1109/TIP.2004.823821.

Petrovic, V., Xydeas, C., 2005. Objective image fusion performance characterisation, in: Tenth IEEE International Conference on Computer Vision (ICCV’05) Volume 1, pp. 1866–1871 Vol. 2. doi:10.1109/ICCV.2005.175.

Prabhakar, K., Srikar, V.S., Babu, R.V., 2017. Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs. 2017 IEEE In- ternational Conference on Computer Vision (ICCV) , 4724–4732

Roberts, J., van Aardt, J.A.N., Ahmed, F.B., 2008. Assessment of image fu- sion procedures using entropy, image quality, and multispec- tral classification. Journal of Applied Remote Sensing 2.

Sandler, M., Howard, A.G., Zhu, M., Zhmoginov, A., Chen, L.C., 2018.Mo- bilenetv2: Inverted residuals and linear bottlenecks.2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition , 4510–4520

Singh, R., Vatsa, M., Noore, A., 2008. Integrated multilevel image fusion and match score fusion of visible and infrared face images for ro- bust face recognition. Pattern Recognit. 41, 880–893.

Tang, L., Zhang, H., Xu, H., Ma, J., 2023a. Rethinking the necessity of image fusion in high-level vision tasks: A practical infrared and visible image fusion network based on progressive semantic injection and scene fidelity. Information Fusion 99, 101870.

Tang, W., He, F., Liu, Y., 2023b. Ydtr: Infrared and visible image fusion via y-shape dynamic transformer. IEEE Transactions on Multimedia 25, 5413–5428. doi:10.1109/TMM.2022.3192661.

Vs, V., Jose Valanarasu, J.M., Oza, P., Patel, V.M., 2022. Image fusion transformer, in: 2022 IEEE International Conference on Image Processing (ICIP), pp. 3566–3570. doi:10.1109/ICIP46576.2022.9897280.

Zhang, K., Huang, Y., Yuan, X., Ma, H., Zhao, C., 2020. Infrared and visible image fusion based on intuitionistic fuzzy sets. Infrared Physics & Technology

Zhang, Q., Liu, Y., Blum, R.S., Han, J., Tao, D., 2018. Sparse representa- tion based multi-sensor image fusion for multi-focus and multi- modality images: A review. Inf. Fusion 40, 57–75.

Zhang, X., Ye, P., Qiao, D., Zhao, J., Peng, S., Xiao, G., 2019. Object fusion tracking based on visible and infrared images using fully convolutional siamese networks, in: 2019 22th International Conference on Information Fusion (FUSION), pp. 1–8. doi:10.23919/FUSION43075.2019.9011253.

Zhang, X., Zhou, X., Lin, M., Sun, J., 2017. Shufflenet: An extremely efficient convolu- tional neural network for mobile devices. 2018 IEEE/CVF Confer- ence on Computer Vision and Pattern Recognition , 6848–6856URL: https://api.semanticscholar.org/CorpusID:24982157.

Zhao, Z., Bai, H., Zhu, Y., Zhang, J., Xu, S., Zhang, Y., Zhang, K.,Meng, D., Timofte, R., Gool, L.V., 2023. Ddfm: Denoising diffusion model for multi-modality image fusion. 2023 IEEE/CVF Interna- tional Conference on Computer Vision (ICCV) , 8048–8059

Zhao, Z., Xu, S., Zhang, C., Liu, J., Li, P., Zhang, J., 2020. Didfuse: Deep image decomposition for infrared and visible image fusion, in: International Joint Conference on Artificial Intelligence. URL: https://api.semanticscholar.org/CorpusID:214605606.

Zhao, Z., Xu, S., Zhang, J., Liang, C., Zhang, C., Liu, J., 2022. Efficient and model- based infrared and visible image fusion via algorithm unrolling. IEEE Transactions on Circuits and Systems for Video Technology 32, 1186–1196. doi:10.1109/TCSVT.2021.3075745.