# DEVELOPING A TRANSFORMER-BASED APPROACH FOR FUSING INFRARED AND VISIBLE IMAGES FOR IMPROVED OBJECT DETECTION

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS OF
THE MIDDLE EAST TECHNICAL UNIVERSITY
BY

AYTEKIN ERDOGAN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
THE DEPARTMENT OF INFORMATION SYSTEMS

AUGUST 2023

**Developing A Transformer-Based Approach for Fusing Infrared and Visible Images for Improved Object Detection**

submitted by **AYTEKIN ERDOGAN** in partial fulfillment of the requirements for the degree of **Master of Science in Information Systems Department, Middle East Technical University** by,

Prof. Dr. Director of Institute
Dean, **Graduate School of Informatics** ____________________

Prof. Dr. Head of Department
Head of Department, **Information Systems** ____________________

Assoc. Prof. Dr. Supervisor
Supervisor, **Department, School** ____________________

Assoc. Prof. Dr. Co-supervisor if Exists
Co-supervisor, **Department, School** ____________________

**Examining Committee Members:**

Prof. Dr. Committee Member 1
Department, School ____________________

Assoc. Prof. Dr. Committee Member 2
Department, School ____________________

Assoc. Prof. Dr. Committee Member 3
Department, School ____________________

Assist. Prof. Dr. Committee Member 4
Department, School ____________________

Assist. Prof. Dr. Committee Member 5
Department, School ____________________

**Date: 28.08.2019**

**I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.**

**Name, Surname: Aytekin Erdogan**

**Signature :**

# ABSTRACT

## DEVELOPING A TRANSFORMER-BASED APPROACH FOR FUSING INFRARED AND VISIBLE IMAGES FOR IMPROVED OBJECT DETECTION

Erdogan, Aytekin
M.S., Department of Information Systems
Supervisor: Assoc. Prof. Dr. Supervisor
Co-Supervisor: Assoc. Prof. Dr. Co-supervisor if Exists

August 2023, **??** pages

English abstract here

Keywords: A keyword, another keyword, some other keywords

# ÖZ

## TÜRKÇE BAŞLIK

Erdogan, Aytekin
Yüksek Lisans, Bilişim Sistemleri Bölümü
Tez Yöneticisi: Doç. Dr. Supervisor
Ortak Tez Yöneticisi: Doç. Dr. Co-supervisor if Exists

Ağustos 2023, **??** sayfa

Türkçe öz buraya

Anahtar Kelimeler: Bir anahtar kelime, başka bir anahtar kelime, başka anahtar kelimeler

To the memories of my beloved friends Murat Tekin and Ragip Enes Katran

# ACKNOWLEDGMENTS

Acknowledgments here

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# LIST OF ABBREVIATIONS

| | |
|---|---|
| IF | Image Fusion |
| VIF | Visual and Infrared Image Fusion |
| AI | Artificial Intelligence |
| CNN | Convolutional Neural Networks |
| GAN | General Adveserial Networks |
| SR | Sparse Representation |
| MST | Multi-Scale Transformation |
| LRR | Low-Rank Representation |
| NAS | Neaural Architecture Search |
| SoTA | State-of-The-Art |
| PSNR | Peak Signal-to-Noise Ratio |
| SSIM | Sructural Similarity Index |

# CHAPTER 1

# INTRODUCTION

Image Fusion is a computer vision task that has been taken place for many years. Gathering all the complementary usefull informations into single image is called image fusion, *a.k.a* IF. Visual and Infrared Image Fusion, *henceforth will be referred to as VIF*, is a subfield of iamge fusion. Since the first study [**?**] in 1989, VIF is actively studied. In the era of AI, new methods such as CNN, GAN, auto-encoder, transformers are also applied to the VIF problem.

## 1.1 Research Questions

## 1.2 Contributions of the Study

## 1.3 Organization of the Thesis

# CHAPTER 2

# RELATED WORK

Image fusion algorithms can be categorized based on several factors, including whether they employ learning methods or hand-crafted steps, based on predefined loss functions, and whether labeled datasets are involved in the process. Learning-based methods involve the use of machine learning techniques, such as CNN, GAN, Transformers, Auto-encoders, to learn the features and relationships between input images. In contrast, hand-crafted approaches involve the manual selection and design of specific features and fusion rules.Whether a methods is end-to-end is defined as do they require one or more handcrafted steps. Loss functions are commonly used in learning-based methods to measure the quality of the fused image and to guide the training process. Based on the loss function, the process can be classified as self-supervised, supervised or unsupervised. Finally, labeled datasets can be used to train learning-based methods, where the some form of label is available and ground-truth labels are known, or for evaluating the performance of fusion algorithms. By considering these factors, researchers can select the most appropriate image fusion algorithm for their specific application requirements.

The available VIF methods can be categorized into two groups: traditional methods, which were widely used before the advent of AI, and learning-based methods. Regardless of the classification, all methods consist of three main components: image feature extraction, fusion of features from multiple images, and reconstruction of the image from the fused features. During the feature extraction stage, features are extracted from multiple images. In the fusion stage, the extracted features are compared, and complementary features are incorporated into a single feature map or set. Finally, in the reconstruction stage, the image is reconstructed from the fused set of features. All related studies aim to improve one or more of these stages in the process.

Learning-based fusion methods are often used to combine information from multiple sources to obtain more accurate and informative results. These methods can be categorized based on whether they require ground truth labels or not. The ground truth annotations don't have to be a form of fused image. If ground truth labels are needed, they are known as supervised methods. On the other hand, if no ground truth labels are required, they are referred to as unsupervised or self-supervised methods. These fusion techniques are commonly employed in convolutional-based methods, which are widely used in image processing and computer vision applications. Regardless of the specific method used, the goal of learning-based fusion methods is to improve the quality and accuracy of the resulting output by leveraging information from multiple sources.

Learning-based image fusion algorithms are commonly categorized based on the type of network used in the algorithm. There are several types of networks that are frequently used, including CNN-based, auto-encoder based, GAN-based, transformer-based networks and others which include hybrid or handcrafted steps. Convolutional neural network (CNN)-based methods are often used due to their ability to extract features from input images and produce high-quality results. Auto-encoder-based methods are also popular, as they can effectively compress and decompress image information to obtain a fused image. Generative adversarial network (GAN)-based methods use a combination of generator and discriminator networks to produce high-quality fused images. Transformer-based methods have recently gained popularity due to their ability to process long sequences of input data efficiently. Other methods, such as sparse representation-based methods and wavelet-based methods, are also used in learning-based image fusion. Ultimately, the choice of network depends on the specific requirements of the task at hand and the available resources.

In addition to categorizing fusion algorithms based on the type of network used, another way to differentiate them is whether they are end-to-end. An end-to-end algorithm is one that can take raw input data and produce the desired output directly, without any intermediate hand-crafted steps. In the context of image fusion, end-to-end algorithms are those that can take multiple input images and produce a fused image without the need for manual feature extraction or other preprocessing steps. These algorithms are often preferred, as they can be more efficient and less prone to errors compared to non-end-to-end methods that require manual intervention. In contrast, non-end-to-end algorithms may require additional processing steps, such as registration and feature extraction, to produce the final fused image. While these steps can provide additional control over the fusion process, they can also introduce additional complexity and reduce the overall efficiency of the algorithm.

## 2.1 Traditional Fusion Algorithms

The traditional image fusion algorithms have been extensively studied in the literature. However, they are not without shortcomings. One of the major issues with these methods is the presence of handcrafted steps, which can lead to suboptimal results. In addition, the time complexity of some methods can be quite high, making them impractical for real-world applications.

Sparse representation (SR)[**?**] based methods are a popular choice for image fusion. However, they suffer from several limitations. For example, methods such as [**?**] and [**?**] require dictionary learning, which can significantly increase the time complexity of the algorithm. Furthermore, these methods include handcrafted steps, which can limit their generalizability.

Another commonly used approach for image fusion is multi-scale transformation (MST) based methods. These methods, such as [**?**] and [**?**], can be quite effective at capturing various characteristics of images at different scales. However, they too suffer from limitations. One major issue is their lack of generalizability, which can make them less effective in certain scenarios.

Low-rank representation (LRR) based methods are another popular choice for image fusion. These methods, such as [**?**], are particularly effective at dealing with noise and other forms of image degradation. However, like the other methods, they too have limitations. For example, they may not be suitable for all types of images, particularly those with complex textures or patterns.

In summary, the success of traditional image fusion algorithms heavily relies on the quality of the feature extraction method used. While there are many different methods available, each with its own strengths and weaknesses, it is important to carefully consider the specific requirements of the problem at hand before selecting a particular method.

## 2.2 CNN Based Deep Learning Algorithms

First image fusion method that utilizes CNN is created by Liu et al [?]. The paper proposed a method to fuse infrared and visible images using Convolutional Neural Networks (CNNs). The approach consisted of four main steps: preprocessing, feature extraction, fusion strategy, and reconstruction. First, the input images were resized and normalized. Then, a CNN model was trained to extract features from both the visible and infrared images. The extracted features were then fused using a weighted sum fusion strategy, and the weights were learned during training. Finally, the fused feature maps were transformed back into the image domain using a deconvolutional network. The proposed method was evaluated on various datasets and outperformed other state-of-the-art fusion methods in terms of objective quality metrics and subjective visual quality. The proposed method in the paper also utilizes a multiscale approach to fuse the source images. The images are decomposed into different scales using Laplacian and Gaussian pyramids. This allows the model to better preserve details in both the visible and infrared images during the fusion process. To train the model, high-quality images and their blurred versions are used. These blurred versions are generated by applying multiscale Gaussian filtering and random sampling, which helps the model to learn features that are robust to image blurring and noise. It's also important to note that the proposed method is a pioneering deep learning-based approach to VIF, which is a metric used to evaluate the visual quality of the fused images. By introducing CNNs to VIF, the proposed method achieved better fusion results than other state-of-the-art methods available at that time.

CNN-based methods can be categorized into two groups: supervised and unsupervised methods. While both methods use CNNs to extract features from the input images, supervised methods require labeled data during the training process, whereas unsupervised methods do not. The majority of CNN-based image fusion methods are unsupervised, as labeled data is often difficult or expensive to obtain. However, to improve the performance of supervised methods, extra steps, such as data augmentation or transfer learning, are often employed. Despite the success of both supervised and unsupervised CNN-based methods, there is still room for improvement in image fusion techniques, particularly in scenarios where the input images have significant differences in terms of illumination, resolution, or noise.

### 2.2.1 Supervised CNNs

Image fusion is a challenging task, as obtaining ground truth data is often difficult or impossible. However, there are several workarounds that have been proposed to generate pseudo annotations and ground truth data. One approach is to use the results of another fusion method as ground truth data. This can be done by comparing the output of the method being evaluated to the output of a known or well-established fusion method. Another approach is to generate blurred versions of the high-quality input images and use these blurred images as labels. For example, in the study by Liu et al. [?], blurred versions of the high-quality input images were used to train a deep learning-based fusion model. Similarly, in the study by An et al. [?], the results of other fusion methods were used as annotations to train a convolutional neural network for image fusion. While these approaches are not a substitute for true ground truth data, they provide a means to evaluate and improve image fusion methods, especially in scenarios where obtaining actual ground truth is challenging or impossible.

It is also wise to state that many CNN-based image fusion methods rely on transfer learning, where pre-trained models, such as ResNet50 [?, ?], VGG19[?, ?], VGG16 [?], and DenseNet-201 [?], are used to extract features from source images. The extracted features are then processed before being fused using manually designed rules, such as weighted averaging or combining the fused base and detail parts. Some studies generate weight maps based on the extracted features [?, ?, ?], while others optimize a loss function computed based on the extracted features using the L-BFGS method [?]. These transfer learning-based methods provide a means to leverage pre-existing knowledge in large-scale datasets and improve the performance of image fusion methods.

### 2.2.2 Unsupervised CNNs

In the field of infrared visual image fusion, most related studies are unsupervised. Due to the lack of a ground truth for this task, loss functions are typically defined as a function of source images and related evaluation metrics. This means that the performance of the fusion method is evaluated based on how well the fused image aligns with the intended objective, rather than how closely it matches a predefined target. While this approach can be challenging due to the subjective nature of image quality evaluation, it remains a popular method for evaluating unsupervised image fusion techniques.

Several studies in the field of image fusion have explored the use of unsupervised convolutional neural networks (CNNs) to improve the process. These CNNs can be applied to either a single part of the process or the entire process. For instance, Liu et al. [?] decompose source images into a base and a detail parts using CNNs, while Hou et al. [?] uses CNNs in the feature extraction and reconstruction parts of the process. Some studies, such as Xu et al. [?] and Mustafa et al. [?], use unsupervised CNNs for the entire image fusion process. By utilizing unsupervised CNNs, these studies aim to improve the overall performance of the image fusion process, particularly for tasks such as infrared visual image fusion.

In addition to unsupervised CNNs, there are other methods that can be used to improve the performance of infrared and visual image fusion. Residual connections [?] and dense connections [?] are examples of techniques that have been used to enhance feature propagation within neural networks. Attention mechanisms [?] have also been applied to focus on important features and exclude irrelevant ones, while multi-scale and multilevel features have been used to capture details across different spatial and frequency ranges. Contrastive learning [?] and neural architecture search [?] are other approaches that have been used to improve the performance of image fusion methods. Image and feature decomposition techniques can also be used to decompose the source images into different components and extract features from them. By combining these various methods, researchers can develop more sophisticated image fusion systems that produce higher-quality results.

## 2.3 Autoencoder Based Deep Learning Algorithms

Autoencoder is a type of neural network that is widely utilized for unsupervised learning tasks, including but not limited to dimensionality reduction, data compression, and anomaly detection. The fundamental principle of autoencoder is to compress the input data into a lower-dimensional representation, which is commonly referred to as the latent space. The compressed representation is then utilized to reconstruct the original input data. The objective is to minimize the difference between the input data and the reconstructed output, which requires the autoencoder to learn a compressed representation that captures the most essential features of the input data. The concept of autoencoder was first introduced by Hinton et al.[?].

In the context of infrared visual image fusion, the autoencoder technique is leveraged to extract features from source images using the encoder stage, while the decoder stage reconstructs the fused image. The training process typically involves two stages: firstly, the autoencoder is trained using source images, either infrared, visual or both, without any fusion. Subsequently, the fusion step is integrated and the entire model is trained. It is also common practice to employ large datasets for the first training stage. It is worth noting that the autoencoder approach, which is used in this method, differs from transfer learning methods discussed in Section **??** in that an autoencoder is trained from scratch, while pre-trained models are utilized in transfer learning with minimal fine tuning. One of the pioneering works in infrared and visual image fusion using autoencoder is DenseFuse [?]. To pre-train the network without the fusion step, the well-known MS-COCO dataset [?] is utilized in the first stage training. Other noteworthy studies include Raza et al. [?], Fu et al. [?], Jian et al. [?], Wang et al. [?], and finally, Zhao et al. [?].

There are still open research questions for this part since rgb and infrared images are different in structure and only rgb image is used in autoencoder's training. There are studies that focuses on this difference. IVFENet [?], an encoder and two decoders are used in the pretraining stage to alleviate the vital information loss. Liu et al. [?] use two encoders and one unified decoder, which utilizes a NAS technique for the model. Measures for performance increase in Section **??** are also used with autoencoder based models such as dense connections [?, ?, ?], attention mechanism [?], multiscale features [?, ?], multi-level features [?].

### 2.4 GAN Based Deep Learning Algorithms

Since its introduction by Goodfellow et al. [?] in 2014, General Adversarial Networks, also known as GAN, have found a wide range of applications. In 2019, Ma et al. [?] introduced GAN to the image fusion task, after which many GAN-based image fusion methods have been proposed. Typically, these methods are unsupervised and utilize various combinations of one-generator-to-one-discriminator, one-generator-to-more-discriminators, and two-generators-to-two-discriminators. Additionally, some supervised versions of GAN-based image fusion methods have been developed.

#### 2.4.1 Unsupervised GANs

Most of the GAN-based VIF methods are unsupervised methods, which means that they do not require labeled data for training. Instead, the training process is driven by a loss function that measures the difference between the fused image and the source images. This loss function typically contains several terms that reflect the difference from different perspectives. For example, one term may measure the pixel-wise difference between the fused image and the source images, while another term may measure the structural similarity between the fused image and the source images.

In addition to the loss function, evaluation metrics are also used to assess the quality of the fused image. These metrics provide quantitative measures of various aspects of the fused image, such as its spatial frequency, contrast, and sharpness. Some commonly used evaluation metrics for VIF include mutual information, entropy, edge preservation, and spatial frequency. The choice of evaluation metrics can depend on the specific application of the VIF method.

Overall, the unsupervised nature of GAN-based VIF methods allows them to learn from data without the need for manual labeling, making them flexible and adaptable to a wide range of applications. The use of loss functions and evaluation metrics further enhances the performance and effectiveness of these methods, enabling them to produce high-quality fused images that preserve the most important features of the source images.

In Section **??**, researchers have explored various methods to enhance the performance of fusion techniques. Numerous approaches have been proposed in the literature to improve the fusion process and achieve better results. For instance, in the work of Xu et al. [**?**], they incorporated a local binary pattern loss during the training phase. This loss function allowed the model to capture local texture information, leading to improved fusion outcomes. Another study by Xu et al. [**?**] introduced the use of residual blocks and skip connections in the generator. By leveraging these architectural components, the model was able to learn residual features and establish direct connections between different layers, facilitating the flow of information. This architectural design helped to alleviate the vanishing gradient problem and enhanced the overall fusion performance.

In a different approach, Fu et al. [**?**] proposed the utilization of dense blocks to further augment the capabilities of the generator. Dense blocks enabled the model to learn richer feature representations by densely connecting layers within the generator. This architecture not only encouraged the fusion of features from different depths but also facilitated the learning of intricate patterns and relationships within the input data. To leverage the valuable information contained in the visible image, Fu et al. [**?**] took an innovative step by incorporating the visible image at each layer of the generator. This strategy enabled the network to effectively capture visible information and incorporate it into the fusion process. By fusing features from the visible image with those learned from the input images, the model was able to enhance the visibility of important details and improve the overall quality of the fused image. Furthermore, attention mechanisms [**?**] and residual connections [**?**] have also been employed as supplementary techniques in fusion methods. Attention mechanisms allow the model to focus on informative regions and weight their contributions accordingly, while residual connections facilitate the flow of gradients and aid in training deep networks. These techniques have proven to be effective in enhancing fusion performance and achieving state-of-the-art results in various fusion tasks. By integrating these innovative approaches and techniques, researchers aim to overcome the challenges of information fusion and improve the quality and effectiveness of fusion models. These advancements pave the way for more accurate and visually appealing fused images in a wide range of applications, such as remote sensing, medical imaging, and computer vision.

The previously mentioned methods utilize original visible and infrared images as inputs for the generator. However, these approaches only focus on the primary information found in each modality and neglect the supplementary information. Consequently, the resulting fused images tend to resemble sharpened infrared images. To tackle this challenge, Ma et al. introduced the GANMcC method [**?**]. GANMcC incorporates a two-branch architecture within the generator, where each branch (gradient and contrast) employs different combinations of source images as input. This design allows the generator to capture both the main and auxiliary information. Another notable method is MFEIF [**?**], which does not require well-aligned image pairs during training. MFEIF utilizes a coarse-to-fine deep architecture to leverage multiscale features and integrates a cross-domain edge-guided attention mechanism, which directs the model's focus towards common structures to preserve finer details. Moreover, Liao et al. proposed a technique [?] that employs VGG19 to extract features from both the visible and fused images generated by the generator. They then minimize the Wasserstein distance within the feature space. A major drawback of most GAN-based methods is their reliance on a single discriminator to enforce similarity either to the visible or infrared image. This approach often leads to the loss of details from the other modality during the adversarial training process. To overcome this limitation, Ma et al. [?] introduced a multiclassification-based discriminator, aiming to strike a balance between the visible and infrared distribution. Some researchers have also proposed the use of multiple discriminators as a potential solution to this issue.

To address the limitations of considering a single source image in the discriminator, researchers have extended GAN-based methods to incorporate two or more discriminators, aiming to preserve features from both source images. For instance, Xu et al. proposed DDcGAN, a VIF method with two discriminators that can fuse source images of different resolutions [?]. Ma et al. further extended this method by using a densely connected CNN in the generator, taking the image itself as input in the discriminator, and employing a deconvolution layer for upsampling the infrared image [?]. Other researchers also recognized the benefits of employing multiple discriminators. Li et al. developed unsupervised GAN-based VIF methods with one generator and two discriminators, such as MD-WGAN, D2WGAN, and MgAN-Fuse [?, ?, ?]. These methods introduced various improvements, including the use of texture loss, Wasserstein distance, multiscale attention, and separate encoders for visible and infrared images. Li et al. also proposed AttentionFGAN, incorporating multiscale attention mechanisms in the generator and discriminators [?]. Additionally, Zhang et al. utilized a full-scale skip connection-based generator and two Markovian discriminators to retain useful information from both visible and infrared source images [?]. Song et al. recently introduced a VIF method with one generator and three discriminators, including a difference image discriminator [?]. These advancements involving multiple discriminators aim to enhance the fusion process and consider various aspects of the input images.

### 2.4.2 Supervised GANs

In the context of supervised GAN-based methods, similar to the supervised CNN methods discussed in Section ??, there are several approaches that utilize different types of ground truth. The first type involves using fused images generated by other methods as the reference. Lebedev et al. [?] proposed a method with one generator and one discriminator, where fused images obtained through the Laplacian pyramid algorithm accompanied by MultiScale Retinex served as the ground truth. Another approach, RCGAN, introduced by Li et al.[?], is based on coupled GAN and features two generators and two discriminators. Notably, RCGAN incorporates pre-fused images generated by GFF [?] for optimization within the coupled generators. However, it is important to consider that the performance of RCGAN can be influenced by the specific method used to generate the pre-fused images.

## 2.5 Transformer Based Deep Learning Algorithms

Transformers have emerged as versatile tools capable of handling long-range dependencies in diverse domains, including natural language processing and computer vision tasks [?, ?, ?]. Their entry into the realm of image fusion in 2021 has opened up exciting prospects for advancing image fusion techniques such as [?, ?, ?, ?, ?, ?, ?]. By harnessing the capabilities of transformers, a range of transformer-based methods have been developed to enhance various aspects of image fusion, encompassing both general fusion scenarios [?], [?], [?], [?]. These methods capitalize on transformers' inherent ability to capture contextual relationships and dependencies across different regions within images, enabling a more comprehensive fusion of visual information. Leveraging the self-attention mechanism, these methods effectively model interactions between pixels or patches, facilitating the fusion of relevant features while preserving crucial details. In short, the integration of transformers into the field of image fusion has sparked notable advancements, empowering the creation of transformer-based methods for image fusion applications. Leveraging transformers' prowess in capturing long-range dependencies and contextual information, these methods pave the way for more sophisticated and accurate fusion of visual data, elevating the possibilities in the field of image fusion.

In the realm of feature fusion, certain methods leverage the transformer architecture for this purpose. One approach, presented by VS et al. [?], utilizes a transformer-based multiscale fusion strategy to merge local and global information effectively. This method acknowledges the importance of incorporating both local and global context in the fusion process. Similarly, Zhao et al. [?] propose DNDT, a method that employs a dual transformer framework for fusion. Their approach involves an encoder to extract features from source images and a decoder to construct the fused image, allowing for the integration of information from multiple sources.

Another line of research focuses on transformer-based fusion blocks. Liu et al. [?] introduce a transformer fusion block that utilizes focal self-attention to fuse multiscale features obtained from a multiscale encoder network. This fusion approach demonstrates the effectiveness of transformer-based techniques in incorporating multiscale information. Recent advancements have extended the application of transformers to image fusion tasks. Rao et al. [?] propose a VIF method that combines transformers and GAN. Their approach integrates a spatial transformer and a channel transformer within the generator, enabling the fusion of spatial and channel-wise information. Additionally, Ma et al. introduce SwinFusion [?], a general image fusion method based on the Swin Transformer. This approach highlights the significance of global information in image fusion and provides valuable visualizations to better understand its impact. Collectively, these transformer-based methods provide innovative solutions for feature fusion in image fusion tasks. They leverage the strengths of transformers to effectively incorporate information from multiple sources and emphasize the importance of global context in the fusion process. By harnessing the capabilities of transformers, these methods contribute to advancements in image fusion techniques and offer new insights into the significance of global information.

In alternative approaches, transformers are not only utilized for feature fusion but also play a role in other stages of VIF methods. For instance, Fu et al.[?] introduced a pyramid patch transformer method, which incorporates a transformer-based feature extraction module and an MLP-based decoder within an autoencoder framework. The fusion process in this method follows an average strategy. Similarly, Wang et al. [?] developed SwinFuse, an autoencoder-based framework where a transformer-based encoder is employed to extract global features. Tang et al. [?] proposed YDTR, which combines CNN and transformers in both the encoding and decoding branches. Visible and infrared image features are merged together. Another example is CGTF [117], which utilizes transformer and convolution feature extraction modules in both the encoding and decoding branches. Additionally, Tang et al. [151] designed a parallel transformer-based global feature extraction branch alongside their CNN-based local feature extraction branch. Yang et al. [?] employed a combination of transformer blocks and convolution blocks to generate fused images from source images.

The architectural diversity among transformer-based methods reflects the flexibility and adaptability of these approaches in addressing various image fusion challenges. Researchers have explored different ways to integrate transformers and CNNs within the fusion pipeline to capitalize on their respective strengths. By combining the local and global modeling capabilities of CNNs and transformers, these methods aim to achieve a more comprehensive and informative fusion of visual data.

It is noteworthy that existing transformer-based VIF methods operate in an unsupervised manner. This means that they do not rely on explicit ground truth labels during training. Instead, the loss function is derived from the comparison between the fused image and the source images themselves. This unsupervised nature presents both advantages and challenges. On one hand, it enables the methods to be applicable to a wide range of scenarios without the need for annotated training data. On the other hand, it poses difficulties in evaluating and objectively comparing different methods, as there is no direct reference for the quality of the fusion output. As researchers delve deeper into transformer-based image fusion, there is ongoing exploration of novel architectural designs and techniques. This includes investigating the integration of transformers at different stages of the fusion pipeline, exploring variations in the combination of transformers and CNNs, and exploring the potential of leveraging additional information or guidance to further enhance the fusion process. By pushing the boundaries of transformer-based methods and continually refining their designs, researchers strive to improve the quality and accuracy of fused images, advancing the field of image fusion across diverse domains and applications.

## 2.6 Datasets

In the image fusion task, the absence of ground truth presents a significant hurdle when it comes to developing supervised VIF methods as discussed in the Section **??** and Section **??**. The majority of deep learning-based VIF methods lean towards unsupervised approaches due to the unavailability of reliable ground truth data. However, researchers have not been deterred by this challenge and have made efforts to address it through various means. They have explored alternative methods to generate pseudo ground truth or devised innovative ways of utilizing different forms of labels to facilitate supervised training in the image fusion domain. To tackle the lack of ground truth, researchers have adopted creative strategies to generate pseudo ground truth that approximates the desired fusion result. These approaches often involve using existing fusion methods as a reference to create synthesized fused images, which can then serve as pseudo ground truth during the training process. By utilizing these pseudo ground truth images, supervised fusion methods can be trained in a manner that mimics the desired fusion outcomes, enabling the models to learn and optimize fusion performance based on the generated reference images. In addition to pseudo ground truth generation, researchers have explored alternative forms of labels that can be utilized in supervised training for image fusion. These labels may not directly correspond to ground truth fusion images but can provide valuable information and guidance during the training process. For instance, researchers have employed object masks or segmentation maps as labels, allowing the models to focus on specific regions or objects of interest during the fusion process. By leveraging such labels, supervised VIF methods can be trained to prioritize the preservation of salient features or specific visual elements in the fused images, leading to improved fusion quality and performance. In short, despite the absence of ground truth in the image fusion task, researchers have demonstrated their ingenuity by developing methods that effectively handle the challenge through the use of pseudo ground truth and alternative forms of labels. These approaches enable supervised VIF methods to leverage available information and guidance during the training process, enhancing their ability to learn and optimize fusion outcomes.

Unsupervised VIF methods employ a range of training data strategies to tackle the challenge of lacking supervision. These strategies offer diverse approaches to learn from available data and enhance the performance of image fusion techniques. One approach involves utilizing pairs of visible-infrared images as training data [**?**] [75], [143]. By leveraging the inherent correlation between these image modalities, the models can learn to extract and fuse relevant features effectively.

Another approach focuses on using independent visible and infrared images, which may not be directly paired, as the training data. This strategy is commonly observed in Autoencoder (AE)-based methods [28], [53], [72], [76], [78]. These methods leverage the inherent structures within individual modalities to learn meaningful representations and subsequently generate fused images. Although these training data do not possess explicit correspondences, the models aim to capture the underlying patterns and correlations between the modalities.

Additionally, some approaches make use of all-clear visible images, primarily employed in AE-based methods. For example, Li et al. [42] utilize the MSCOCO dataset to train the encoder and decoder components. By leveraging the rich visual content present in these images, the models can learn to encode and decode the relevant information necessary for fusion. This data-driven approach benefits from the availability of large-scale visible datasets, enabling the models to capture diverse visual characteristics.

Furthermore, a hybrid strategy combines visible images with visible-infrared image pairs to train different modules of the model. For instance, Jian et al. [90] employ visible images to train an image decomposition module and visible-infrared image pairs to train a stacked sparse autoencoder for local saliency map extraction. This approach leverages the complementary nature of the training data, allowing the model to learn both local saliency and decomposition features for effective fusion.

Lastly, transfer learning emerges as a viable solution to address the scarcity of training data. It involves utilizing pre-trained models that have been trained on large-scale RGB datasets, as mentioned in Section 2.4.3. By leveraging the knowledge learned from these pre-trained models, the models can acquire generic visual representations that can be adapted and fine-tuned for the task of VIF. This approach offers a practical and effective way to overcome the limitations of limited available training data.

Overall, these diverse training data strategies enable unsupervised VIF methods to learn from different sources of information and exploit inherent correlations, leading to enhanced fusion performance.

### 2.6.1 Methods to Create Labeled Dataset

Addressing the lack of ground truth in image fusion task is approached through various methods. One method involves using fused images generated by other techniques as a substitute for ground truth. For instance, Li et al. [24] utilize the Generative Fusion Framework (GFF) [146] to generate labels, while Lebedev et al. [22] employ the Laplacian pyramid algorithm in combination with MultiScale Retinex

[144] to produce ground truth images. However, this approach may have limitations on the learning process [101].

Another approach involves utilizing clear images and their corresponding blurred versions. This technique has been applied in several studies [23], [29], [47], [56], [82], where RGB images and their blurred counterparts are utilized. Recently, Zhu et al. [112] generated blurred versions for both RGB and infrared images. Nevertheless, the training data produced using this approach may lack realism and differ from actual visible-infrared image pairs.

A different strategy involves utilizing manually-labeled object masks available in existing image fusion datasets. Some researchers [31], [94]–[96] use these masks to retain semantic information in the fused images. However, this method is labor-intensive and not always convenient for obtaining the required masks.

Another technique involves transforming the image fusion task, which lacks ground truth, into a task that has ground truth for a specific part of the loss function. This is achieved by incorporating labels from downstream applications. For example, Shopovska et al. [44] use a pre-trained pedestrian detector to generate pedestrian labels and include these labels as an auxiliary detection loss during training. Tang et al. [108] employ scene segmentation as a downstream task and integrate a segmentation loss term into the overall loss function to guide training. The scene segmentation labels are manually labeled by the authors of the segmentation dataset. Similarly, Liu et al. [115] utilize general object detection as a downstream task and introduce an object detection loss term to the loss function, leveraging object detection labels provided by the creator of the detection dataset.

Lastly, the Y channel of RGB images in the YCbCr color space can be used as a form of ground truth. Synthetic infrared and visible images are generated to facilitate training [102].

In summary, researchers have adopted diverse methods to overcome the absence of ground truth in image fusion tasks. These methods include using fused images from other techniques, clear images and their blurred versions, manually-labeled object masks, labels from downstream applications, or synthetic images. Each approach has its own advantages and considerations, providing alternative means to enable supervised training and enhance the performance of image fusion algorithms.

# CHAPTER 3

# USER EXPERIMENT

In this chapter, the details of the user experiment are presented.

## 3.1 Research Method and Experiment Design

# CHAPTER 4

# USER EXPERIMENT

In this chapter, the details of the user experiment are presented.

## 4.1 Research Method and Experiment Design

# CHAPTER 5

# CONCLUSION AND FUTURE WORK

# REFERENCES

[1] A. Toet, L. J. Van Ruyven, and J. M. Valeton, "Merging thermal and visual images by a contrast pyramid," *Optical engineering*, vol. 28, no. 7, pp. 789–792, 1989.

[2] C. Liu, Y. Qi, and W. Ding, "Infrared and visible image fusion method based on saliency detection in sparse domain," *Infrared Physics & Technology*, vol. 83, pp. 94–102, 2017.

[3] Y. Bin, Y. Chao, and H. Guoyu, "Efficient image fusion with approximate sparse representation," *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 14, no. 04, p. 1650024, 2016.

[4] Q. Zhang, Y. Fu, H. Li, and J. Zou, "Dictionary learning method for joint sparse representation-based image fusion," *Optical Engineering*, vol. 52, no. 5, pp. 057006–057006, 2013.

[5] H.-M. Hu, J. Wu, B. Li, Q. Guo, and J. Zheng, "An adaptive fusion algorithm for visible and infrared videos based on entropy and the cumulative distribution of gray levels," *IEEE Transactions on Multimedia*, vol. 19, no. 12, pp. 2706–2719, 2017.

[6] K. He, D. Zhou, X. Zhang, R. Nie, Q. Wang, and X. Jin, "Infrared and visible image fusion based on target extraction in the nonsubsampled contourlet transform domain," *Journal of Applied Remote Sensing*, vol. 11, no. 1, pp. 015011–015011, 2017.

[7] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 171–184, 2012.

[8] Y. Liu, X. Chen, J. Cheng, H. Peng, and Z. Wang, "Infrared and visible image fusion with convolutional neural networks," *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 16, no. 03, p. 1850018, 2018.

[9] W.-B. An and H.-M. Wang, "Infrared and visible image fusion with supervised convolutional neural network," *Optik*, vol. 219, p. 165120, 2020.

[10] H. Li, X.-j. Wu, and T. S. Durrani, "Infrared and visible image fusion with resnet and zero-phase component analysis," *Infrared Physics & Technology*, vol. 102, p. 103039, 2019.

[11] G. Li, Y. Lin, and X. Qu, "An infrared and visible image fusion method based on multi-scale transformation and norm optimization," *Information Fusion*, vol. 71, pp. 109–129, 2021.

[12] H. Li, X.-J. Wu, and J. Kittler, "Infrared and visible image fusion using a deep learning framework," in *2018 24th international conference on pattern recognition (ICPR)*, pp. 2705–2710, IEEE, 2018.

[13] X. Ren, F. Meng, T. Hu, Z. Liu, and C. Wang, "Infrared-visible image fusion based on convolutional neural networks (cnn)," in *Intelligence Science and Big Data Engineering: 8th International Conference, IScIDE 2018, Lanzhou, China, August 18–19, 2018, Revised Selected Papers 8*, pp. 301–307, Springer, 2018.

[14] Y. Yang, J.-X. Liu, S.-Y. Huang, H.-Y. Lu, and W.-Y. Wen, "Vmdm-fusion: a saliency feature representation method for infrared and visible image fusion," *Signal, Image and Video Processing*, pp. 1–9, 2021.

[15] Y. Li, J. Wang, Z. Miao, and J. Wang, "Unsupervised densely attention network for infrared and visible image fusion," *Multimedia Tools and Applications*, vol. 79, no. 45-46, pp. 34685–34696, 2020.

[16] Y. Liu, L. Dong, Y. Ji, and W. Xu, "Infrared and visible image fusion through details preservation," *Sensors*, vol. 19, no. 20, p. 4556, 2019.

[17] R. Hou, D. Zhou, R. Nie, D. Liu, L. Xiong, Y. Guo, and C. Yu, "Vif-net: an unsupervised framework for infrared and visible image fusion," *IEEE Transactions on Computational Imaging*, vol. 6, pp. 640–651, 2020.

[18] H. Xu, M. Gong, X. Tian, J. Huang, and J. Ma, "Cufd: An encoder–decoder network for visible and infrared image fusion based on common and unique feature decomposition," *Computer Vision and Image Understanding*, vol. 218, p. 103407, 2022.

[19] H. T. Mustafa, J. Yang, H. Mustafa, and M. Zareapoor, "Infrared and visible image fusion based on dilated residual attention network," *Optik*, vol. 224, p. 165409, 2020.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[21] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

[22] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," in *Advances in neural information processing systems*, pp. 2204–2212, 2014.

[23] G. E. Hinton and R. R. Salakhutdinov, "Dimensionality reduction by learning an invariant mapping," *arXiv preprint arXiv:0704.2550*, 2006.

[24] B. Zoph and Q. V. Le, "Neural architecture search with reinforcement learning," in *International Conference on Learning Representations*, 2017.

[25] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[26] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755, Springer, 2014.

[27] A. Raza, H. Huo, and T. Fang, "Pfaf-net: Pyramid feature network for multimodal fusion," *IEEE Sensors Letters*, vol. 4, no. 12, pp. 1–4, 2020.

[28] Y. Fu and X.-J. Wu, "A dual-branch network for infrared and visible image fusion," in *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 10675–10680, IEEE, 2021.

[29] L. Jian, X. Yang, Z. Liu, G. Jeon, M. Gao, and D. Chisholm, “Sedrfuse: A symmetric encoder–decoder with residual block network for infrared and visible image fusion,” *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–15, 2020.

[30] Z. Wang, Y. Wu, J. Wang, J. Xu, and W. Shao, “Res2fusion: Infrared and visible image fusion based on dense res2net and double nonlocal attention models,” *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–12, 2022.

[31] Z. Zhao, S. Xu, J. Zhang, C. Liang, C. Zhang, and J. Liu, “Efficient and model-based infrared and visible image fusion via algorithm unrolling,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1186–1196, 2021.

[32] F. Zhao, W. Zhao, L. Yao, and Y. Liu, “Self-supervised feature adaption for infrared and visible image fusion,” *Information Fusion*, vol. 76, pp. 189–203, 2021.

[33] J. Liu, Y. Wu, Z. Huang, R. Liu, and X. Fan, “Smoa: Searching a modality-oriented architecture for infrared and visible image fusion,” *IEEE Signal Processing Letters*, vol. 28, pp. 1818–1822, 2021.

[34] Y. Pan, D. Pi, I. A. Khan, Z. U. Khan, J. Chen, and H. Meng, “Densenetfuse: A study of deep unsupervised densenet to infrared and visual image fusion,” *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–13, 2021.

[35] Z. Wang, J. Wang, Y. Wu, J. Xu, and X. Zhang, “Unfusion: A unified multi-scale densely connected network for infrared and visible image fusion,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 6, pp. 3360–3374, 2021.

[36] Z. Li, H. Wu, L. Cheng, S. Luo, and M. Chen, “Infrared and visible fusion imaging via double-layer fusion denoising neural network,” *Digital Signal Processing*, vol. 123, p. 103433, 2022.

[37] Y. Peng, G. Liu, X. Xu, D. P. Bavirisetti, X. Gu, and X. Zhang, “Mfdetection: A highly generalized object detection network unified with multilevel heterogeneous image fusion,” *Optik*, vol. 266, p. 169599, 2022.

[38] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial networks,” in *Advances in neural information processing systems*, pp. 2672–2680, 2014.

[39] J. Ma, W. Yu, P. Liang, C. Li, and J. Jiang, “Fusiongan: A generative adversarial network for infrared and visible image fusion,” *Information fusion*, vol. 48, pp. 11–26, 2019.

[40] J. Xu, X. Shi, S. Qin, K. Lu, H. Wang, and J. Ma, “Lbp-began: A generative adversarial network architecture for infrared and visible image fusion,” *Infrared Physics & Technology*, vol. 104, p. 103144, 2020.

[41] D. Xu, Y. Wang, S. Xu, K. Zhu, N. Zhang, and X. Zhang, “Infrared and visible image fusion with a generative adversarial network and a residual network,” *Applied Sciences*, vol. 10, no. 2, p. 554, 2020.

[42] Y. Fu, X.-J. Wu, and T. Durrani, “Image fusion based on generative adversarial network consistent with perception,” *Information Fusion*, vol. 72, pp. 110–125, 2021.

[43] J. Wang, Y. Li, and Z. Miao, "A new infrared and visible image fusion method based on generative adversarial networks and attention mechanism," in *2021 The 4th International Conference on Image and Graphics Processing*, pp. 109–119, 2021.

[44] J. Ma, H. Zhang, Z. Shao, P. Liang, and H. Xu, "Ganmcc: A generative adversarial network with multiclassification constraints for infrared and visible image fusion," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–14, 2020.

[45] J. Liu, X. Fan, J. Jiang, R. Liu, and Z. Luo, "Learning a deep multi-scale feature ensemble and an edge-attention guidance for image fusion," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 1, pp. 105–119, 2021.

[46] B. Liao, Y. Du, and X. Yin, "Fusion of infrared-visible images in ue-iot for fault point detection based on gan," *IEEE Access*, vol. 8, pp. 79754–79763, 2020.

[47] H. Xu, P. Liang, W. Yu, J. Jiang, and J. Ma, "Learning a generative model for fusing infrared and visible images via conditional generative adversarial network with dual discriminators.," in *IJCAI*, pp. 3954–3960, 2019.

[48] J. Ma, H. Xu, J. Jiang, X. Mei, and X.-P. Zhang, "Ddcgan: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion," *IEEE Transactions on Image Processing*, vol. 29, pp. 4980–4995, 2020.

[49] J. Li, H. Huo, K. Liu, and C. Li, "Infrared and visible image fusion using dual discriminators generative adversarial networks with wasserstein distance," *Information Sciences*, vol. 529, pp. 28–41, 2020.

[50] J. Li, H. Huo, C. Li, R. Wang, C. Sui, and Z. Liu, "Multigrained attention network for infrared and visible image fusion," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–12, 2020.

[51] J. Li, H. Huo, C. Li, R. Wang, and Q. Feng, "Attentionfgan: Infrared and visible image fusion using attention-based generative adversarial networks," *IEEE Transactions on Multimedia*, vol. 23, pp. 1383–1396, 2020.

[52] H. Zhang, J. Yuan, X. Tian, and J. Ma, "Gan-fm: Infrared and visible image fusion using gan with full-scale skip connection and dual markovian discriminators," *IEEE Transactions on Computational Imaging*, vol. 7, pp. 1134–1147, 2021.

[53] A. Song, H. Duan, H. Pei, and L. Ding, "Triple-discriminator generative adversarial network for infrared and visible image fusion," *Neurocomputing*, vol. 483, pp. 183–194, 2022.

[54] M. Lebedev, D. Komarov, O. Vygolov, and Y. V. Vizilter, "Multisensor image fusion based on generative adversarial networks," in *Image and Signal Processing for Remote Sensing XXV*, vol. 11155, pp. 565–574, SPIE, 2019.

[55] Q. Li, L. Lu, Z. Li, W. Wu, Z. Liu, G. Jeon, and X. Yang, "Coupled gan with relativistic discriminators for infrared and visible images fusion," *IEEE Sensors Journal*, vol. 21, no. 6, pp. 7458–7467, 2019.

[56] S. Li, X. Kang, and J. Hu, "Image fusion with guided filtering," *IEEE Transactions on Image processing*, vol. 22, no. 7, pp. 2864–2875, 2013.

[57] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[58] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

[59] X. Liu, H. Gao, Q. Miao, Y. Xi, Y. Ai, and D. Gao, "Mfst: Multi-modal feature self-adaptive transformer for infrared and visible image fusion," *Remote Sensing*, vol. 14, no. 13, p. 3233, 2022.

[60] H. Zhao and R. Nie, "Dndt: Infrared and visible image fusion via densenet and dual-transformer," in *2021 International Conference on Information Technology and Biomedical Engineering (ICITBE)*, pp. 71–75, IEEE, 2021.

[61] D. Rao, T. Xu, and X.-J. Wu, "Tgfuse: An infrared and visible image fusion approach based on transformer and generative adversarial network," *IEEE Transactions on Image Processing*, 2023.

[62] J. Li, J. Zhu, C. Li, X. Chen, and B. Yang, "Cgtf: Convolution-guided transformer for infrared and visible image fusion," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–14, 2022.

[63] W. Tang, F. He, and Y. Liu, "Ydtr: infrared and visible image fusion via y-shape dynamic transformer," *IEEE Transactions on Multimedia*, 2022.

[64] Z. Wang, Y. Chen, W. Shao, H. Li, and L. Zhang, "Swinfuse: A residual swin transformer fusion network for infrared and visible images," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–12, 2022.

[65] X. Yang, H. Huo, R. Wang, C. Li, X. Liu, and J. Li, "Dglt-fusion: A decoupled global–local infrared and visible image fusion transformer," *Infrared Physics & Technology*, vol. 128, p. 104522, 2023.

[66] W. Tang, F. He, and Y. Liu, "Tccfusion: An infrared and visible image fusion method based on transformer and cross correlation," *Pattern Recognition*, p. 109295, 2023.

[67] V. Vs, J. M. J. Valanarasu, P. Oza, and V. M. Patel, "Image fusion transformer," in *2022 IEEE International Conference on Image Processing (ICIP)*, pp. 3566–3570, IEEE, 2022.

[68] Y. Fu, T. Xu, X. Wu, and J. Kittler, "Ppt fusion: Pyramid patch transformerfor a case study in image fusion," *arXiv preprint arXiv:2107.13967*, 2021.

[69] J. Ma, L. Tang, F. Fan, J. Huang, X. Mei, and Y. Ma, "Swinfusion: Cross-domain long-range learning for general image fusion via swin transformer," *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 7, pp. 1200–1217, 2022.

[70] L. Qu, S. Liu, M. Wang, S. Li, S. Yin, Q. Qiao, and Z. Song, "Transfuse: A unified transformer-based image fusion framework using self-supervised learning," *arXiv preprint arXiv:2201.07451*, 2022.