FuseFormer: A Transformer For Visual and Infrared Image Fusion

1st Aytekin Erdogan
Graduate School of Informatics
Middle East Technical University
Ankara, Turkey
aytekin.erdogan@metu.edu.tr

2nd Erdem Akagunduz
Graduate School of Informatics
Middle East Technical University
Ankara, Turkey
akaerdem@metu.edu.tr

***Abstract*—Image fusion is a process where images obtained from different sensors are combined to generate a single image that benefits from complementary information. Recently, there has been a growing interest in image fusion, which involves fusing images from diverse sensors to produce an enhanced image. Although deep learning methods have been widely employed in state-of-the-art techniques to extract meaningful features for image fusion, these methods primarily focus on integrating local features while disregarding the broader context within the image. To overcome this limitation, Transformer-based models have emerged as a promising solution, aiming to capture general context dependencies through attention mechanisms. Inspired by this, we propose a novel image fusion approach that incorporates a transformer-based multi-scale fusion strategy, effectively considering both local and general context information, thus enhancing the overall fusion process. Our proposed method follows a two-stage training approach, where an auto-encoder is initially trained to extract deep features at multiple scales. Subsequently, the multi-scale features are fused using a combination of Convolutional Neural Networks (CNNs) and Transformers. The CNNs are utilized to capture local features, while the Transformer handles the integration of general context features. Notably, in contrast to similar methods, we propose novel loss functions to address the challenges associated with defining a loss function when ground truth for fusion is absent. Through extensive experiments on various benchmark datasets, our proposed method, along with the novel loss function definition, demonstrates superior performance compared to other competitive fusion algorithms. Overall, this thesis presents significant advancements in image fusion techniques, offering innovative approaches and contributing to the state-of-the-art in this field.**

***Index Terms*—Image Fusion, Visual Infrared Image Fusion, Transformer Based Image Fusion, Structural Similarity Metric**

I. Introduction

The intricate evolution of perception systems across species, from minuscule insects to towering apex predators, showcases nature's mastery. As humans delve into technological innovation, we've engineered mechanical eyes that not only emulate but at times exceed their biological counterparts. This intertwining of natural instincts and modern camera technology manifests a limitless realm of perception. Central to computer vision is the spectrum of light, which extends from the visible colors our eyes discern, around 400 to 700 nanometers, to realms beyond human sight, including the warmth of infrared and the energy of ultraviolet waves. While humans are limited to a segment of this spectrum, technology unlocks a wider expanse, particularly through both visible and infrared imagery, heralding a revolutionary phase in computer vision. Image fusion leverages multiple image sources, offering enhanced clarity and understanding, and is pivotal in contemporary computer vision. Specifically, Visual-Infrared Image Fusion (VIF) melds visible and infrared spectrums, proving transformative for applications like night vision and thermal imaging. This fusion augments perception in low-visibility conditions, benefiting sectors like the military and security. Various techniques, from traditional pixel and feature-level fusion to advanced learning methods such as CNN-based and transform-based fusion, are employed to effectively merge these images, amplifying situational comprehension across applications.

A. Problem Definition

While deep learning methods excel at extracting pertinent features for image fusion, their primary focus on local features often neglects the overarching context within the image, a concern raised by Li et al. (2021) [1]. Transformer-based models, employing attention mechanisms, offer a solution, capturing wider context dependencies to refine fusion outcomes.

A significant impediment in fusion techniques is crafting the appropriate loss function. Conventional methods rely heavily on the Structural Similarity Metric (SSIM) to ascertain similarity between the input visible band image and the fused output. This SSIM-centric approach has its limitations, given that its maximum value is only achieved when both images are identical. To address this, we suggest a novel loss function that places equal emphasis on congruence with both the visible band and infrared inputs. This dual-input consideration ensures optimal retention of essential information, bolstering the quality of the fusion.

In response to the pressing need for enhanced image fusion techniques, our research bifurcates into two pivotal studies. The primary study delves deep into innovating the loss function. Central to this exploration is the creation of a loss function that harmonizes the similarity between both visible and infrared inputs. Traditional methods often marginalize this aspect, leading to subpar fusion results. Our proposed function seeks to remedy this, fostering a more efficient and meaningful fusion process.

Through this structure, we aim to fully harness the potential of image fusion techniques, heralding not just advancements in current applications but also paving the way for future innovations.

B. Research Questions and Hypotheses

To gain a comprehensive understanding of the complexities surrounding image fusion, our research seeks to answer the following hypotheses, systematically divided into two studies based on their thematic focus:

1) New Loss Function Proposal:

Hypothesis 1: The new loss function, which emphasizes the similarity between both the visible band input and the infrared image input, will result in more informative and meaningful fused images compared to using Structural Similarity Metric (SSIM) as the guiding criterion.

Hypothesis 2: The proposed approach will surpass existing state-of-the-art image fusion methods in terms of visual quality, providing more detailed, sharper, and visually appealing fused images.

This structured presentation of research questions and hypotheses aims to provide a clear roadmap for our investigation. Through rigorous testing of these hypotheses, our ultimate goal is to validate or challenge them, leading to significant contributions to the field of image fusion. By systematically evaluating each hypothesis within the respective studies, we aim to draw conclusions that are both robust and comprehensive, thereby advancing the current state of knowledge within this domain.

2) A Unique Transformer Based Fusion Strategy:

Hypothesis 3: The combination of Transformer-based models and the new loss function will significantly improve night vision enhancement, medical imaging, and surveillance tasks, allowing for better object detection, classification, and tracking in challenging lighting conditions.

Hypothesis 4: The proposed transformer based approach will achieve a better balance between quantitative and qualitative performance in image fusion, overcoming the compromise between the two that is often observed in traditional deep learning methods.

Hypothesis 5: The proposed approach will demonstrate computational efficiency, making it suitable for real-time applications, such as video surveillance and live medical imaging, without sacrificing the quality of the fused images.

Hypothesis 6: The limitations and challenges associated with implementing Transformer-based image fusion techniques can be mitigated through proper model tuning, regularization, and architecture adjustments, leading to improved overall performance.

II. Related Work

The domain of RGB and infrared image fusion has been a focal point of substantial research, witnessing a plethora of innovative methods from the early 1990s to the cutting-edge Transformer-based models. These techniques span categories like Traditional Fusion Algorithms, CNN, Autoencoder, GAN, and notably, Transformer-based approaches where attention mechanisms optimize the fusion. Within this vast spectrum, a key challenge is the limited availability of labeled datasets due to the inherent complexity of fusing RGB and IR images. As a result, our research navigates the unsupervised scenario, leveraging various Performance Evaluation and Benchmarking metrics to quantify the fusion results, offering insights into the strengths, limitations, and practical applications of each method. This chapter culminates in our exploration of Transformer-based algorithms, representing a comprehensive journey through RGB and IR image fusion techniques and challenges.

Image fusion algorithms can be classified based on factors like the use of learning methods versus hand-crafted steps, predefined loss functions, and the involvement of labeled datasets. While learning-based methods like CNN, GAN, Transformers, and Auto-encoders harness machine learning to understand input images, hand-crafted approaches involve manual feature design and fusion rules. The categorization extends to whether methods are end-to-end or require intermediary handcrafted steps. Furthermore, loss functions guide the training process, leading to classifications like self-supervised, supervised, or unsupervised based on the loss function. Labeled datasets, whether for training learning-based methods where ground-truth is known or for evaluating fusion algorithms, play a pivotal role. These considerations help researchers select the most apt image fusion algorithm tailored to their application requirements.

1) Traditional Fusion Algorithms: Traditional image fusion algorithms, while extensively studied, come with inherent shortcomings, notably the presence of handcrafted steps leading to potentially suboptimal results and high time complexity in some cases. Sparse representation (SR) based methods, such as those by Bin et al [2] and Zhang et al [3], are popular but are limited by requirements like dictionary learning, which increases their time complexity, and handcrafted steps affecting their generalizability. Multi-scale transformation (MST) based methods, exemplified by Hu et al [4] and others [5], are effective at capturing image characteristics at varying scales but can lack generalizability in specific contexts. Low-rank representation (LRR) methods, like those from Liu et al [6], excel in handling noise and image degradation, but may falter with complex textures or patterns. Ultimately, the efficacy of traditional fusion algorithms largely hinges on the chosen feature extraction method, necessitating careful consideration of the problem's specific demands before settling on a technique.

A. Learning Based Algorithms for Image Fusion

1) CNN Based Algorithms: Liu et al. were among the first to introduce a method for image fusion that utilizes CNN, focusing on the fusion of infrared and visible images [7]. Their approach consisted of several steps, from preprocessing to feature extraction, fusion strategy, and reconstruction, achieving better fusion results than other state-of-the-art methods of the

time. CNN-based methods are generally divided into supervised and unsupervised categories. While both utilize CNNs for feature extraction, supervised methods require labeled data, often leveraging techniques like data augmentation or transfer learning. Unsupervised methods, on the other hand, don't require labeled data, making them more flexible for real-world applications. Despite their success, there's room for improvement, especially in situations where input images have significant differences in factors like illumination or resolution.

2) Autoencoder Based Algorithms: Autoencoders, a type of neural network, have been widely used for tasks like dimensionality reduction and data compression. Within the domain of infrared visual image fusion, autoencoders are leveraged to extract features from source images in the encoder stage, while the decoder stage focuses on reconstructing the fused image. The training process involves two stages: firstly, training the autoencoder using source images without any fusion, followed by integrating the fusion step. Notable works in this area include DenseFuse [8] and studies by Raza et al. [9] and Fu et al. [10], among others.

3) GAN Based Algorithms: Since their introduction by Goodfellow et al. [11], GANs have found diverse applications, including image fusion. The majority of GAN-based image fusion methods are unsupervised, focusing on the difference between the fused and source images. Innovations in this space include the use of multiple discriminators, as seen in methods introduced by Ma et al. [12] and Xu et al. [13], and the integration of attention mechanisms and residual connections for improved performance.

4) Transformer Based Algorithms: Transformers have risen as paramount tools in managing long-range dependencies in fields like natural language processing and computer vision [14]–[16]. Their introduction into image fusion in 2021 revolutionized the domain, with several innovative methodologies emerging [17]–[23].

Central to these methodologies is the self-attention mechanism, enhancing fusion outcomes by preserving essential details and efficiently blending features. For instance, VS et al. [24] introduced a multiscale fusion strategy harnessing this architecture, while Zhao et al. [17] advanced a dual transformer approach. Noteworthy strategies also include the integration of transformers with traditional methods, as evidenced by Fu et al. [25], Wang et al. [21], and others.

Transformer-based VIF techniques predominantly operate unsupervised, leveraging loss functions from the comparison of fused and source images. While this eliminates the need for annotated data, it also complicates methodological evaluations due to the absence of a direct quality reference.

With increasing research into transformer-based image fusion, the focus remains on novel designs and methodologies, such as diverse transformer integrations in the fusion pipeline, transformer-CNN combinations, and the utilization of auxiliary information to enrich fusion processes.

III. Methodology

In this chapter, we delve into the realm of infrared and visual band image fusion using transformers, offering a detailed account of our experimental setup and the rationale for our model selection that marries autoencoders with fusion transformers. The robustness of our research is grounded in the carefully curated datasets described in Section **??**, which encompass diverse real-world scenarios and are tailored for compatibility with our fusion model. The infrastructure supporting our research, detailed in Section **??**, boasts state-of-the-art GPUs, enhancing our model's training and inference speeds. Our image fusion model's foundation, the autoencoder, is explored in Section III-A1, highlighting its capability to distill crucial features from both image bands, subsequently processed by our fusion transformer—a novel yet unclaimed pioneering approach in the field—described in Section III-A3. Our commitment to refining image fusion is further emphasized by our tailored loss functions, delineated in Sections III-A2 and III-A4, which play pivotal roles in our model's training to achieve superior fusion results. Through this exploration, we aspire to advance the horizons of multi-modal image processing, heralding advancements in areas like remote sensing and surveillance.

A. Model Selection

Image fusion can be segmented into three components: the feature extractor, feature fuser, and image reconstructor. The feature extractor derives multilevel features from input images, which the feature fuser then amalgamates into unified feature maps for each level. These consolidated features facilitate the final image reconstruction by the image reconstructor. Drawing parallels, the feature extractor aligns with the encoder of autoencoders (Section **??**), and the image reconstructor mirrors autoencoders' decoder. The feature fuser harnesses both Convolutional Neural Networks (CNNs) from Section **??** and transformers from Section II-A4. CNNs adeptly handle local features, while transformers manage global contexts, thus marrying them enriches the fusion process, aiming for enhanced accuracy.

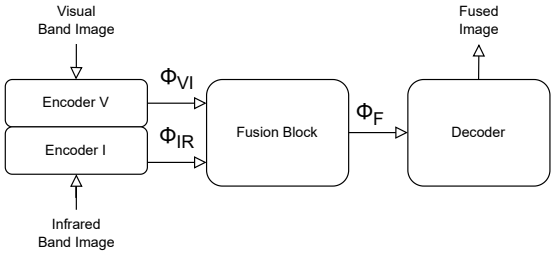


Fig. 1: High-Level Design of Model

Given the problem's complexity and the proposed model configuration, it is essential to conduct in-depth analyses of each aspect. Therefore, the subsequent sections, from III-A1 to III-A4, delve into a comprehensive exploration of the model's inner workings, the datasets employed during experimentation,

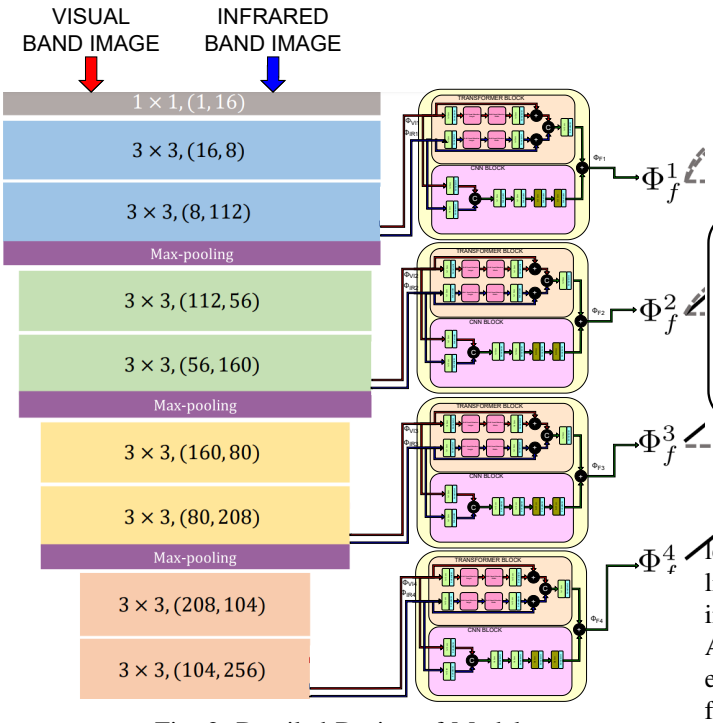


Fig. 2: Detailed Design of Model

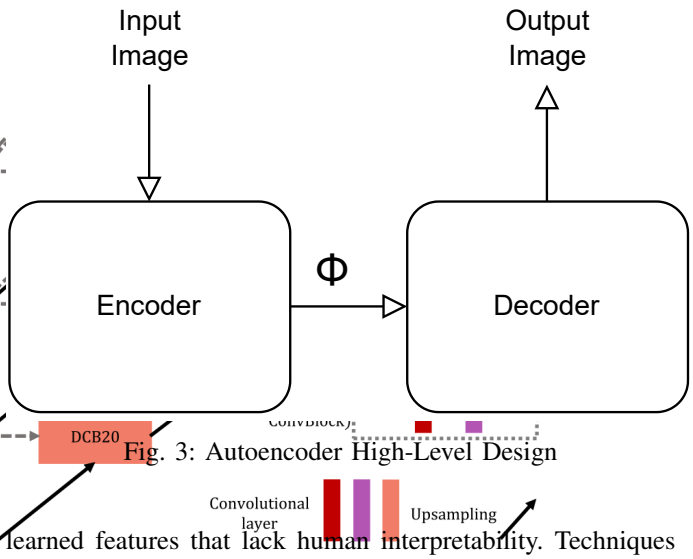


Fig. 3: Autoencoder High-Level Design

and the design of the loss function used to train the model. This deeper understanding will shed light on the model's capabilities and limitations, providing a solid foundation for testing and evaluating the proposed hypothesis.

1) Stage 1: Autoencoder Selection: Training autoencoders presents a set of challenges that require careful consideration and effective solutions. One of the primary obstacles encountered is the vanishing or exploding gradients during backpropagation, which can hinder the convergence of the model. To overcome this, using activation functions like $ReLU$ and employing gradient clipping techniques can stabilize the training process. Another critical issue is overfitting, where the model becomes too specialized to the training data. Regularization methods such as $L1$ or $L2$ regularization and dropout can help prevent overfitting and improve generalization.

Additionally, selecting the right dimension for the latent space is vital in autoencoders. A comprehensive hyperparameter search and evaluation using techniques like cross-validation can guide the choice of an appropriate latent space dimension. Moreover, traditional autoencoders might struggle with multimodal data, producing unimodal outputs.Data preprocessing also plays a crucial role in training autoencoders. Careful normalization, scaling, and augmentation techniques can improve the model's robustness and performance. However, dealing with large and deep autoencoder architectures can be computationally expensive. To address this issue, optimizing the model architecture, reducing the number of parameters, and utilizing hardware accelerators like GPUs or TPUs can significantly speed up training.

Despite their capabilities, autoencoders may produce learned features that lack human interpretability. Techniques like activation maximization and feature visualization can aid in understanding the representations learned by the model. Additionally, enforcing sparsity constraints in the encoder can encourage the extraction of more interpretable and compact features.

By addressing these challenges and applying appropriate solutions, autoencoders can become powerful tools for learning meaningful representations, enabling a wide range of applications in various domains. The combination of theoretical insights and practical strategies is essential for designing robust autoencoder architectures that can tackle complex real-world problems effectively.

Vanishing or Exploding Gradients: Autoencoders, especially deep autoencoders with multiple layers, can suffer from the vanishing or exploding gradient problem during backpropagation. This occurs when the gradients become too small (vanishing) or too large (exploding), leading to slow convergence or instability during training. **Solution**: To mitigate the vanishing gradient problem, use activation functions that are less prone to saturation, such as ReLU (Rectified Linear Unit). Additionally, consider using gradient clipping techniques to prevent gradients from exploding during backpropagation.

Overfitting: Autoencoders are prone to overfitting, especially when the network architecture is too complex or when the training dataset is limited. Overfitting occurs when the model becomes too specific to the training data and fails to generalize well to unseen data. **Solution**: Employ regularization techniques like L1 or L2 regularization to penalize large weights and prevent overfitting. You can also use dropout, where neurons are randomly dropped during training, to reduce interdependence among neurons and improve generalization.

Choice of Latent Space Dimension: Selecting the appropriate dimensionality for the latent space is crucial in autoencoders. If the latent space is too small, it may not capture all the essential features, leading to information loss. Conversely, a large latent space may lead

to the model memorizing the training data, resulting in poor generalization. **Solution**: Perform a hyperparameter search to determine the optimal latent space dimension using techniques like grid search or random search. Use techniques such as cross-validation to evaluate the model's performance with different latent space dimensions and choose the one that strikes a balance between capturing essential features and avoiding overfitting.

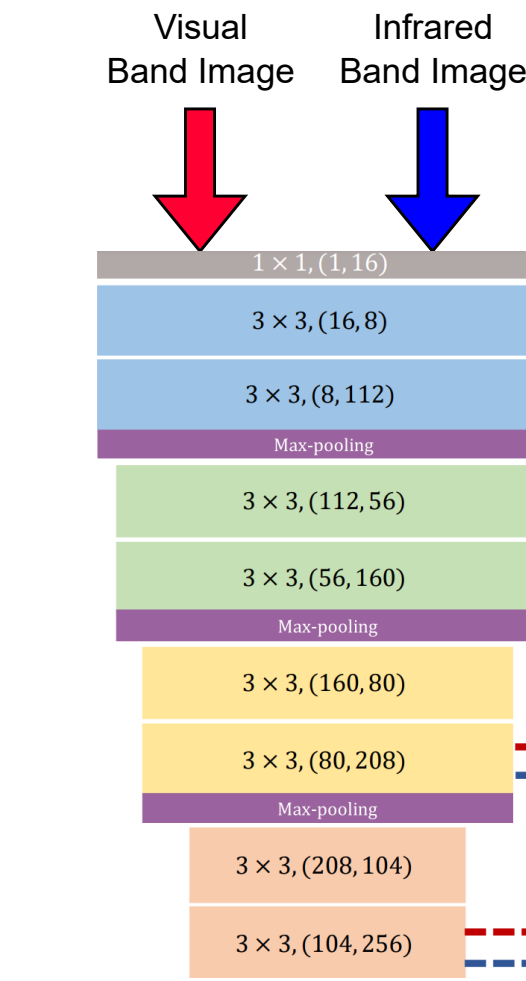
Unimodal Outputs: Standard autoencoders often produce unimodal outputs, which can limit their ability to handle multimodal data effectively. For tasks where multiple plausible outputs exist for a given input, autoencoders may struggle to capture this variability. **Solution**: For tasks involving multimodal data, consider using variational autoencoders (VAEs) or generative adversarial networks (GANs) that can produce diverse and realistic outputs. These models can learn a richer representation of the data, allowing for the generation of multiple plausible outputs for a given input.

High Computational Cost: Deep autoencoders with many layers and parameters can be computationally expensive to train, requiring substantial computational resources and time. **Solution**: Optimize the model architecture and reduce the number of parameters to lower the computational burden. Consider using transfer learning or pre-trained encoders to speed up training. Utilize hardware accelerators, such as GPUs or TPUs, to expedite the training process.

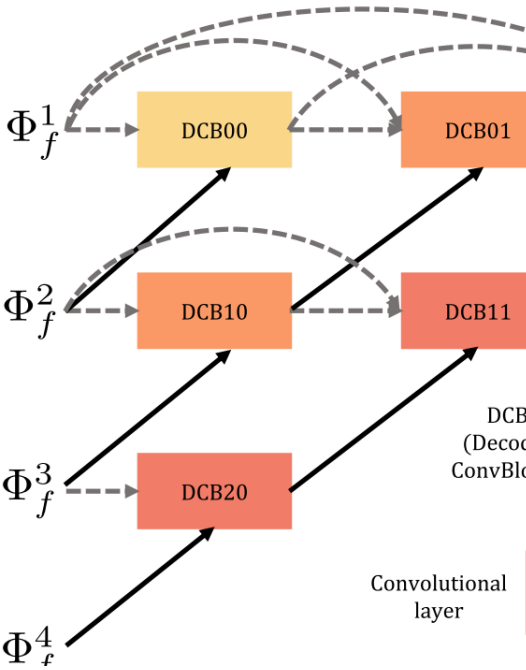
Choosing the Right Loss Function: The choice of the loss function can significantly impact the autoencoder's performance. Selecting an appropriate loss function for a specific task is crucial, and using an unsuitable one may lead to suboptimal results. **Solution**: Select a loss function that aligns with the specific task and data characteristics. For example, mean squared error (MSE) is suitable for continuous data, while binary cross-entropy is appropriate for binary data. Explore custom loss functions tailored to the unique requirements of the problem. In our experiment we have used the Eq 1 to direct the training.

Lack of Interpretable Features: While autoencoders can learn useful representations, the learned features may not always be human-interpretable, making it challenging to understand what specific features the model has captured. **Solution**: Investigate techniques for interpretability, such as activation maximization or feature visualization, to gain insights into the learned representations. Additionally, consider using sparsity constraints in the autoencoder's encoder to encourage the extraction of more interpretable and compact features.

To address the array of challenges outlined earlier, our approach commenced by adopting the autoencoder employed in RFN-Nest [1] as the base model which is also adopted from DenseFuse [8] paper autoencoder. This autoencoder possesses the ability to extract multilevel features from input images and successfully reconstructs the original image by decoding



(a) Encoder



(b) Decoder

Fig. 4: RFN-Nest Model Architecture [1]

these feature maps into a single coherent representation. In order to tailor the model to our specific problem requirements, we re-trained this base autoencoder using the MS-COCO [26] dataset, initializing the network with random weights. The process of training the autoencoder with the MS-COCO dataset resulted in the creation of a pretrained autoencoder, which effectively captures essential features from the input data.

Subsequently, we pursued fine-tuning the pretrained autoencoder with a lower learning rate using the RoadScene [27] dataset. This dataset contains both visual band and infrared images, allowing us to incorporate valuable information from both domains during the fine-tuning process. As a result, the autoencoder underwent a refined training stage, where it fine-tuned its feature extraction and reconstruction capabilities with the RoadScene dataset. The conclusion of this stage yielded a fully capable autoencoder that excels at encoding and decoding input images, achieving minimal loss in the process.

The comprehensive details of the stage 1 training and the subsequent fine-tuning process can be found in Section **??**, providing an in-depth analysis of the autoencoder's performance and the effect of leveraging different datasets during training. Our staged training methodology allowed us to progressively enhance the model's proficiency in feature extraction and image reconstruction, leading to a robust autoencoder model primed for further investigation and integration within our proposed fusion framework. In the subsequent sections, we present the extensive results obtained from this model, demonstrating its efficacy in image fusion and comparing it with existing state-of-the-art fusion methods on the TNO dataset [28]. Moreover, we delve into the implications of these findings, shedding light on the potential applications and future directions for advancing image fusion techniques.

2) Stage 1: Training of The Autoencoder: As mentioned in Section III-A1, the initial phase of the training process involves instructing the encoder network to capture multi-scale deep features. Concurrently, the decoder network is also trained to reconstruct the input image, utilizing the aforementioned multi-scale deep features. The training framework of the autoencoder network is depicted in Figure 4a. Distinguished from previous research, our feature extraction component integrates a down-sampling operation via max pooling, facilitating the extraction of deep features at various scales. These extracted multi-scale deep features are then fed into the decoder network for the purpose of reconstructing the input image. Leveraging short cross-layer connections ensures the comprehensive utilization of the multi-scale deep features in the image reconstruction process.

The loss function, denoted as L_{ae}, serves as the training criterion for the autoencoder network and is defined in the subsequent manner:

$$L_{ae} = L_{pixel} + \alpha L_{SSIM} \tag{1}$$

The terms L_{pixel} and L_{SSIM} refer to the pixel loss and the structural similarity (SSIM) loss, respectively, computed between the input and output images. The parameter α represents the trade-off parameter governing the balance between the contributions of L_{pixel} and L_{SSIM} meanwhile also it handles the order of magnitude difference in the overall loss function in Eq 1.

$$L_{\text{pixel}} = \left|\left|\text{image}_{\text{output}} - \text{image}_{\text{input}}\right|\right|_F^2 \tag{2}$$

L_{pixel} is defined in Eq 2. where $||.||_F$ denotes Frobenius norm. The Frobenius norm, denoted as $\|A\|_F$, is a matrix norm that measures the size or magnitude of a matrix A. For an $m \times n$ matrix A, the Frobenius norm is defined as the square root of the sum of the squares of all the elements of the matrix as in Eq 3:

$$\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2} \tag{3}$$

where a_{ij} represents the element in the ith row and jth column of matrix A.

L_{pixel} ensures that the reconstructed image closely resembles the original input image at the individual pixel level, imposing a constraint on the fidelity of pixel-wise information in the reconstruction process. This constraint helps to maintain fine-grained details and accuracy in the reconstructed image, ensuring that it retains the essential characteristics of the input image at a granular level.

The second term in Eq 1 is the SSIM loss L_{SSIM} is defined as in Eq 4:

$$L_{SSIM} = 1 - SSIM(image_{output}, image_{input}) \tag{4}$$

where $SSIM(.)$ is the structural similarity measure [29] which quantifies the structural similarity of the two images. The structural similarity between Input and Output is constrained by L_{SSIM}. The Structural Similarity Index (SSIM) is a widely used metric for evaluating the similarity between two images. It aims to capture not only the pixel-wise differences but also the structural information and perceptual quality of the images. The $SSIM(.)$ is formulated as in Eq 5 and Figure 5.

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1) \cdot (2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1) \cdot (\sigma_x^2 + \sigma_y^2 + C_2)} \tag{5}$$

where:

x and y represent the two images being compared.
μ_x and μ_y are the local means of x and y respectively.
σ_x^2 and σ_y^2 are the local variances of x and y respectively.
σ_{xy} is the local covariance between x and y.
C_1 and C_2 are constants to stabilize the division with weak denominators. They are often set to small values, such as $C_1 = (k_1 \cdot L)^2$ and $C_2 = (k_2 \cdot L)^2$, where L is the dynamic range of pixel values, and k_1 and k_2 are constants typically set to small positive values.

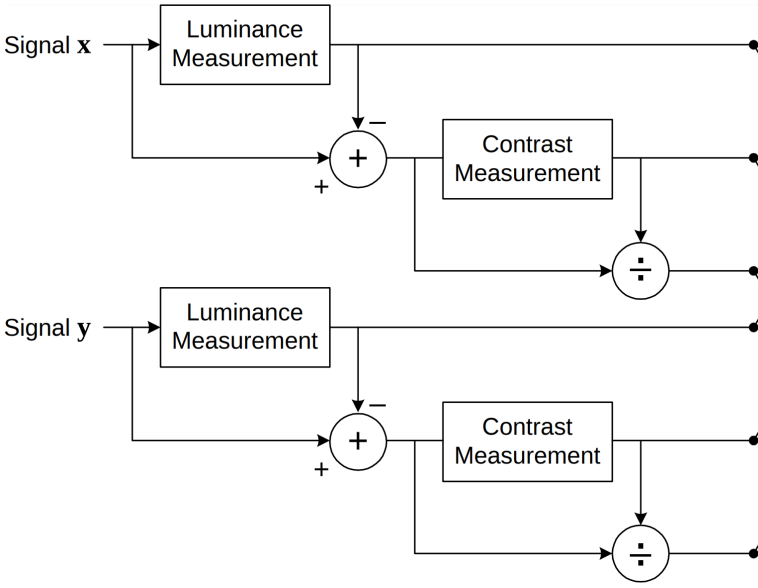
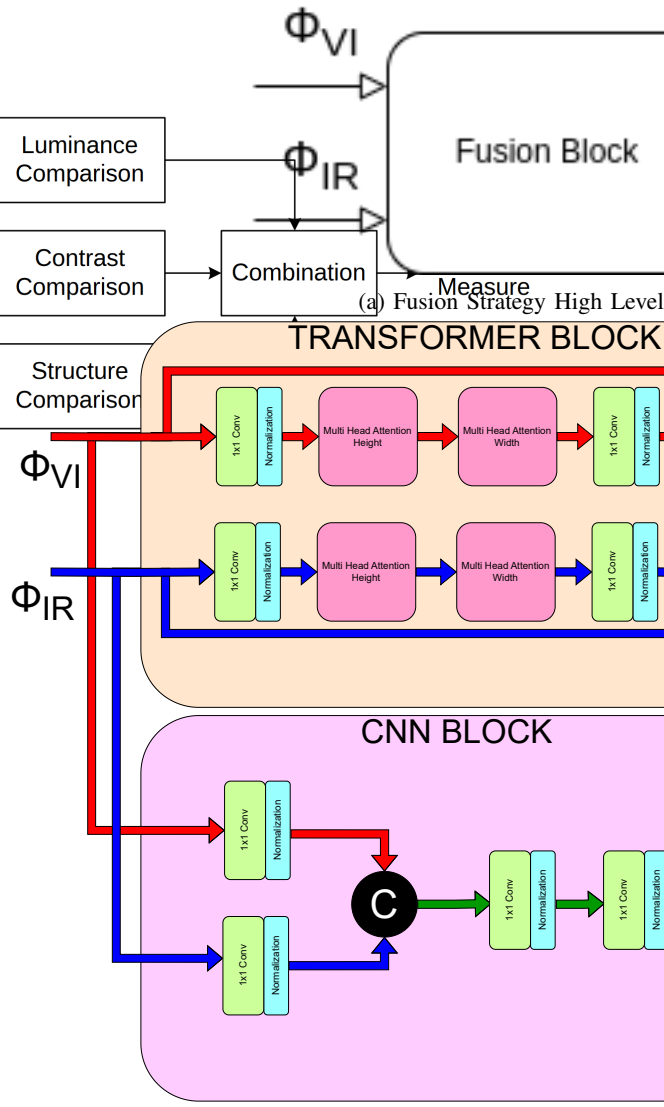


Fig. 5: Diagram of the structural similarity (SSIM) measurement system [29]



(a) Fusion Strategy High Level

(b) Fusion Strategy Detailed Design

Fig. 6: Fusion Strategy

The Structural Similarity Index ($SSIM$) is a metric that quantifies the similarity between two images, yielding values within the range of -1 to 1. A $SSIM(\odot, \odot)$ value of 1 denotes perfect similarity, indicating that the images share **same** characteristics in terms of luminance, contrast, and structure. Conversely, a value close to -1 signifies a substantial dissimilarity between the images. Notably, the $SSIM(\odot, \odot)$ index demonstrates a strong correlation with human perception of image quality, making it widely employed in diverse image processing and computer vision applications [29].

The equation for $SSIM$ ($SSIM(.)$) in Eq. (5) constrains its output to the range of $[-1, 1]$, which consequently bounds the L_{SSIM} loss function (as defined in Eq. (4)) to the interval $[0, 2]$. In this context, lower values of L_{SSIM} indicate better performance with respect to $SSIM$. In contrast, the L_{pixel} loss is unbounded. To balance the impact of both L_{pixel} and L_{SSIM} during training, the trade-off parameter α in Eq. (1) governs their relative magnitudes.

In short, the autoencoder depicted in Figure 4a is subjected to training using the MS-COCO dataset [26] and the Road-Scene dataset [27]. The training process is guided by the loss function presented in Eq. (1). Comprehensive assessments of the autoencoder's performance are presented, encompassing both quantitative and qualitative evaluations. Detailed results can be found in the Section **??**.

3) Stage 2: Fusion Strategy Selection: Referenced in Section III-A1, the process involved in accurately extracting multi-scale feature maps, coupled with the decoding and reconstruction of the original image as illustrated in Figure 3, has been previously discussed. This process is primarily concerned with the extraction of complex features at various scales which are intrinsic to image data. The essence of multi-scale feature extraction is to gather spatially diverse information from the image at different resolutions, thereby allowing for a more robust representation of the image data.

Presently, the emphasis will shift to the application of separate encoders for the extraction of multi-scale features from both visual and infrared band images. The process entails the deployment of these encoders, each uniquely purposed for their respective image band. This is important as different image bands often contain distinct but complimentary information. For instance, **the visual image band, which relies on the visible light spectrum, presents color and texture details, while the infrared band, capturing non-visible light, provides thermal information**.

The outputs from these separate encoders are then merged into a single multi-scale feature map. This fusion process is a crucial step as it combines diverse features from dif-

ferent bands, enhancing the feature representation. It allows the model to leverage the strengths of each band, thereby improving the overall effectiveness of the feature extraction process.

Following the fusion, the resulting multi-scale feature maps are subsequently decoded, leading to the reconstruction of the original image, as depicted in Figure 1. It is important to note that this decoding phase does not merely entail the generation of a visually coherent image. Rather, it reconstructs an image that encapsulates the combined information from both bands. In essence, the reconstructed image, though visually similar to the original, carries a much richer set of features, potentially paving the way for more accurate subsequent analyses or processes.

As delineated in Section I-A, conventional Convolutional Neural Network (CNN) based techniques facilitate image fusion through the amalgamation of local features. However, a significant limitation inherent to these methods is their lack of consideration for the global context that permeates an image. In an attempt to circumvent this limitation, transformer-based models have been introduced, which capitalise on the self-attention mechanism to effectively model the global context.

The development of an innovative approach that synthesises transformer-based models with CNNs is thus postulated. This approach strives to account for local features at multiple scales, paying careful attention to both local and global contexts. As articulated in Section III-A, the method that is proposed adopts a bi-phase training protocol.

The first stage of this training protocol, as elucidated in Section III-A1, necessitates the use of an auto-encoder to extract deep, multi-scale features. In the subsequent stage, these multi-scale features are blended via a fusion strategy that innovatively combines CNNs with Transformers. Comprising a CNN and a transformer branch, the combined fusion blocks capably capture both local and global context features.

Further experimentation with this method was undertaken on a multitude of benchmark datasets, as illustrated in Section **??**. Comparative metrics, as specified in Section **??**, revealed that the method proposed outperformed many existing fusion algorithms. The results demonstrated that the combined CNN-Transformer fusion strategy was effective in capturing a broader context, leading to superior performance in various comparative metrics.

This method's success lies in its ability to leverage the strengths of both CNNs and Transformer models, providing a comprehensive view of an image by capturing both local and global contexts. The strategy's robustness, combined with its efficiency, heralds a new direction for further developments in the field of image fusion. Used combined fusion block details can be seen at Figure 6a.

The fusion network, illustrated in Figure 6b, is characterised by its dual-branch design, which consists of a spatial branch and a transformer branch. The spatial branch integrates convolution layers and a bottleneck layer, specifically tailored to distil local feature representations. Conversely, the transformer branch employs an axial attention-based transformer block to capture the global context embedded within the input data.

The Local Feature Fusion block within the spatial branch operates in a relatively straightforward manner, focusing on exploiting spatial dependencies in the data to extract intricate local features. The structure and functionality of this branch can be observed in Figure 6b.

Meanwhile, for the transformer branch, two alternatives can be considered for attention mechanism deployment: self-attention and axial-attention.

The self-attention mechanism is a strategic process that correlates disparate tokens within a singular sequence to generate a representative sequence. It is an effective method for modelling dependencies without regard to their position in the input. Consider an input feature tensor $x \in \mathbb{R}^{C_{\text{in}} \times H \times W}$ and output feature tensor $y \in \mathbb{R}^{C_{\text{out}} \times H \times W}$. Here, C_{in} and C_{out} respectively denote the quantity of input and output channels, while H and W are indicative of the tensor's height and width, respectively.

The self-attention mechanism can be mathematically formulated as in Eq 6:

$$\begin{aligned} Q &= xW_Q \\ K &= xW_K \\ V &= xW_V \\ A &= \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \\ y &= AV \end{aligned} \tag{6}$$

In these equations, W_Q, W_K, and W_V are weight matrices that are learned. Q, K, and V symbolise the query, key, and value, which are derived from the input tensor x. After obtaining these matrices, the attention scores A are calculated using a softmax function applied to the dot product of Q and K^T, which is further scaled by $1/\sqrt{d}$. The output feature tensor y is then obtained by multiplying the attention scores with the value matrix V.

Alternatively, the axial-attention mechanism, as presented by Ho et al. [30], offers a unique approach to sequence processing. This mechanism, a variant of self-attention, is characterised by its improved computational efficiency. In axial attention, the application of self-attention is executed sequentially over the axes of the feature map's height, followed by the width. This approach significantly reduces computational complexity, thus fostering more efficient operations.

A noteworthy contribution to the axial attention mechanism was proposed by Wang et al. [31], who introduced a learnable positional embedding to the query, key, and value of axial attention. This addition enhances the sensitivity of the affinities to positional information, further improving the performance of the mechanism. These positional embeddings are considered parameters that are learned in conjunction with the training process.

Considering an input x, the self-attention computation along the height axis can be formulated as presented in Eq 7, and along the width axis as shown in Eq 8:

$$y_{ij} = \sum_{h=1}^{H} \text{softmax}\left(q_{ij}^T k_{ih} + q_{ij}^T r_{ih}^q + k_{ij}^T r_{ih}^k\right) \tag{7}$$

Here, r^q, r^k, and $r^v \in \mathbb{R}^{H\times H}$ represent the positional embeddings for the height axis.

$$y_{ij} = \sum_{w=1}^{W} \text{softmax}\left(q_{ij}^T k_{iw} + q_{ij}^T r_{iw}^q + k_{ij}^T r_{iw}^k\right) \tag{8}$$

Here, r^q, r^k, and $r^v \in \mathbb{R}^{W\times W}$ denote the positional embeddings for the width axis.

The axial attention mechanism applies Eq 7 for the height axis and Eq 8 for the width axis, resulting in an efficient self-attention model. This process can be visually examined in Figure 6b.

While dealing with challenges and formulating suitable solutions, axial attention transformers prove to be versatile tools in learning meaningful representations, facilitating a broad array of applications across various fields. The fusion of theoretical insights and practical approaches is vital for constructing resilient axial attention transformer architectures that can proficiently handle complex real-world issues.

Long-Range Dependencies: Axial attention transformers, like most transformer-based models, may struggle with long-range dependencies, especially when processing large sequences or images. This difficulty arises from the transformers' full self-attention mechanism, which may not always capture these dependencies effectively. **Solution**: Introduce position encodings or relative position representations to help the model capture positional relationships between elements. Additionally, axial attention, which factorizes the full attention into separate attention distributions for each dimension, can be used to more effectively capture long-range dependencies.

Quadratic Time Complexity: Transformers, including axial attention transformers, have a quadratic time complexity due to their full self-attention mechanism. This can lead to high computational cost when dealing with large inputs. **Solution**: Apply efficient variants of the attention mechanism, such as Longformer's sliding window attention or Linformer's low-rank approximation, to reduce the time complexity.

Lack of Interpretability: The representations learned by axial attention transformers can be challenging to interpret, similar to other deep learning models. This lack of interpretability makes it difficult to understand what the model has learned and how it makes decisions. **Solution**: Use explainability techniques, such as feature visualization or attention visualization, to gain insights into the model's learned representations and decision-making process.

Overfitting: Axial attention transformers can overfit, particularly when the model is overly complex or the training dataset is limited. Overfitting happens when the model becomes too specialized to the training data, losing its ability to generalize well to unseen data. **Solution**: Incorporate regularization techniques, such as weight decay or dropout, to discourage overfitting. Techniques like early stopping can also be useful in preventing the model from overtraining.

Training Stability: Transformers, including axial attention transformers, can sometimes be difficult to train due to their high complexity and susceptibility to issues like exploding gradients. **Solution**: Use gradient clipping techniques to prevent gradients from exploding during backpropagation. Additionally, adaptive optimization algorithms like Adam, which have built-in mechanisms for dealing with sparse gradients and other challenges, can help ensure stable training.

To tackle the challenges outlined above, our approach commenced with the adoption of an axial attention transformer in Image Fusion Transformer [24] as the base model. This axial transformer can extract multi-level features from input data and reconstruct the original data by decoding these feature maps into a single unified representation. To adjust the model to our specific problem requirements, we re-trained this base axial transformer using the RoadScene [27] dataset, initializing the network with random weights. This training process resulted in the development of a pretrained axial attention transformer that effectively captures crucial features from the input data.

In summary, the dual-branch design of the network offers complementary functionality. The spatial branch focuses on capturing fine-grained local features using a convolutional block and a bottleneck layer. Simultaneously, the transformer branch employs axial attention to distil global context-related features, enabling a comprehensive feature representation from both local and global perspectives.

4) Stage 2: Training of The Autoencoder with Fusion Strategy: As elaborated in Section III-A1, the initial phase of the training protocol is geared towards enabling the encoder network to capture multi-scale deep features. In tandem, the decoder network is trained to reconstruct the input image, utilizing these multi-scale deep features, which the encoder network has been trained to extract. The structure and training scheme for the auto-encoder network can be observed in Figure 4a.

In Section III-A, it was discussed that the second phase of the training regimen involves incorporating the fusion block between the encoder and the decoder, as portrayed in Figure 6a. At this point, we must reassess the prior loss function utilized during the autoencoder training stage, as delineated by Eq 1 in Section III-A2. The goal here is to scrutinize its efficacy for this stage of training.

The fusion loss function, L_{fuse}, can be formulated as in Eq 9:

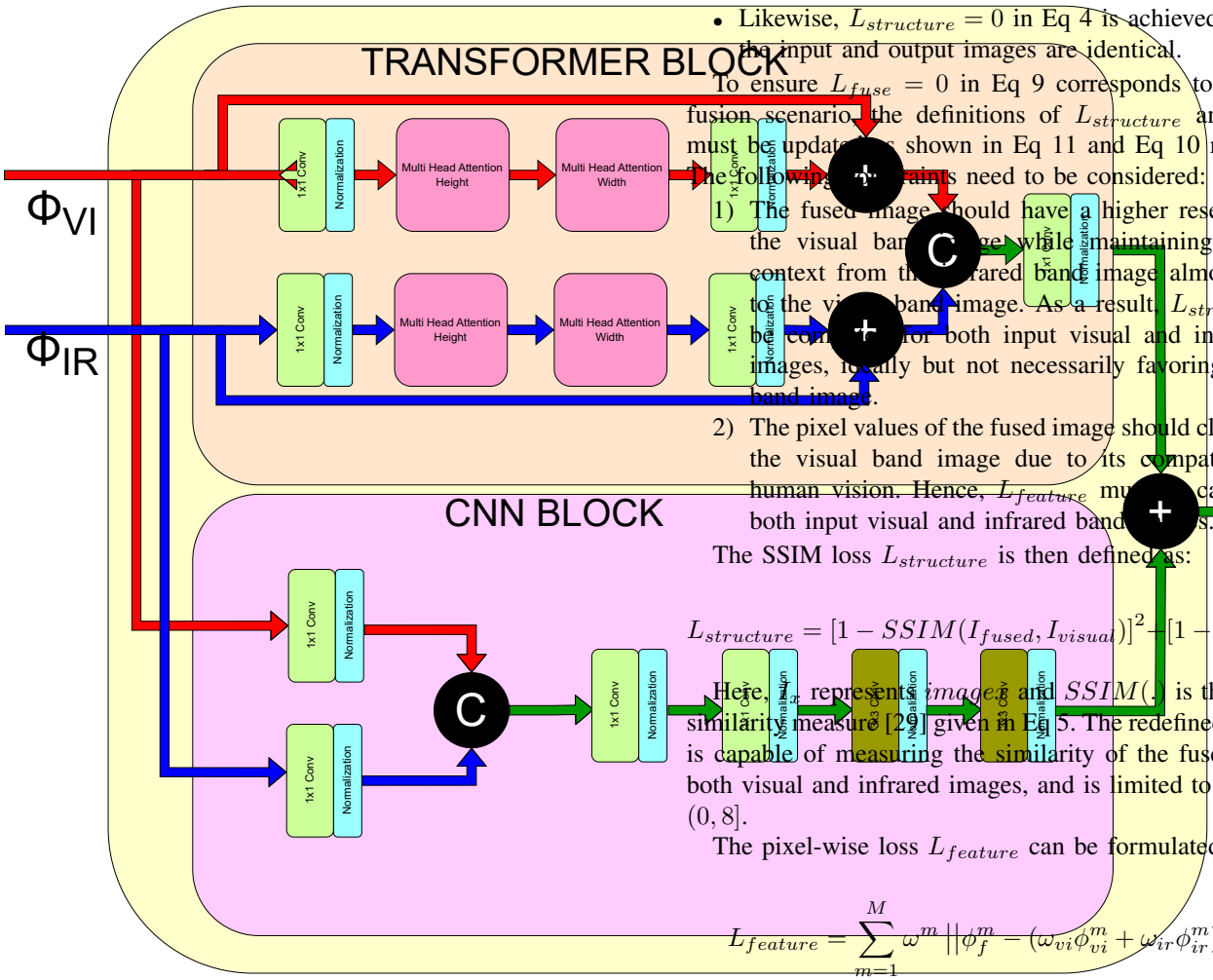


Fig. 7: Detailed Fusion Strategy

$$L_{fuse} = L_{pixel} + \alpha L_{structure} \tag{9}$$

This fusion loss function aims to balance the contribution from pixel-level losses, denoted as $L_{feature}$, and structural similarity losses, $L_{structure}$, modulated by a trade-off factor, α. This mathematical construct becomes critical in the context of our analysis and potential modification of the loss function used in the preceding training stage, thereby introducing an additional degree of intricacy to the model's optimization procedure.

As referenced in Section I-A and Section III-A2, it is clear that the autoencoder loss function is insufficient to meet the needs of the fusion strategy. This is due to the following reasons:

- $L_{fuse} = 0$ in Eq 9 signifies an optimal fusion condition, excluding the overfitting case. This implies that both $L_{feature}$ in Eq 2 and $L_{structure}$ in Eq 4 must independently be equal to zero.
- $L_{feature} = 0$ in Eq 2 occurs only when the input and output images are identical.
- Likewise, $L_{structure} = 0$ in Eq 4 is achieved only when the input and output images are identical.

To ensure $L_{fuse} = 0$ in Eq 9 corresponds to an optimal fusion scenario, the definitions of $L_{structure}$ and $L_{feature}$ must be updated as shown in Eq 11 and Eq 10 respectively. The following constraints need to be considered:

1) The fused image should have a higher resemblance to the visual band image while maintaining the global context from the infrared band image, almost identical to the visual band image. As a result, $L_{structure}$ must be computed for both input visual and infrared band images, ideally but not necessarily favoring the visual band image.
2) The pixel values of the fused image should closely match the visual band image due to its compatibility with human vision. Hence, $L_{feature}$ must be calculated on both input visual and infrared band images.

The SSIM loss $L_{structure}$ is then defined as:

$$L_{structure} = [1 - SSIM(I_{fused}, I_{visual})]^2 + [1 - SSIM(I_{fused}, I_{infrare} \tag{10}$$

Here, I_x represents $image_x$ and $SSIM(.)$ is the structural similarity measure [29] given in Eq 5. The redefined $L_{structure}$ is capable of measuring the similarity of the fused image to both visual and infrared images, and is limited to the interval $(0, 8]$.

The pixel-wise loss $L_{feature}$ can be formulated as:

$$L_{feature} = \sum_{m=1}^{M} \omega^m \left|\left|\phi_f^m - (\omega_{vi}\phi_{vi}^m + \omega_{ir}\phi_{ir}^m)\right|\right|_F^2 \tag{11}$$

Here, M refers to the number of scales for deep feature extraction, while f, vi, and ir denote the fused image, the input visual band image, and the input infrared band image respectively. ω^m, ω_{vi}, and ω_{ir} represent trade-off parameters employed to harmonize the magnitudes of the losses. ϕ_x^m corresponds to the feature maps of $imagex$, which could be either the input or output feature maps of the fusion block, as depicted in Figure 6b.

This loss function restricts the fused deep features to preserve significant structures, thereby enriching the fused feature space with more conspicuous features and preserving detailed information.

REFERENCES

[1] H. Li, X.-J. Wu, and J. Kittler, "Rfn-nest: An end-to-end residual fusion network for infrared and visible images," *Information Fusion*, vol. 73, pp. 72–86, 2021.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

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

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[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] H. Xu, J. Ma, Z. Le, J. Jiang, and X. Guo, "Fusiondn: A unified densely connected network for image fusion," in *proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020.

[28] A. Toet *et al.*, "Tno image fusion dataset¡ https://figshare. com/articles," *TN_Image_Fusion_Dataset/1008029*, 2014.

[29] K. Ma, K. Zeng, and Z. Wang, "Perceptual quality assessment for multi-exposure image fusion," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3345–3356, 2015.

[30] J. Ho, N. Kalchbrenner, D. Weissenborn, and T. Salimans, "Axial attention in multidimensional transformers," *arXiv preprint arXiv:1912.12180*, 2019.

[31] H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille, and L.-C. Chen, "Axial-deeplab: Stand-alone axial-attention for panoptic segmentation," in *European conference on computer vision*, pp. 108–126, Springer, 2020.