# 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 combines data from different sensors to create a single image with enriched information. Despite the widespread use of deep learning in image fusion, current methods often focus on local features, neglecting broader contextual information. Addressing this limitation, we introduce a novel image fusion approach employing a transformer-based multi-scale strategy. This method captures both local and general context, improving overall fusion. Our two-stage training involves an auto-encoder for deep feature extraction and subsequent fusion of multi-scale features using Convolutional Neural Networks (CNNs) and Transformers. Unlike comparable approaches, we propose innovative loss functions to tackle the challenge of fusion without ground truth. Through extensive experiments on benchmark datasets, our method, coupled with the newly defined loss functions, outperforms other fusion algorithms.**

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

## I. Introduction

The evolution of perception systems across species, from insects to apex predators, underscores nature's prowess. Human technological innovation has given rise to mechanical eyes that not only emulate but sometimes surpass their biological counterparts. This union of natural instincts and modern camera technology unleashes a vast realm of perception. Central to computer vision is the light spectrum, extending from visible colors (around 400 to 700 nanometers) to the invisible realms of infrared and ultraviolet waves. While humans are confined to a segment of this spectrum, technology unlocks a wider expanse, particularly through visible and infrared imagery, ushering in a revolutionary phase in computer vision. Image fusion, a pivotal aspect of computer vision, combines multiple image sources to enhance clarity and understanding, with Visual-Infrared Image Fusion (VIF) proving transformative for applications such as night vision and thermal imaging. Various techniques, from traditional pixel and feature-level fusion to advanced methods like CNN-based and transformer-based fusion, are employed to merge visible and infrared images effectively, amplifying situational comprehension across applications.

The advancement of deep learning methods in extracting pertinent features for image fusion necessitates addressing their primary focus on local features, often neglecting the overarching context within the image, as highlighted by Li et al. (2021) [1]. Transformer-based models, leveraging attention mechanisms, offer a solution by capturing wider context dependencies to refine fusion outcomes.

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

In response to the need for enhanced image fusion techniques, our research focuses on several key questions. Firstly, we explore the hypothesis that the new loss function, emphasizing 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 [1]. Secondly, we aim to determine if 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 [1]. Additionally, we investigate whether 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 [2], [3]. Furthermore, we explore whether the proposed transformer-based approach will achieve a better balance between quantitative and qualitative performance in image fusion, overcoming the compromise between the two often observed in traditional deep learning methods [2], [3]. Lastly, we assess whether 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 [2], [3]. Through these research questions, we aim to gain a comprehensive understanding of the complexities surrounding image fusion and contribute to advancements in the field.

## II. Related Work

The RGB and infrared image fusion domain has undergone extensive research, traversing from Traditional Fusion Algorithms to cutting-edge Transformer-based models [4]–[8]. In the early 1990s, methods like Sparse Representation and Multi-scale Transformation were explored, each with its inherent limitations. Traditional algorithms, relying on handcrafted steps, faced challenges in adaptability and time complexity [4]–[8]. The scarcity of labeled datasets for RGB-IR fusion prompted a shift towards unsupervised scenarios, guiding our exploration of Performance Evaluation metrics.

With the advent of deep learning, learning-based algorithms became predominant, categorized by learning methods, loss functions, and the use of labeled datasets [9]–[13]. CNN-based approaches, both supervised and unsupervised, exhibited success in feature extraction for image fusion, yet challenges persisted in scenarios with significant differences in factors like illumination or resolution [9]. Autoencoder-based algorithms, utilizing neural networks for dimensionality reduction, showcased advancements in works such as DenseFuse, Raza et al., and Fu et al. [10]–[12].

GAN-based methods, introduced by Goodfellow et al., focused on unsupervised fusion, integrating attention mechanisms and residual connections for improved performance [13]–[15]. While these approaches demonstrated promise, challenges persisted in effectively handling the inherent differences between fused and source images.

A transformative shift in RGB-IR image fusion occurred with the introduction of Transformer-based algorithms in 2021 [2], [3], [16]. These methodologies, driven by the self-attention mechanism, marked a paradigm shift by efficiently managing long-range dependencies in images. Innovative designs, such as multiscale fusion strategies and dual transformer approaches, were introduced, emphasizing the seamless integration of Transformers with traditional methods [17]–[20]. Unsupervised Transformer-based techniques, reliant on loss functions, eliminated the need for labeled data but posed challenges in methodological evaluations [17], [18]. Ongoing research explores diverse Transformer integrations, Transformer-CNN combinations, and the utilization of auxiliary information to further enrich the fusion process [17]–[20].

This section provides a comprehensive overview of the evolution of RGB-IR image fusion techniques, emphasizing the pivotal role of Transformer-based approaches as the forefront of current research in the field [17]–[20]. The integration of deep learning methods has not only enhanced feature extraction capabilities but has also paved the way for more adaptive and robust solutions in the challenging realm of image fusion.

## III. Methodology

Image fusion can be decomposed into three key components: the feature extractor, feature fuser, and image reconstructor. The feature extractor is responsible for extracting multilevel features from input images. Subsequently, the feature fuser amalgamates these extracted features into unified feature maps for each level. These consolidated features play a crucial role in the final image reconstruction, orchestrated by the image reconstructor.

In terms of analogy, the feature extractor corresponds to the encoder in autoencoders, while the image reconstructor aligns with the decoder in autoencoders. The feature fuser leverages both Convolutional Neural Networks (CNNs) and transformers, with CNNs adeptly managing local features and transformers overseeing global contexts. This amalgamation enhances the fusion process, ultimately aiming to improve accuracy.

Our approach involves a two-stage training process. In the initial stage, we train an autoencoder, originally derived from RFN-Nest [1] as illustrated in Figure **??**. Subsequently, in the second stage, we train our fusion block, depicted in Figure **??**, in conjunction with the previously trained autoencoder.

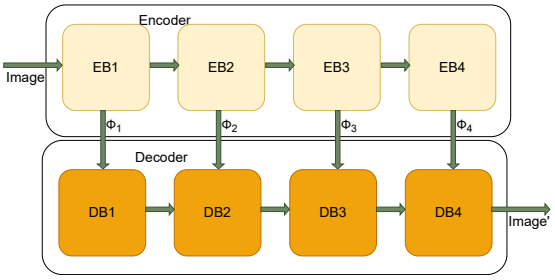
## IV. Stage 1: Autoencoder Selection

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.

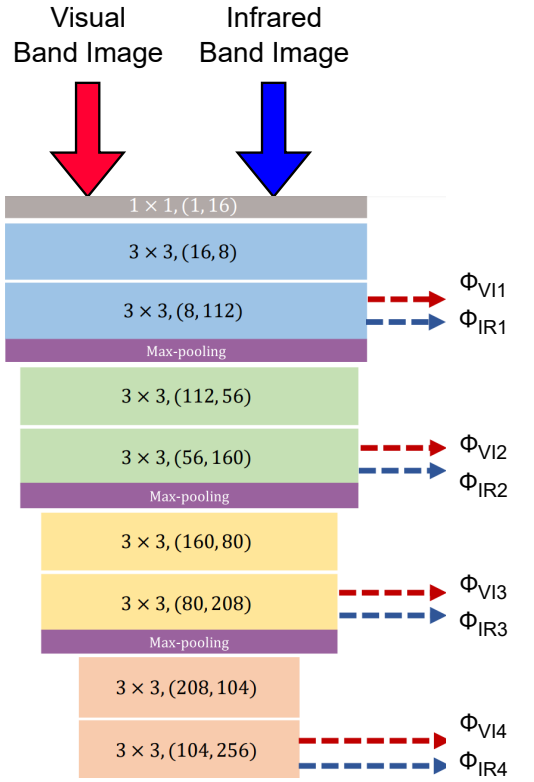
In the realm of autoencoders, choosing the right latent space dimension is crucial. A thorough hyperparameter search, including techniques like cross-validation, guides this selection. Autoencoders face challenges with multimodal data and computational demands, necessitating effective data preprocessing and optimization strategies.

Despite their capabilities, autoencoders may yield features lacking human interpretability. Techniques like activation maximization and feature visualization aid in deciphering model representations. Our approach adopts an autoencoder from RFN-Nest and DenseFuse, fine-tuning it on the MS-COCO and RoadScene datasets for robust feature extraction. Detailed in Section **??**, this staged training produces a proficient autoencoder. Examined on the TNO dataset, the findings illuminate potential applications and future directions for advancing image fusion techniques.

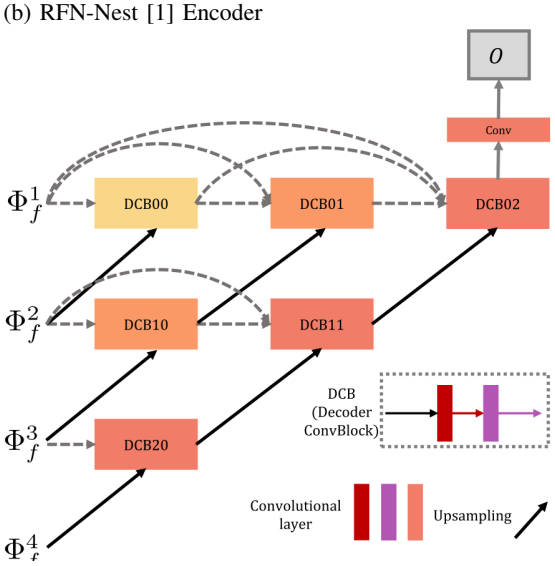
The initial phase of the training process involves instructing the encoder network to capture multi-scale deep features. Concurrently, the decoder network is also trained to reconstruct the input image, utilizing the aforementioned multi-scale deep features. The training framework of the auto-encoder network is depicted in Figure **??**. 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




(a) Block Diagram




(b) RFN-Nest [1] Encoder




(c) RFN-Nest [1] Decoder

Fig. 1: Stage 1 Configuration

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 $\alpha$ 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 $i$th row and $j$th 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 [21] 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 **??**.

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

where:

$x$ and $y$ represent the two images being compared.
$\mu_x$ and $\mu_y$ are the local means of $x$ and $y$ respectively.
$\sigma_x^2$ and $\sigma_y^2$ are the local variances of $x$ and $y$ respectively.
$\sigma_{xy}$ 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.

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 [21].

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 $\alpha$ in Eq. (1) governs their relative magnitudes.

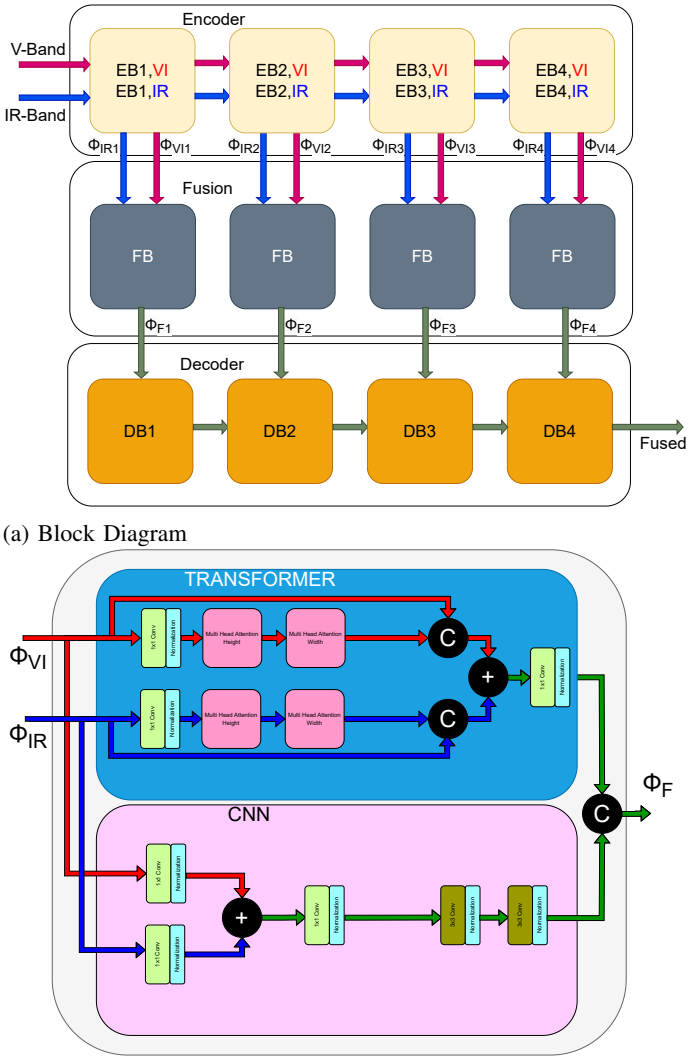
In short, the autoencoder depicted in Figure 1b is subjected to training using the MS-COCO dataset [22] and the Road-Scene dataset [23]. The training process is guided by the loss function presented in Eq. (1). Comprehensive assessments of the autoencoder's performance are presented, encompassing both quantitative and qualitative evaluations.

## V. Stage 2: Fusion Strategy Selection

The process involved in accurately extracting multi-scale feature maps, coupled with the decoding and reconstruction of the original image as illustrated in Figure 1a, 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 with fusion block depicted in Figure 2b per scale. This fusion process is a crucial step as it combines diverse features from different 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.




(a) Block Diagram

(b) Fusion Block

Fig. 2: Stage 2 Configuration

Following the fusion, the resulting multi-scale feature maps are subsequently decoded, leading to the reconstruction of the original image, as depicted in Figure **??**. 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.

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 **??**, the method that is proposed adopts a bi-phase training protocol.

The first stage of this training protocol 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 **??**.

The fusion network, illustrated in Figure **??**, 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 **??**.

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. [24], 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. [25], 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, employing specific equations for height and width axes, as depicted in Eq 7 and

Eq 8, respectively, contributes to the efficacy of a self-attention model, visually illustrated in Figure **??**. This mechanism proves versatile in learning meaningful representations, demonstrating applicability across diverse fields. Constructing resilient axial attention transformer architectures requires a fusion of theoretical insights and practical approaches, a crucial combination for proficiently addressing complex real-world challenges.

Addressing potential issues, axial attention transformers, like other transformer-based models, may encounter difficulties with long-range dependencies, quadratic time complexity, lack of interpretability, overfitting, and training stability. Solutions involve introducing position encodings, efficient attention variants, interpretability techniques, regularization, and gradient clipping. Our approach involves adopting an axial attention transformer as the base model in the Image Fusion Transformer [17], fine-tuned on the RoadScene [23] dataset, resulting in a pretrained model adept at capturing essential features. In summary, the dual-branch design of the network, incorporating a spatial branch for local features and a transformer branch for global context, provides a comprehensive feature representation, making it a promising tool for various applications.

In the initial training phase, the focus is on empowering the encoder network to capture multi-scale deep features. Simultaneously, the decoder network is trained to reconstruct the input image by leveraging these multi-scale deep features, extracted by the trained encoder network. The autoencoder network's structure and training scheme are depicted in Figure 1b, illustrating the coordinated learning process.

Moving forward to the second phase, detailed in Section **??**, the training regimen introduces the fusion block between the encoder and decoder, illustrated in Figure **??**. This crucial step prompts a reassessment of the initial loss function employed during the autoencoder training stage, specified by Eq 1 in Section **??**. The objective is to scrutinize the effectiveness of this loss function for the current stage of training, ensuring its alignment with the fusion block integration.

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

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

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, $\alpha$. 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.

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:

$$L_{structure} = [1 - SSIM(I_f, I_v)]^2 + [1 - SSIM(I_f, I_i)]^2 \tag{10}$$

Here, $I_x$ represents $imagex$ and $SSIM(.)$ is the structural similarity measure [21] 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. $\omega^m$, $\omega_{vi}$, and $\omega_{ir}$ represent trade-off parameters employed to harmonize the magnitudes of the losses. $\phi_x^m$ 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 2.

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.

## VI. Results

After carefully executed Stage-I and Stage-II training, in this section, we will design comprehensive experiments to probe the hypotheses detailed below. With the stated hypotheses, we aim to extend the envelope of existing understanding and contribute novel insights into the domain of image fusion. We expect that the carefully designed experimental procedure

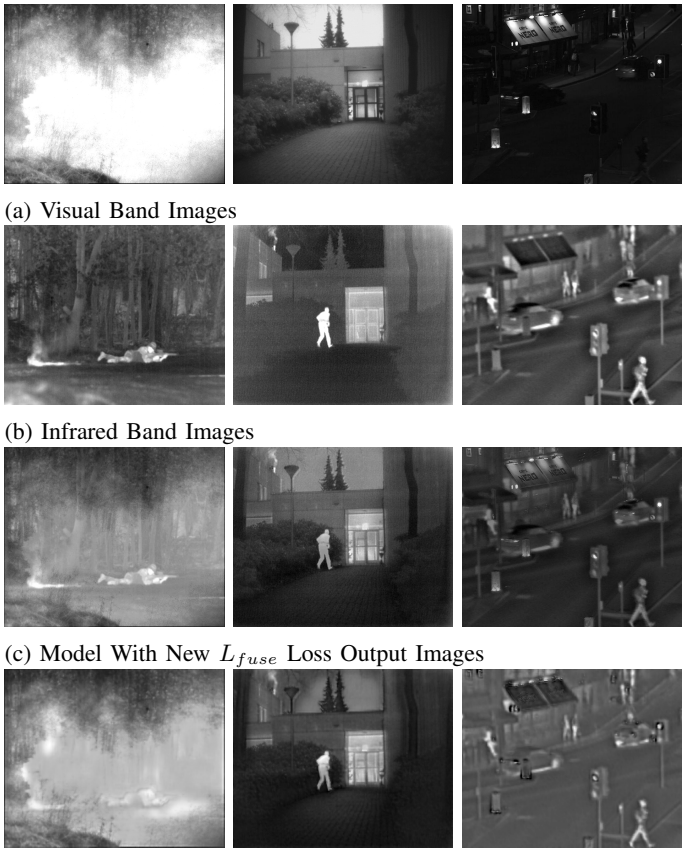
will offer a robust platform to scrutinize these hypotheses, shed light on the latent aspects, and help us further refine our model's capability and performance. We will be testing the capabilities of the model into two different approach: capabilities of the new loss function and the model.

### A. Study I: New Loss Function

Employing our novel loss function, which prioritizes the similarity between visible and infrared inputs, is anticipated to yield more informative fused images compared to using the Structural Similarity Metric (SSIM). To validate this hypothesis, the proposed loss function ($L_{fuse}$) was implemented, emphasizing the preservation of significant features from both input sources. Multiple model versions were trained using different loss functions—one utilizing the traditional SSIM loss and the other employing our newly proposed loss function, with all other parameters held constant for a fair comparison. The evaluation involved qualitative analysis, assessing details, contrast, sharpness, and overall perceptual quality of the fused images, as well as quantitative comparisons using metrics such as $pSSIM$ and mutual information. By scrutinizing the quality and information retention in the resulting fused images, this comprehensive approach allows us to ascertain the validity of the hypothesis. The newly proposed loss function exhibits commendable abilities in both qualitative and quantitative results, as evident from Figure 3 and Table I. Notably, when evaluated at nearly identical SSIM scores, the loss function proposed in Eq. 9 demonstrates superior performance compared to the conventionally adopted loss function detailed in Eq. 1. The inclusion of RFN-Nest results ensures a comprehensive assessment, highlighting the potential of our methodology to outperform traditional standards, especially in terms of the SSIM metric. This comparison underscores the significance of the innovative approach embodied by the newly proposed loss function, offering promising insights and paving the way for advancements in image fusion and related fields.

The comparison our model with leading image fusion methods are given in Figure 6. We'll evaluate them based on visual quality and the quantitative metrics. As highlighted in Section II, current top methods include transformer-based techniques like M3FD [29] and IFT [17], GAN-based approaches such as DenseFuse [10] and SwinFusion [30], and the autoencoder method RFN-Nest [1]. Our goal is to provide a clear and thorough comparison to understand the strengths and limitations of each method in the field of image fusion.

Upon a qualitative assessment of Figure **??**, several observations come to the fore. The first column distinctly showcases the imaging of a soldier amidst smoke, and this rendition, with its meticulous portrayal of the surrounding details, stands qualitatively superior to its counterparts. A similar superiority can be discerned in the second column, where the depiction of the man concealed behind the tree is accentuated by the intricate details of the tree's broader context, setting it apart from other state-of-the-art methods presented. The final column further underscores the prowess of our method. Here, one can distinctly discern the brand name on a shop awning



(a) Visual Band Images

(b) Infrared Band Images

(c) Model With New $L_{fuse}$ Loss Output Images

(d) Model With $L_{fuse} = L_{ae}$ Loss Output Images

Fig. 3: Loss Functions Comparison

and identify pedestrians on the street, all the while maintaining the fidelity of details from the original visual band images. Turning our attention to Table II, it becomes evident that our method eclipses others in performance across almost all metrics. The sole exception is the (Entropy [26]) metric. It's worth noting that entropy gauges the degree to which pixel values in an image are non-redundant. Consequently, original infrared images, as depicted in Figure 3b, would naturally register the highest entropy scores. This distinction provides a nuanced understanding of where and how our method stands in relation to others in the domain.

### B. Study II: A Unique Transformer Based Fusion Strategy

In this part, our research explores the application of transformer-based models for visual and infrared image fusion, focusing on autoencoders for feature extraction and representation learning. Rooted in hypotheses, our study aims to enhance night vision, medical imaging, and surveillance tasks (**Hypothesis II-1**), achieve a balance between quantitative and qualitative performance (**Hypothesis II-2**), demonstrate computational efficiency for real-time applications (**Hypothesis II-3**), and address challenges through model tuning and regularization (**Hypothesis II-4**). Our proposed approach combines autoencoder-derived features from visual and infrared images using a transformer-based model, aiming to balance perfor-

TABLE I: Loss Functions Comparison

| **Folder** | **Entropy [26]↑** | **SCD [27]↓** | **MI [28]↑** | **SSIM [21]↑** |
|---|---|---|---|---|
| Proposed $L_{fuse}$ in Eq 9 | 4.536 | **5.433** | **1.591** | **0.884** |
| $L_{ae}$ as $L_{fuse}$ Exp | 4.559 | 6.466 | 0.552 | 0.879 |
| RFN-Nest [1] | **4.729** | 7.062 | 0.602 | 0.541 |

TABLE II: Comparison with SoTA

| **Method** | **Entropy [26]↑** | **SCD [27]↓** | **MI [28]↑** | **SSIM [21]↑** |
|---|---|---|---|---|
| Ours | 4.536 | **5.433** | **1.591** | **0.884** |
| SwinFusion [30] | 4.605 | 6.760 | 0.804 | 0.690 |
| M3FD [29] | 4.625 | 6.858 | 0.742 | 0.659 |
| IFT [17] | 4.644 | 6.864 | 0.684 | 0.630 |
| DenseFuse [10] | 4.724 | 6.455 | 0.853 | 0.588 |
| RFN-Nest [1] | **4.729** | 7.062 | 0.602 | 0.541 |

(a) Visual Band Images

(b) Infrared Band Images

(c) Our Output Images

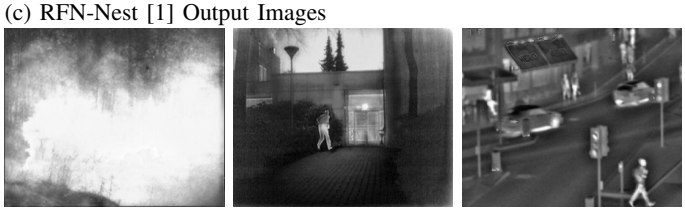
(d) SwinFusion [30] Output Images

Fig. 4: Comparison with SoTA

(a) M3FD [29] Output Images

(b) IFT [17] Output Images

(c) RFN-Nest [1] Output Images



(d) DenseFuse [10] Output Images

Fig. 5: Comparison with SoTA *cont'd*

mance and efficiency. Experiments, conducted on the TNO dataset, evaluate these hypotheses, covering diverse tasks and assessments. Results will yield insights into the effectiveness and efficiency of transformer-based image fusion techniques, advancing our understanding in this domain.

From Figure 6, the first column showcases an instance of night vision imagery. This particular example serves as a suitable representation for illustrating long-range dependencies and the global context inherent in such images. The second and third columns, on the other hand, predominantly depict scenarios captured under low light conditions. While at a qualitative glance, the distinctions between the images might appear subtle, a closer examination reveals nuanced differences. Turning our attention to Table **??**, a comprehensive evaluation indicates that our proposed method consistently delivers superior results. In comparison to state-of-the-art (SoTA) techniques, our method exhibits marked improvements across nearly every evaluated performance metric.

(a) Visual Band Images

(b) Infrared Band Images

(c) Our Output Images

(d) SwinFusion [30]Output Images

Fig. 6: Night Vision Enhancement

(a) M3FD [29] Output Images

(b) IFT [17] Output Images

(c) RFN-Nest [1] Output Images

(d) DenseFuse [10] Output Images

Fig. 7: Night Vision Enhancement *cont'd*

## 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] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[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] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[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] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

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

[25] 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.

[26] J. W. Roberts, J. A. Van Aardt, and F. B. Ahmed, "Assessment of image fusion procedures using entropy, image quality, and multispectral classification," *Journal of Applied Remote Sensing*, vol. 2, no. 1, p. 023522, 2008.

[27] V. Aslantas and E. Bendes, "A new image quality metric for image fusion: The sum of the correlations of differences," *Aeu-international Journal of electronics and communications*, vol. 69, no. 12, pp. 1890–1896, 2015.

[28] G. Qu, D. Zhang, and P. Yan, "Information measure for performance of image fusion," *Electronics letters*, vol. 38, no. 7, p. 1, 2002.

[29] J. Liu, X. Fan, Z. Huang, G. Wu, R. Liu, W. Zhong, and Z. Luo, "Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5802–5811, 2022.

[30] 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.