TAFIE: Transformer-Assisted Fusion with Integrated Entropy Attention for Multimodal Medical Imaging

Abhinav Sagar[1]
[1]Vrije Universiteit Brussel (VUB)
abhinav.sagar@vub.be

Abstract

Multimodal medical image fusion aims to integrate complementary information from different imaging modalities to enhance clinical diagnosis and surgical navigation. While deep learning-based approaches have significantly advanced fusion quality over traditional methods by leveraging powerful feature extraction, challenges such as blurring, noise, and artifacts persist in the fused results. To address these issues, we propose a novel fusion framework that incorporates an entropy-based attention module to emphasize salient image regions. Our architecture is designed in a multi-scale manner, utilizing an adaptive gating mechanism to effectively extract and combine salient features across different scales. Additionally, we introduce a Top-K token vision transformer to enable efficient global feature extraction while reducing computational overhead by restricting the context space. We further demonstrate the effectiveness of our fused representations in the downstream task of oocyte quality prediction, showing improved accuracy over individual focal images as well as over other approaches. Extensive experiments on diverse medical imaging datasets demonstrate that our method achieves competitive performance compared to state-of-the-art techniques, both quantitatively and visually. Ablation studies underscore the importance of each proposed component.

Introduction

The goal of image fusion is to integrate the common and complementary information from the source images into one high-quality fused image. The image obtained from a single sensor typically lacks sufficient information to be useful in real-world applications. For example, visible sensors can produce images with rich texture details, while infrared sensors produce images with distinguished targets. Image fusion has a wide range of applications, such as medical diagnosis, object tracking, surveillance, and semantic segmentation.

Multimodal medical image fusion has drawn a lot of attention of late due to its significant clinical applications for cell classification, tumor segmentation, and treatment for high-grade gliomas. PET (Positron Emission Tomography) and SPECT (Single-Photon Emission Computed Tomography) images reflect metabolic information and aid in the diagnosis of vascular diseases and tumor detection, but suffer from low resolution. On the other hand, an MRI (Magnetic Resonance Imaging) image contains rich anatomical information in high resolution to facilitate adequate recognition of soft tissues. Due to complementary information in both images, physicians analyze the PET/SPECT and MRI images separately, which is not only time-consuming but also error-prone. To solve this problem, image fusion could help retain the complementary information from both images in a single fused image.

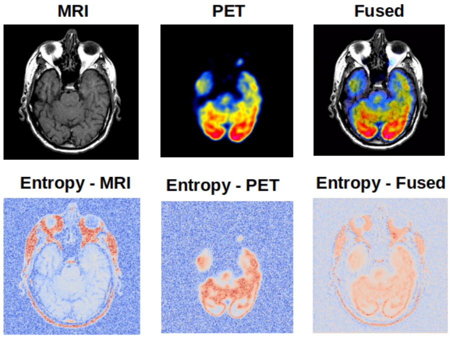

Figure 1: Entropy map visualization using a sample of the PET-MRI dataset. The top row (left to right) shows the MRI, PET, and fused image produced by our model. The bottom row (left to right) presents the corresponding entropy maps for the MRI, PET, and fused image. In the entropy maps, regions highlighted in red indicate higher entropy, corresponding to areas with greater information content.

Despite the potential of existing multimodal image fusion techniques, achieving effective integration of complementary information remains challenging. Differences in imaging physics and acquisition protocols lead to a modality gap, making direct pixel-wise fusion difficult. Furthermore, medical images often contain noise, artifacts, and redundant information, which can degrade the quality of the fused image if not properly handled. Conventional methods often struggle to preserve salient structures while suppressing irrelevant information, and many deep learning models fail to explicitly model the information richness or uncertainty of different regions within the images. These limitations motivate the need for a fusion framework that not only integrates complementary modalities but also selectively emphasizes informative regions, ensuring both structural fidelity and diagnostic relevance.

The primary contribution of this work is the introduction of an entropy-based attention module designed to selectively emphasize salient and information-rich regions within the image. An example entropy map generated from the PET-MRI dataset is provided in Figure 1, illustrating the module's ability to highlight structurally relevant areas.

The three main contributions of our paper are as follows:

- A dual-branch network architecture is proposed that combines an entropy-attention based CNN and a top-k token vision transformer for extracting both the local and global features efficiently.
- A dynamic convolution with learnable gates is suggested to further enhance the feature extraction at multiple scales.
- Using PET-MRI, SPECT-MRI, CT-MRI, and our private oocytes datasets, we demonstrate that our method performs comparably to other state-of-the-art methods using both visual quality and quantitative metrics.
- We show that incorporating our method leads to significant gains in the downstream task of oocyte image quality prediction.

Related Work

Existing image fusion methods using deep learning can be broadly categorized into three main types: autoencoder-based, CNN-based, and transformer-based.

Autoencoder-Based Image Fusion

The autoencoder-based method typically trains the encoder and decoder by reconstructing images from the training dataset. The encoder is used as a feature extractor, and the decoder is used to obtain the fused image with a fusion block in between the encoder and decoder to fuse the features. The encoder and decoder are used for feature extraction and image reconstruction, respectively. Manually designed fusion rules, including addition, concatenation, and multiplication, are used in the fusion block. (Li and Wu 2018) proposed a DenseNet-based end-to-end method using an encoder, a decoder, and a fusion block, with the dense connections being used to obtain more complementary information from the images by reusing the features.

CNN-Based Image Fusion

Dense connections in the U-shaped model were used by (Li, Wu, and Durrani 2020) to enhance the salient features in the fused image while retaining the detailed information from the source images, along with a channel and spatial attention module for further improving the feature extraction. An end-to-end fully convolutional-based image fusion network was suggested by (Zhang et al. 2020), using feature extraction, feature fusion, and image reconstruction blocks, along with perceptual loss for training to incorporate more textural information. (Xu et al. 2020) designed an unsupervised multi-task image fusion method using a Densenet-based model with adaptive information preservation.

Transformer-Based Image Fusion

CNN typically has a small receptive field, thus making it challenging to model long-range dependencies. Transformers solve this problem using self-attention by breaking down the image into patches and modeling the relationship between patches. (Tang et al. 2022c) designed an adaptive convolution and an adaptive transformer-based architecture for medical image fusion. Similarly, (Tang et al. 2023b) proposed a dual attention transformer-based method to model long-range dependencies and pay more attention to vital features. (Park, Vien, and Lee 2023) came up with a cross-modal transformer to preserve complementary information from the source images by removing redundancies in both spatial and channel domains.

Similarly, (Wang et al. 2024) proposed a CNN-Transformer-based method using self and cross attention for effective feature communication and merging complementary characteristics across spatial and channel locations. On the other hand, (Tang and He 2024) suggested a dual transformer-based model to effectively communicate and leverage multi-scale properties for local and global feature extraction and fusion. A transformer-based model was proposed by (Tang, He, and Liu 2024) using interactive attention and residual attention for salient feature extraction and a cross-modal attention to model long-range contextual information. Hybrid CNN-transformer architectures have been proposed to simultaneously capture shallow local features via CNNs and global spatial-channel relationships via transformers (Chen et al. 2023).

A dual CNN and Restormer-based network was introduced by (Zhao et al. 2023) for fusion of both local and global features to better capture the distinct modality-specific and modality-shared features, along with a correlation-based decomposition loss to make the cross-modality low-frequency features correlated while decorrelating the high-frequency shared features. (Sun et al. 2025) incorporated a Differential Convolution Amplification Module-based network to extract both local and global features to prevent spatial information loss by integrating complementary and shared features. Similarly, (Sun, Dong, and Zhu 2025) proposed a method using KAN for local feature extraction by focusing on the localized areas of the image and Mamba for global feature extraction by modeling the information exchange between different channels.

Methodology

Problem Definition

In cases of PET-MRI and SPECT-MRI datasets, one image is in RGB ($I_1 \in R^{H \times W \times 3}$) format, while the other is in grayscale ($I_2 \in R^{H \times W \times 1}$) format. The fusion objective is to combine the original input images into a single RGB fused image ($I_f \in R^{H \times W \times 3}$) that preserves key information from the inputs. We address the channel mismatch issue of fusing a 3-channel image with a 1-channel image by converting the RGB image to YUV and extracting its Y (grayscale or brightness), U (chrominance), and V (chrominance) components. The Y component of the YUV image is fed along with the grayscale image into our network. The fused image is obtained by converting from YUV back to RGB. In the case of the CT-MRI dataset, both images are in GrayScale ($I_1 \in R^{H\times W\times 1}$) format. The fusion objective is to combine the original input images into a single gray-scale fused image ($I_f \in R^{H\times W\times 1}$). Finally, in the case of the Oocytes dataset, 11 images are in GrayScale ($I_1 \in R^{H\times W\times 1}$) format. The objective is to combine the 11 images into a single gray-scale fused image ($I_f \in R^{H\times W\times 1}$).

Model Architecture

We design a novel network architecture using edge encoders, dynamic convolution, entropy-based attention, and top-k vision transformers as detailed in the following subsections:

Edge Encoders We use separate edge encoders for the two modalities to extract edge-based features from input images using Sobel filtering, followed by a convolutional transformation. The module uses fixed Sobel kernels to compute gradients along the horizontal (edge_x) and vertical (edge_y) directions. These filters are not learnable and serve to highlight intensity changes (edges) in the input image. The horizontal and vertical gradients are combined using the Euclidean norm to yield an overall edge magnitude map as in:

$$\text{edge} = \sqrt{\text{edge}_x^2 + \text{edge}_y^2} \tag{1}$$

The resulting edge map is then passed through a learnable convolutional layer.

By combining fixed edge detectors with trainable layers, our model is expected to retain strong inductive biases for edge structures while enabling flexible learning.

Dynamic Convolution The dynamic convolutional block is made up of convolutional layers that dynamically select and fuse convolutional kernels with different receptive fields based on input features. Our design is able to adaptively capture multi-scale spatial features for better representation.

Let $X \in R^{b\times c\times h\times w}$ and b, c, h, and w are the batch size, number of channels, height, and width of the image, respectively. A 3×3 and 5×5 convolution with padding 1 and 2, respectively, is used for spatial size preservation, denoted in:

$$\mathbf{Y}_1 = \text{Conv}_{3\times 3}(\mathbf{X}), \quad \mathbf{Y}_2 = \text{Conv}_{5\times 5}(\mathbf{X}) \tag{2}$$

Where $Y_1 \in R^{b\times c\times h\times w}$, $Y_2 \in R^{b\times c\times h\times w}$. A gating network is used for adaptively weighing the contribution of each convolutional layer. A Global Average Pooling (GAP) layer is used to reduce each channel's information to a single scalar. Then, the pooled features are flattened to a vector as in Equation 3:

$$\mathbf{z} = \text{Flatten(AdaptiveAvgPool(X))} \in R^{B\times C} \tag{3}$$

A linear layer is used to convert this vector to a 2-dimensional vector representing weights (one for each convolution). A Softmax activation normalizes these weights to sum to 1, as shown below:

$$\mathbf{w} = \text{Softmax}\left(\mathbf{W}_g\mathbf{z} + \mathbf{b}_g\right) \in R^{B\times 2} \tag{4}$$

Where W_g and b_g are learnable parameters of the gating linear layer. The output of each convolutional kernel is multiplied by its corresponding weight from the gating mechanism, and the results are summed to produce the final output as in Equation 5:

$$\mathbf{Y} = \text{ReLU}\left(w_1\mathbf{Y}_1 + w_2\mathbf{Y}_2\right) \tag{5}$$

Where $Y \in R^{b\times c\times h\times w}$.

Entropy-Guided Attention Module The entropy-guided attention module is designed to emphasize spatial locations of high uncertainty by leveraging entropy as a signal for feature ambiguity. Given an input tensor $\mathbf{X} \in \mathbb{R}^{B\times C\times H\times W}$, the module first applies a channel-wise softmax normalization to transform raw activations into a probabilistic distribution at each spatial location:

$$p_{b,c,h,w} = \frac{e^{X_{b,c,h,w}}}{\sum_{c'} e^{X_{b,c',h,w}}} \tag{6}$$

The entropy map, which quantifies uncertainty over the channel dimension, is then computed as:

$$E_{b,h,w} = -\sum_c p_{b,c,h,w} \log(p_{b,c,h,w} + \varepsilon) \tag{7}$$

Here, $\varepsilon = 10^{-6}$ is a small constant added for numerical stability to avoid the logarithm of zero. This entropy map is broadcast along the channel dimension and used to scale the input tensor, emphasizing regions of high uncertainty:

$$\tilde{X}_{b,c,h,w} = X_{b,c,h,w} \cdot E_{b,h,w} \tag{8}$$

The scaled features $\tilde{\mathbf{X}}_{b,c,h,w}$ are passed through a lightweight projection network comprising two 1×1 convolutional layers with a ReLU activation in between, which learns a refined attention-aware representation:

$$\mathbf{Y} = \text{Conv}_{1\times 1}\big(\text{ReLU}(\text{Conv}_{1\times 1}(\tilde{\mathbf{X}}))\big) + \mathbf{X} \tag{9}$$

The small convolutional block enables the network to learn complex transformations of the scaled features, effectively learning to convert uncertainty cues into attention weights. The residual connection helps facilitate gradient flow and preserves original information.

Top K Token Vision Transformer CNN-based models alone are good at local feature extraction but are unable to model long-range dependencies. Vision transformers solve this problem, but have a high computational complexity. To balance global modeling capacity with computational efficiency, we propose a token importance-based sparse attention mechanism that selects a subset of the most informative tokens for attention computation.

Given an input tensor $\mathbf{X} \in \mathbb{R}^{B\times N\times C}$, where B is the batch size, N is the number of tokens, and C is the embedding dimension, the input is linearly projected into queries, keys, and values using a shared projection layer:

$$[\mathbf{Q}; \mathbf{K}; \mathbf{V}] = \text{reshape}\left(W_{\text{qkv}}\mathbf{X}\right), \quad W_{\text{qkv}} \in \mathbb{R}^{C\times 3C} \tag{10}$$

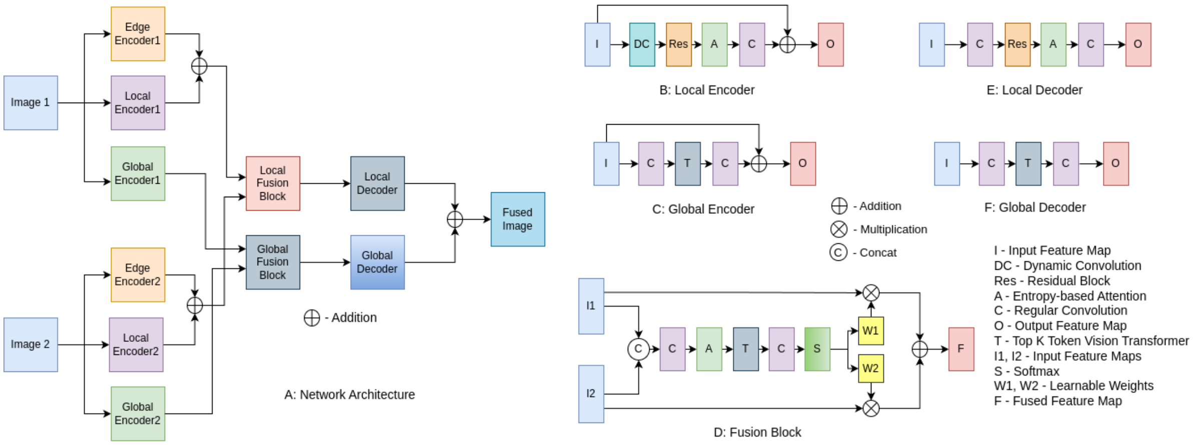

Figure 2: A: Schematic of the complete network architecture. B–F: Detailed illustrations of the individual components, including the local encoder, global encoder, fusion block, local decoder, and global decoder.

These projections are reshaped and split into H heads, each operating in a subspace of dimension $D = C/H$. To guide sparsity, an importance score is assigned to each token through a learnable linear projection:

$$s_i = W_{\text{imp}} x_i, \quad s_i \in \mathbb{R} \tag{11}$$

where $W_{\text{imp}} \in \mathbb{R}^{C\times 1}$ is the learned token scoring weight. For each input sequence, the top-K tokens with the highest scores are selected. The keys and values are then gathered using the top-K indices, resulting in:

$$\mathbf{K}_{\text{topk}}, \mathbf{V}_{\text{topk}} \in \mathbb{R}^{B\times H\times K\times D} \tag{12}$$

We use a K value of 8 in our work. Attention is then computed between all queries and only the selected top-K keys:

$$\text{Attn} = \text{softmax}\left(\frac{\mathbf{Q}\cdot\mathbf{K}_{\text{topk}}^{\top}}{\sqrt{D}}\right) \tag{13}$$

The attention output is the weighted sum over the corresponding values:

$$\mathbf{O} = \text{Attn}\cdot\mathbf{V}_{\text{topk}}, \quad \mathbf{O} \in \mathbb{R}^{B\times H\times N\times D} \tag{14}$$

The outputs from all heads are then concatenated and passed through a final linear projection to restore the original embedding dimension:

$$\text{Output} = W_{\text{out}}\cdot\text{Concat}(\mathbf{O}_1,\ldots,\mathbf{O}_H), \quad W_{\text{out}} \in \mathbb{R}^{C\times C} \tag{15}$$

This mechanism dynamically allocates attention resources to the most informative tokens, reducing attention complexity from $O(N^2)$ to $O(NK)$ per head. By learning token importance during training, the model adapts to focus computation on semantically rich regions, achieving a favorable trade-off between efficiency and performance.

Overall Architecture (TAFIE) We propose a dual-branch architecture named TAFIE that separates local and global feature extraction to better capture complementary information from source images. Each input image is processed by three parallel encoders: a local encoder, a global encoder, and an edge encoder. Features from the local and edge encoders are combined and passed through a local fusion block, while features from the global encoders are directed to a global fusion block. The outputs of these fusion blocks are then decoded by separate local and global decoders, and their reconstructions are summed to produce the final fused image. Using separate decoder and fusion blocks allows each type of feature to be fused using mechanisms tailored to their nature (i.e., local fusion handles fine-grained textures and global fusion captures more of the semantic structures). This, in a way, addresses the issue of the semantic gap that exists between the two types of features. After fusion, the separate decoders (Local Decoder and Global Decoder) can independently reconstruct detailed and coarse components of the final image. This separation allows the network to learn domain-specific reconstruction cues without interference between local and global semantics.

The local encoder consists of a dynamic convolution layer followed by a residual block, an entropy-based attention module, and a 1×1 convolution. Residual connections are employed to preserve the integrity of the original feature representations.

The global encoder includes a 3×3 convolutional layer, a Top-K token vision transformer with an embedding dimension of 96, and a 1×1 convolution. Residual connections are similarly used to maintain original feature information. Complementing the transformer in the global encoder with a convolutional layer injects inductive bias (locality and translation invariance) into the network.

In the fusion block, the respective features extracted from the two source images are concatenated and processed through a 3×3 convolution, an entropy-based attention module, a Top-K token vision transformer, and another 3×3 convolution. The resulting features are passed through a softmax layer to generate learnable weights W_1 and W_2, which are element-wise multiplied with the respective source features I_1 and I_2. The final fused feature F_{out} is given as:

$$F_{out} = I_1 \odot W_1 + I_2 \odot W_2, \tag{16}$$

where $\odot$ denotes the elementwise multiplication.

The local decoder consists of a 3×3 convolutional layer, followed by a residual block, an entropy-based attention module, and another 3×3 convolutional layer. This design enables effective reconstruction of locally fused features while preserving fine-grained details.

The global decoder is composed of a 3×3 convolutional layer, a Top-K token vision transformer with an embedding dimension of 96 for capturing long-range dependencies, and a final 3×3 convolutional layer. This configuration facilitates the reconstruction of globally fused features with enhanced contextual understanding.

The complete network architecture diagram is shown in Figure 2.

While the original TAFIE architecture was designed for fusing two source images, we extend it to handle 11 images captured at different focal lengths for tasks such as oocyte quality prediction. In this setup, each of the 11 input images is independently processed by the local, global, and edge encoders. Features from all local and edge encoders are aggregated through a multi-image local fusion block, while features from all global encoders are similarly fused via a multi-image global fusion block. The aggregated features are then decoded separately by the local and global decoders, and their reconstructions are summed to generate the final fused image. This extension allows the network to simultaneously leverage complementary information from multiple focal planes, enhancing the preservation of fine-grained textures and global semantic structures. By explicitly separating local and global fusion for multiple inputs, the network can effectively address the increased semantic gap and maintain domain-specific reconstruction cues across all images, improving the quality and robustness of the fused representation for downstream prediction tasks.

Loss Function

We combine several loss functions for training our model in an unsupervised manner. Structural similarity (SSIM) loss retains the structural information from the source images by quantifying the similarity between the fused image and the source images in terms of brightness, contrast, and structure. The SSIM loss is defined as in Equation 17 and Equation 18:

$$\operatorname{SSIM}(X, Y) = \frac{\left(2 \mu_X \mu_Y + C_1\right)\left(2 \sigma_{XY} + C_2\right)}{\left(\mu_X{}^2 + \mu_Y{}^2 + C_1\right)\left(\sigma_X^2 + \sigma_Y^2 + C_2\right)}, \tag{17}$$

$$L_{\text{ssim}} = 1 - \operatorname{SSIM}\left(I_{\mathrm{Y}}, I_{\mathrm{X}}\right), \tag{18}$$

where X and Y represent the source and fused images, respectively. (μ_X, σ_X) denote the mean and standard deviation of X and (μ_Y, σ_Y) denote the mean and standard deviation of Y. The correlation between X and Y is represented as $\sigma_{X,Y}$. $C_1 = 0.01$ and $C_2 = 0.03$ are set as constants in our experiments.

Gradient loss is used for keeping the edge details. The gradient loss, computed using the Sobel operator, is defined as:

$$L_{\text{gradient}} = \frac{1}{HW} \left\| \left| \nabla I_f \right| - \max\left(\left| \nabla I_{inp1} \right|, \left| \nabla I_{inp2} \right| \right) \right\|_1, \tag{19}$$

where ∇ denotes the gradient operator that measures the texture information in an image and that is computed using the Sobel operation. H and W represent the height and width of the two source images (I_{inp1}, I_{inp2}), respectively. I_f represents the fused image.

The intensity loss retains the pixel intensity information and is defined as in:

$$L_{int} = \frac{1}{HW} \left\| I_f - \max\left(I_{inp1}, I_{inp2}\right) \right\|_1, \tag{20}$$

where max denotes elementwise maximum operation, and $||.||$ denotes L_1 norm.

The final loss ensemble L_{total} is therefore computed as in Equation 21 with 50, 100, and 20 as the weights given for the SSIM loss, gradient loss, and intensity loss, respectively.

$$\begin{aligned} L_{total} = 50 L_{SSIM} + 100 L_{Gradient} \\ + 20 L_{Intensity} \end{aligned} \tag{21}$$

To extend this approach for 11 source images, we use the SSIM Loss which is computed for each source image–fused image pair and averaged across all images to maintain global consistency. Similarly, the Gradient Loss is computed as the mean of the absolute difference between the fused image gradients and the maximum gradient magnitude among all 11 inputs. The Intensity Loss for 11 images is computed as the mean absolute difference between the fused image and the elementwise maximum pixel intensity across all 11 input images.

Experiment

Dataset

The following datasets were used to test and compare the performance of our method:

1. **PET-MRI:** This dataset contains 311 PET-MRI image pairs, split into 269 images for training and 42 images for testing with a resolution of 256×256 (Summers 2003).
2. **SPECT-MRI:** This dataset contains 430 SPECT-MRI image pairs, split into 357 images for training and 73 images for testing with a resolution of 256×256 (Summers 2003).
3. **CT-MRI:** This dataset contains 205 CT-MRI image pairs, split into 184 images for training and 21 images for testing with a resolution of 256×256 (Summers 2003).
4. **Oocytes:** This private dataset contains 2167 multi-focal (11 focal) images, split into 1606 images for training and 561 images for testing with a resolution of 800×800.

Implementation Details

We use the Adam optimizer for training our model. We employ an initial learning rate of 10^{-4}, and decay the learning rate using the MultiStepLR scheduler by reducing the learning rate by a factor of 0.5 every 50 epochs. A batch size of 4 is used, and our training is restricted to 200 epochs on an Nvidia A100 GPU. The images are randomly cropped into patches of size 64×64 for training.

Evaluation Metrics

The performance of our model is evaluated using various metrics, each reflecting a different aspect of fusion quality. Entropy (EN) is a measure of the amount of unique information present in the image. Standard Deviation (SD) measures the color variation, Spatial Frequency (SF) quantifies edge information, and Average Gradient (AG) assesses edge sharpness. Additionally, Mutual Information (MI) indicates how much information is retained from the original images, the Structural Content Difference (SCD) assesses information preservation, and the Correlation Coefficient (CC) analyzes the linear correlation between source and fused images. Higher values are preferred for all metrics. We note that EN, SD, SF, and AG are no-reference metrics, which means that they do not require a ground-truth, whereas MI, SCD, and CC are reference-based, comparing the fused image with the source images.

For evaluating the computational complexity of our model, we use the number of parameters in the model (Param) in millions and floating point operations per second in the model (FLOPS) in GigaFLOPS. The FLOPS is calculated on images with a resolution of 256×256. A lower value is preferred for Param and FLOPS.

Comparison Approaches

We compare our model TAFIE with other state-of-the-art image fusion methods, including AITFuse (Wang et al. 2024), ATFuse (Jian et al. 2024), CDDFuse (Zhao et al. 2023), CMTFusion (Park, Vien, and Lee 2023), DATFuse (Tang et al. 2023b), FusionMamba (Xie et al. 2024) ITFuse (Tang, He, and Liu 2024), MBHFuse (Sun et al. 2025), PMKFuse (Sun, Dong, and Zhu 2025), PSLPT (Wang, Deng, and Vivone 2024), and U2Fusion (Xu et al. 2020). The performance of these methods is reproduced using open-sourced implementations provided by the authors, following similar experimental settings described in their respective papers. We train the model proposed in PSLPT (Wang, Deng, and Vivone 2024) using an unsupervised manner to maintain consistency with other approaches and ours.

Quantitative Performance Table 1 presents a comprehensive quantitative comparison of the proposed method against several state-of-the-art fusion models across four benchmark datasets: PET-MRI, SPECT-MRI, CT-MRI, and Oocytes. As shown in Subtable A, the proposed method achieves competitive performance on the PET-MRI dataset. On the SPECT-MRI dataset (Subtable B), our model attains the highest SD value and exhibits robust information fusion capability. For the CT-MRI dataset (Subtable C), our approach demonstrates superior performance in EN, SF, AG, and SCD, indicating improved contrast, texture richness, and edge enhancement compared to competing methods. Lastly, on the Oocytes dataset (Subtable D), the proposed model delivers the best EN and SF scores, confirming its effectiveness in capturing fine-grained biological textures while retaining strong global consistency.

A: PET-MRI Dataset

Method	EN ↑	SD ↑	SF ↑	AG ↑	MI ↑	SCD ↑	CC ↑	Param ↓	FLOPS ↓
AIT (Wang et al. 2024)	**6.41**	55.78	8.41	**3.48**	2.20	1.24	0.62	6.50	82.09
ATF (Jian et al. 2024)	4.01	71.41	8.22	2.86	2.37	1.22	0.61	1.05	5.40
CDD (Zhao et al. 2023)	4.07	61.01	8.10	2.77	2.34	**1.29**	**0.65**	1.19	77.68
CMT (Park, Vien, and Lee 2023)	4.04	47.64	6.73	2.40	1.94	1.19	0.59	0.62	13.10
DAT (Tang et al. 2023b)	4.18	**87.38**	7.45	2.53	2.07	1.19	0.59	**0.01**	**2.32**
FMB (Xie et al. 2024)	4.30	67.80	**8.73**	3.06	1.94	1.20	0.60	225.42	26.48
ITF (Tang, He, and Liu 2024)	4.34	38.53	5.82	1.97	1.69	1.23	0.61	0.08	5.68
MBH (Sun et al. 2025)	3.93	60.26	8.11	2.78	**2.53**	1.29	0.64	0.30	28.92
PMK (Sun, Dong, and Zhu 2025)	3.91	61.42	8.10	2.75	2.47	1.28	0.64	0.05	3.29
PSL (Wang, Deng, and Vivone 2024)	4.06	66.94	8.07	2.79	2.39	1.26	0.63	1.26	24.56
U2F (Xu et al. 2020)	4.49	47.00	5.28	2.02	1.82	1.20	0.60	0.66	43.17
Ours	4.30	64.82	8.10	2.82	2.46	1.26	0.63	1.09	108.14

B: SPECT-MRI Dataset

Method	EN ↑	SD ↑	SF ↑	AG ↑	MI ↑	SCD ↑	CC ↑	Param ↓	FLOPS ↓
AIT (Wang et al. 2024)	4.09	58.38	7.02	2.30	3.07	1.70	0.85	6.50	82.09
ATF (Jian et al. 2024)	3.77	57.57	7.18	2.35	2.74	1.71	0.85	1.05	5.40
CDD (Zhao et al. 2023)	3.79	58.27	7.08	2.28	**3.13**	1.69	0.85	1.19	77.68
CMT (Park, Vien, and Lee 2023)	3.75	43.60	5.13	1.74	2.36	**1.74**	**0.87**	0.62	13.10
DAT (Tang et al. 2023b)	**5.09**	49.80	7.11	2.57	2.33	1.71	0.86	**0.01**	**2.32**
FMB (Xie et al. 2024)	3.87	57.21	**8.05**	**2.63**	2.32	1.67	0.84	225.42	26.48
ITF (Tang, He, and Liu 2024)	4.72	48.75	5.51	1.91	2.20	1.70	0.85	0.08	5.68
MBH (Sun et al. 2025)	3.87	55.79	7.26	2.41	2.55	1.10	0.55	0.30	28.92
PMK (Sun, Dong, and Zhu 2025)	3.75	40.20	6.91	2.26	2.33	1.31	0.66	0.05	3.29
PSL (Wang, Deng, and Vivone 2024)	3.77	57.95	6.96	2.24	2.91	1.71	0.86	1.26	24.56
U2F (Xu et al. 2020)	3.90	45.30	4.01	1.37	2.22	1.74	0.87	0.66	43.17
Ours	4.00	**59.28**	7.00	2.27	2.54	1.73	0.86	1.09	108.14

C: CT-MRI Dataset

Method	EN ↑	SD ↑	SF ↑	AG ↑	MI ↑	SCD ↑	CC ↑	Param ↓	FLOPS ↓
AIT (Wang et al. 2024)	4.30	**87.38**	7.41	2.59	2.26	1.62	0.81	6.50	82.09
ATF (Jian et al. 2024)	4.36	81.66	7.84	2.85	2.32	1.60	0.80	1.05	5.40
CDD (Zhao et al. 2023)	4.68	79.27	8.10	3.03	2.32	1.61	0.80	1.19	77.68
CMT (Park, Vien, and Lee 2023)	4.60	57.14	6.88	2.47	2.35	**1.68**	**0.84**	0.62	13.10
DAT (Tang et al. 2023b)	4.13	83.00	7.36	2.53	2.11	1.58	0.79	**0.01**	**2.32**
FMB (Xie et al. 2024)	4.32	81.87	**8.92**	**3.24**	1.88	1.61	0.80	225.42	26.48
ITF (Tang, He, and Liu 2024)	**5.24**	25.86	6.85	2.63	1.60	1.33	0.66	0.08	5.68
MBH (Sun et al. 2025)	4.48	79.81	7.98	2.95	**2.41**	1.59	0.79	0.30	28.92
PMK (Sun, Dong, and Zhu 2025)	4.54	78.53	7.88	2.92	2.30	1.61	0.80	0.05	3.29
PSL (Wang, Deng, and Vivone 2024)	4.69	81.36	7.81	2.89	2.33	1.61	0.81	1.26	24.56
U2F (Xu et al. 2020)	4.89	44.85	4.50	1.67	1.99	1.63	0.81	0.66	43.17
Ours	5.03	80.18	8.12	3.10	2.19	1.63	0.81	1.09	108.14

D: Oocytes Dataset

Method	EN ↑	SD ↑	SF ↑	AG ↑	MI ↑	SCD ↑	CC ↑	Param ↓	FLOPS ↓
EMMA (Zhao et al. 2024)	7.21	64.61	3.02	1.28	15.42	**10.23**	0.93	1.34	28.42
PMK (Sun, Dong, and Zhu 2025)	7.27	45.42	3.51	**1.47**	12.64	9.12	0.83	**0.31**	**22.09**
PSL (Wang, Deng, and Vivone 2024)	7.22	**65.70**	3.12	1.31	15.27	10.20	0.93	4.79	84.25
U2F (Xu et al. 2020)	7.19	64.92	2.92	1.24	**15.47**	10.23	0.93	0.66	43.40
Ours	**7.29**	64.66	**3.58**	1.38	15.38	10.22	**0.93**	17.30	2284.81

Table 1: A: Quantitative comparison on the PET-MRI dataset. B: Quantitative comparison on the SPECT-MRI dataset. C: Quantitative comparison on the CT-MRI dataset. D: Quantitative comparison on the Oocytes dataset. The best values are highlighted in bold, while the second-best values are highlighted in red.

Our approach, while effective, requires greater computational resources compared to some alternative techniques.

Qualitative Performance The qualitative performance comparison of our approach with other state-of-the-art methods is presented in Figure 3, Figure 4, Figure 5, and Figure 6. We note that the fused images obtained using DATFuse and FusionMamba do not maintain the salient information of the MRI image accurately, while the fused images obtained using ITFuse and U2Fusion do not preserve the detailed information of the PET/SPECT image accurately. ATFuse and PSLPT overcapture the functional information from the PET/SPECT images and do not effectively capture the MRI's anatomical information, while MBHFuse and PMKFuse overcapture the MRI's anatomical information but have limited effectiveness in capturing PET's/SPECT's metabolic information. Our method not only retains the richer information from the PET/SPECT image but also keeps the salient information from the MRI/CT image Figure 3 and Figure 4.

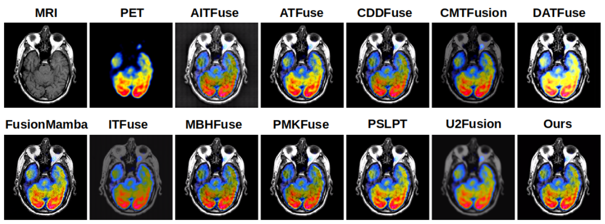

Figure 3: Qualitative comparison on the PET-MRI dataset with other state-of-the-art methods.

.

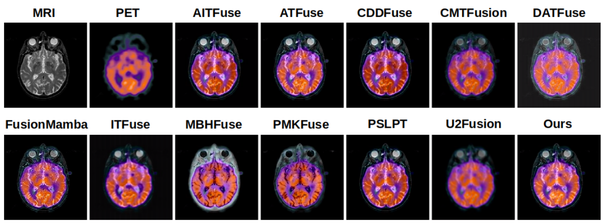

Figure 4: Qualitative comparison on the SPECT-MRI dataset with other state-of-the-art methods.

.

As shown in Figure 5, our method effectively preserves both structural and textural details, achieving enhanced edge sharpness and superior contrast in the fused CT-MRI images. The results exhibit clearer anatomical boundaries and more consistent intensity distributions, highlighting the model's capability to integrate complementary modality information. In Figure 6, which presents results on the Oocytes dataset across multiple focal lengths (F0, F15, F30, F45, F60, F75, F-15, F-30, F-45, F-60, and F-75), the proposed model consistently delivers sharper focus transitions and superior depth consistency compared to other methods.

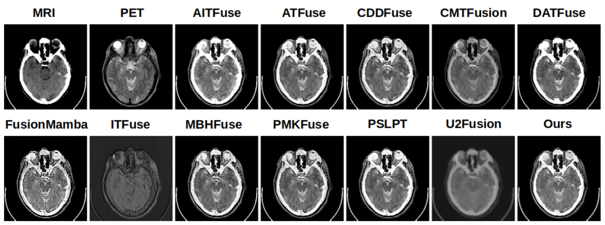

Figure 5: Qualitative comparison on the CT-MRI dataset with other state-of-the-art methods.

.

Application of Downstream Tasks We compare the fused image using our approach against the 11 individual source images as well as other approaches for the four-class oocyte quality prediction task, using ResNet50 as the backbone model. The reported values are under 5-fold cross-validation, and we demonstrate an improvement using our method using multiple metrics, as in Table 2.

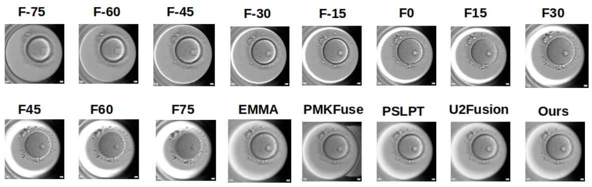

Figure 6: Qualitative comparison on the Oocytes dataset with other state-of-the-art methods. Here, F0, F15, F30, F45, F60, F75, F-15, F-30, F-45, F-60, and F-75 denote 11 different focal lengths.

.

Method	Acc ↑	Pre ↑	Rec ↑	F1 ↑	AUC ↑	MCC ↑
F0	0.3222	0.2975	0.2987	0.2838	0.5447	0.0810
F15	0.3270	0.3175	0.3037	**0.2974**	0.5574	0.0816
F30	0.3058	0.2889	0.2916	0.2782	0.5530	0.0613
F45	0.2862	0.2725	0.2695	0.2643	0.5418	0.0314
F60	0.3442	0.3050	0.3088	0.2920	0.5676	0.0922
F75	0.2993	0.2723	0.2754	0.2648	0.5423	0.0406
F-15	0.3295	0.3045	0.3070	0.2955	0.5692	0.0893
F-30	0.3140	0.3076	0.2959	0.2903	0.5601	0.0725
F-45	0.3090	**0.3225**	0.2848	0.2747	0.5543	0.0567
F-60	0.3214	0.3042	0.3028	0.2967	0.5462	0.0744
F-75	0.3263	0.3067	0.3129	0.2895	0.5767	0.0927
EMMA (Zhao et al. 2024)	0.3524	0.2958	0.3147	0.2519	0.5726	**0.1283**
PMK (Zhao et al. 2024)	0.3269	0.2711	0.2978	0.2635	0.5482	0.1187
PSL (Zhao et al. 2024)	0.3415	0.2864	0.3083	0.2748	0.5751	0.1129
U2F (Zhao et al. 2024)	0.3478	0.2789	0.3125	0.2796	0.5733	0.1098
Ours	**0.3541**	0.2795	**0.3154**	0.2812	**0.5810**	0.1069

Table 2: Quantitative results on the Oocytes dataset for the downstream task of four-class quality prediction (image classification) using the ResNet50 model. Here, F0, F15, F30, F45, F60, F75, F-15, F-30, F-45, F-60, and F-75 denote 11 different focal lengths. Acc denotes Accuracy, Pre denotes Precision, Rec denotes Recall, F1 denotes F1 Score, AUC denotes class-wise average AUC Score, and MCC denotes Matthews Correlation Coefficient. The best values are highlighted in bold.

Ablation Study We perform an ablation study using the entropy-based attention as in Table 3. Using the entropy-based attention module in the encoder, decoder, and fusion block results in the highest EN, AG, MI, SCD, and CC scores.

The quantitative ablation study by removing various components from our network architecture is exhibited in Table 4. We note that using only the global branch leads to the best performance using SCD and CC as the metric. Not using the residual block gives the best result using SF as the metric. The best result on the SD metric was obtained by not using a transformer in the fusion block. Our model performs the best using EN, AG, and MI metrics.

Table 5 reports a quantitative ablation study on the PET-MRI dataset, evaluating key Vision Transformer hyperparameters. Increasing TopK from 4 to 8 improves most metrics with moderate computational cost, while further increases yield marginal gains. Using 8 attention heads achieves the best balance between performance and efficiency, and an embedding dimension of 96 provides the strongest overall results, whereas smaller or larger sizes ei-

A: Entropy Based Attention Module										
E/D	Fus	EN ↑	SD ↑	SF ↑	AG ↑	MI ↑	SCD ↑	CC ↑	Param ↓	FLOPS ↓
×	×	3.96	65.32	8.09	2.79	2.40	1.26	0.63	**1.00**	**96.89**
×	✓	4.01	**67.39**	**8.12**	2.81	2.41	1.24	0.62	1.07	106.63
✓	×	4.00	66.41	8.10	2.81	2.39	1.26	0.63	1.02	98.40
✓	✓	**4.30**	64.82	8.10	**2.82**	**2.46**	**1.26**	**0.63**	1.09	108.14

Table 3: A: Quantitative ablation study on the PET-MRI dataset using an entropy-based attention module. Enc/Dec stands for the encoder-decoder block, and Fusion stands for the fusion block. The best values are highlighted in bold.

ther reduce performance or increase FLOPS. These findings highlight the importance of carefully tuning TopK, attention heads, and embedding dimensions for optimal Transformer performance in PET-MRI fusion.

Finally, a loss function-based ablation study is presented in Table 6, illustrating the individual contributions of the SSIM, Intensity, and Gradient terms to the overall fusion performance across multiple quantitative metrics.

Network	EN ↑	SD ↑	SF ↑	AG ↑	MI ↑	SCD ↑	CC ↑	Param ↓	FLOPS ↓
W/O G	3.97	64.60	8.09	2.80	2.42	1.27	0.63	0.88	57.55
W/O L	3.94	61.56	7.91	2.70	2.32	**1.28**	**0.64**	**0.77**	**50.60**
W/O RB	3.98	66.20	**8.13**	2.81	2.39	1.26	0.63	0.89	94.77
W/O EE	3.98	67.23	8.07	2.78	2.44	1.25	0.62	1.09	108.03
W/O DC	3.98	63.66	8.08	2.78	2.34	1.26	0.63	1.07	106.83
W/O TF	4.10	**67.60**	8.11	2.81	2.40	1.25	0.62	0.95	88.79
Ours	**4.30**	64.82	8.10	**2.82**	**2.46**	1.26	0.63	1.09	108.14

Table 4: Ablation study using different components in the network architecture. ”W/O G” denotes without the global branch; ”W/O L” denotes without the local branch; ”W/O RB” means without the residual block; ”W/O EE” means without the edge encoders; ”W/O DC” means using a regular convolutional layer in place of dynamic convolution; and ”W/O TF” means without the transformer in the fusion block. The best values are highlighted in bold.

A: K in TopK of Transformer									
K	EN ↑	SD ↑	SF ↑	AG ↑	MI ↑	SCD ↑	CC ↑	Param ↓	FLOPS ↓
4	4.12	63.45	7.84	2.73	2.39	1.21	0.61	**1.03**	**101.28**
8	**4.30**	64.82	**8.10**	**2.82**	2.46	**1.26**	**0.63**	1.09	108.14
16	4.18	**64.90**	7.91	2.75	**2.48**	1.23	0.62	1.15	116.32
B: No. of Heads in Transformer									
H	EN ↑	SD ↑	SF ↑	AG ↑	MI ↑	SCD ↑	CC ↑	Param ↓	FLOPS ↓
4	4.11	63.32	7.80	2.71	2.37	1.20	**0.65**	**1.04**	**102.47**
8	4.30	**64.82**	8.10	2.82	**2.46**	**1.26**	0.63	1.09	108.14
16	**4.34**	64.05	7.92	**2.87**	2.42	1.23	0.62	1.14	115.90
C: Transformer Embedding Dimension									
D	EN ↑	SD ↑	SF ↑	AG ↑	MI ↑	SCD ↑	CC ↑	Param ↓	FLOPS ↓
48	4.05	62.90	7.65	2.68	2.34	1.18	**0.64**	**0.98**	**96.20**
96	**4.30**	64.82	**8.10**	2.82	**2.46**	**1.26**	0.63	1.09	108.14
192	4.25	**65.13**	7.95	**2.88**	2.43	1.24	0.62	1.20	123.47

Table 5: Quantitative ablation study on the PET-MRI dataset. A: K in TopK of the Vision Transformer. B: Number of attention heads in the Vision Transformer. C: Embedding dimension size in the Vision Transformer. The best-performing values are highlighted in bold.

SSIM	Gradient	Intensity	EN ↑	SD ↑	SF ↑	AG ↑	MI ↑	SCD ↑	CC ↑
10	50	100	4.04	63.25	7.81	2.63	2.34	1.18	0.58
30	50	100	4.11	64.03	8.04	2.74	2.40	1.22	0.61
20	30	100	4.04	63.54	7.87	2.70	2.36	1.21	0.60
20	70	100	4.17	64.52	8.01	**2.84**	2.46	1.23	**0.66**
20	50	50	3.96	63.23	7.72	2.61	2.29	1.18	0.57
20	50	200	4.22	64.74	**8.15**	2.83	**2.48**	1.25	0.63
20	50	100	**4.30**	**64.82**	8.10	2.82	2.46	**1.26**	0.63

Table 6: Quantitative ablation study on the PET-MRI dataset, evaluating the impact of different loss function components: SSIM, Intensity, and Gradient. The best-performing values are highlighted in bold.

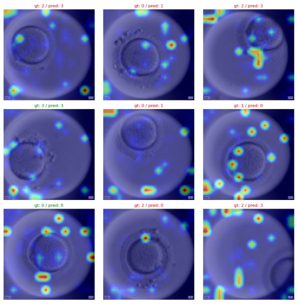

Figure 7: Attention heatmaps produced by the ViT on fused oocyte images for the quality prediction task. Each image is outlined according to prediction correctness: **green** indicates correct predictions, while **red** indicates incorrect ones. Within the heatmaps, warmer colors represent stronger attention responses, with **red** denoting the highest attention.

Failure Cases We replaced the ResNet50 backbone with a Vision Transformer (ViT) and visualized the corresponding attention heatmaps in Figure 7. Several failure cases were observed, where the model not only produced incorrect quality predictions but also focused attention on irrelevant regions of the image. These results suggest that the fused images alone may be insufficient for reliable quality prediction.

Conclusions

We propose a dual-branch image fusion network that enhances feature extraction and fusion. It combines an entropy-based attention module with a Top-K token vision transformer to highlight salient regions and preserve global context efficiently. Edge encoders and dynamic convolutions in the local branch further strengthen feature representation. Applied to oocyte quality prediction, our fused representations outperform individual focal images and existing methods. Experiments on multiple medical imaging datasets show competitive performance, and ablation studies confirm the contribution of each component.

Acknowledgements

We would like to thank Godwin Komla Krampah, Tan Lu, and Ann Dooms for their insightful discussions and valuable feedback. This study was funded by the FWO Project Grant “Joint R&D 2024: Freezing of Reliable Oocytes after Screening with AI for women awaiting Cancer Treatment”.

References

Cai, J.; Gu, S.; and Zhang, L. 2018. Learning a deep single image contrast enhancer from multi-exposure images. *IEEE Transactions on Image Processing*, 27(4): 2049–2062.

Chen, J.; Ding, J.; Yu, Y.; and Gong, W. 2023. THFuse: An infrared and visible image fusion network using transformer and hybrid feature extractor. *Neurocomputing*, 527: 71–82.

Chollet, F. 2017. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1251–1258.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ICLR*.

Fang, L.; Hou, M.; Huang, B.; Chen, G.; and Yang, J. 2025. DCAFusion: A novel general image fusion framework based on reference image reconstruction and dual-cross attention mechanism. *Information Sciences*, 698: 121772.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141.

Jian, L.; Xiong, S.; Yan, H.; Niu, X.; Wu, S.; and Zhang, D. 2024. Rethinking cross-attention for infrared and visible image fusion. *arXiv preprint arXiv:2401.11675*.

Li, H.; and Wu, X.-J. 2018. DenseFuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing*, 28(5): 2614–2623.

Li, H.; and Wu, X.-J. 2024. CrossFuse: A novel cross attention mechanism based infrared and visible image fusion approach. *Information Fusion*, 103: 102147.

Li, H.; Wu, X.-J.; and Durrani, T. 2020. NestFuse: An infrared and visible image fusion architecture based on nest connection and spatial/channel attention models. *IEEE Transactions on Instrumentation and Measurement*, 69(12): 9645–9656.

Li, J.; Song, H.; Liu, L.; Li, Y.; Xia, J.; Huang, Y.; Fan, J.; Lin, Y.; and Yang, J. 2025. MixFuse: An iterative mix-attention transformer for multi-modal image fusion. *Expert Systems with Applications*, 261: 125427.

Li, X.; Liu, W.; Li, X.; Zhou, F.; Li, H.; and Nie, F. 2024. All-weather multi-modality image fusion: Unified framework and 100k benchmark. *arXiv preprint arXiv:2402.02090*.

Liu, J.; Fan, X.; Huang, Z.; Wu, G.; Liu, R.; Zhong, W.; and Luo, Z. 2022. 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*, 5802–5811.

Liu, J.; Liu, Z.; Wu, G.; Ma, L.; Liu, R.; Zhong, W.; Luo, Z.; and Fan, X. 2023. Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8115–8124.

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

Ma, J.; Tang, L.; Fan, F.; Huang, J.; Mei, X.; and Ma, Y. 2022. SwinFusion: Cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA Journal of Automatica Sinica*, 9(7): 1200–1217.

Mei, L.; Hu, X.; Ye, Z.; Tang, L.; Wang, Y.; Li, D.; Liu, Y.; Hao, X.; Lei, C.; Xu, C.; et al. 2024. GTM-Fuse: Group-attention transformer-driven multiscale dense feature-enhanced network for infrared and visible image fusion. *Knowledge-Based Systems*, 293: 111658.

Nejati, M.; Samavi, S.; and Shirani, S. 2015. Multi-focus image fusion using dictionary-based sparse representation. *Information fusion*, 25: 72–84.

Park, S.; Vien, A. G.; and Lee, C. 2023. Cross-modal transformers for infrared and visible image fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(2): 770–785.

Qu, L.; Liu, S.; Wang, M.; Li, S.; Yin, S.; and Song, Z. 2024. Trans2Fuse: Empowering image fusion through self-supervised learning and multi-modal transformations via transformer networks. *Expert Systems with Applications*, 236: 121363.

Qu, L.; Liu, S.; Wang, M.; and Song, Z. 2022. Transmef: A transformer-based multi-exposure image fusion framework using self-supervised multi-task learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 2126–2134.

Ram Prabhakar, K.; Sai Srikar, V.; and Venkatesh Babu, R. 2017. Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs. In *Proceedings of the IEEE international conference on computer vision*, 4714–4722.

Sagar, A. 2025. LightFusionNet: Lightweight Dual-Stream Network with Predictive Context Attention for Efficient Medical Image Fusion. *medRxiv*, 2025–10.

Summers, D. 2003. Harvard Whole Brain Atlas: www. med. harvard. edu/AANLIB/home. html. *Journal of Neurology, Neurosurgery & Psychiatry*, 74(3): 288–288.

Sun, Y.; Dong, M.; Yu, M.; and Zhu, L. 2025. MBHFuse: A multi-branch heterogeneous global and local infrared and visible image fusion with differential convolutional amplification features. *Optics & Laser Technology*, 181: 111666.

Sun, Y.; Dong, M.; and Zhu, L. 2025. Rethinking the approach to lightweight multi-branch heterogeneous image fusion frameworks: Infrared and visible image fusion via the parallel Mamba-KAN framework. *Optics & Laser Technology*, 185: 112612.

Tang, L.; Deng, Y.; Ma, Y.; Huang, J.; and Ma, J. 2022a. SuperFusion: A versatile image registration and fusion network with semantic awareness. *IEEE/CAA Journal of Automatica Sinica*, 9(12): 2121–2137.

Tang, L.; Yuan, J.; and Ma, J. 2022. Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network. *Information Fusion*, 82: 28–42.

Tang, L.; Yuan, J.; Zhang, H.; Jiang, X.; and Ma, J. 2022b. PIAFusion: A progressive infrared and visible image fusion network based on illumination aware. *Information Fusion*, 83: 79–92.

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

Tang, W.; and He, F. 2024. FATFusion: A functional–anatomical transformer for medical image fusion. *Information Processing & Management*, 61(4): 103687.

Tang, W.; He, F.; and Liu, Y. 2022. YDTR: Infrared and visible image fusion via Y-shape dynamic transformer. *IEEE Transactions on Multimedia*, 25: 5413–5428.

Tang, W.; He, F.; and Liu, Y. 2023. TCCFusion: An infrared and visible image fusion method based on transformer and cross correlation. *Pattern Recognition*, 137: 109295.

Tang, W.; He, F.; and Liu, Y. 2024. ITFuse: An interactive transformer for infrared and visible image fusion. *Pattern Recognition*, 156: 110822.

Tang, W.; He, F.; Liu, Y.; and Duan, Y. 2022c. MATR: Multimodal medical image fusion via multiscale adaptive transformer. *IEEE Transactions on Image Processing*, 31: 5134–5149.

Tang, W.; He, F.; Liu, Y.; Duan, Y.; and Si, T. 2023b. DATFuse: Infrared and visible image fusion via dual attention transformer. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(7): 3159–3172.

Tang, Z.; Xiao, G.; Guo, J.; Wang, S.; and Ma, J. 2023c. Dual-attention-based feature aggregation network for infrared and visible image fusion. *IEEE Transactions on Instrumentation and Measurement*, 72: 1–13.

Vaswani, A. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; and Hu, Q. 2020. ECA-Net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11534–11542.

Wang, W.; Deng, L.-J.; and Vivone, G. 2024. A general image fusion framework using multi-task semi-supervised learning. *Information Fusion*, 108: 102414.

Wang, X.; Girshick, R.; Gupta, A.; and He, K. 2018. Nonlocal neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7794–7803.

Wang, X.; Hua, Z.; and Li, J. 2023. Cross-UNet: dual-branch infrared and visible image fusion framework based on cross-convolution and attention mechanism. *The Visual Computer*, 39(10): 4801–4818.

Wang, Z.; Yang, F.; Sun, J.; Xu, J.; Yang, F.; and Yan, X. 2024. AITFuse: Infrared and visible image fusion via adaptive interactive transformer learning. *Knowledge-Based Systems*, 299: 111949.

Woo, S.; Park, J.; Lee, J.-Y.; and Kweon, I. S. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, 3–19.

Xie, X.; Cui, Y.; Tan, T.; Zheng, X.; and Yu, Z. 2024. Fusionmamba: Dynamic feature enhancement for multimodal image fusion with mamba. *Visual Intelligence*, 2(1): 37.

Xu, H.; Ma, J.; Jiang, J.; Guo, X.; and Ling, H. 2020. U2Fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1): 502–518.

Zhang, H.; Le, Z.; Shao, Z.; Xu, H.; and Ma, J. 2021. MFF-GAN: An unsupervised generative adversarial network with adaptive and gradient joint constraints for multi-focus image fusion. *Information Fusion*, 66: 40–53.

Zhang, H.; and Ma, J. 2021. SDNet: A versatile squeeze-and-decomposition network for real-time image fusion. *International Journal of Computer Vision*, 129(10): 2761–2785.

Zhang, X. 2021. Benchmarking and comparing multi-exposure image fusion algorithms. *Information Fusion*, 74: 111–131.

Zhang, X.; and Demiris, Y. 2023. Visible and infrared image fusion using deep learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8): 10535–10554.

Zhang, Y.; Liu, Y.; Sun, P.; Yan, H.; Zhao, X.; and Zhang, L. 2020. IFCNN: A general image fusion framework based on convolutional neural network. *Information Fusion*, 54: 99–118.

Zhao, Z.; Bai, H.; Zhang, J.; Zhang, Y.; Xu, S.; Lin, Z.; Timofte, R.; and Van Gool, L. 2023. Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5906–5916.

Zhao, Z.; Bai, H.; Zhang, J.; Zhang, Y.; Zhang, K.; Xu, S.; Chen, D.; Timofte, R.; and Van Gool, L. 2024. Equivariant multi-modality image fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 25912–25921.