RIFNet: Bridging Modalities for Accurate and Detailed Ocular Disease Analysis

Yuqing Li[1,2,†], Qingshan Hou[1,2,†,(✉)], Peng Cao[1,2,(✉)], Jianguo Ju[4], Tianqi Wang[1,2], Meng Wang[5], Ke Zou[5], Yih Chung Tham[5], Huazhu Fu[6], Osmar R. Zaiane[3]

[1] Computer Science and Engineering, Northeastern University, Shenyang, China
[2] Key Laboratory of Intelligent Computing in Medical Image of Ministry of Education, Northeastern University, Shenyang, China
caopeng@cse.neu.edu.cn
houqingshancv@gmail.com
[3] Alberta Machine Intelligence Institute, University of Alberta, Edmonton, Canada
[4] College of Artificial Intelligence and Computer Science, Xi'an University of Science and Technology, Xi'an, China
[5] Ophthalmology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore
[6] Institute of High Performance Computing, Agency for Science, Technology and Research, Singapore

Abstract. Color fundus photography (CFP) is widely used in clinical practice for its convenience and accessibility. However, it faces challenges such as low image quality, limited depth information, susceptibility to artifacts and low contrast, which reduce diagnostic accuracy and hinder the detection of small lesions. Fluorescein angiography (FA), on the other hand, effectively highlights features such as vascular leakage and non-perfusion. However, it also has drawbacks, including health risks and the lack of color information. To address these challenges, we propose a multi-stage retinal image fusion framework, RIFNet, to improve image quality and diagnostic efficacy by integrating multimodal information from CFP and FA. First, to address the problem of missing modalities due to the difficulty of accessing FA as an intrusive inspection, we design a bi-stream generative subnetwork to generate pseudo FA images by pre-training with real CFP images as the generating condition, which effectively supplements the modality information. Subsequently, the color representations of different modalities are unified by color coding, and fed into the multimodal discriminative fusion network to generate the fused color-coded images. Finally, a multiscale reconstruction method is used to generate a high-resolution and high-contrast enhanced image. Experiments demonstrate that this multimodal fusion framework supplements FA information, reduces medical costs, and reveals lesion details unobservable with a single modality, supporting accurate ocular disease diagnosis.

† Yuqing Li and Qingshan Hou contribute equally to this work.

Keywords: Multimodal Fusion · Ocular Disease Diagnosis · Color Fundus Photography · Fluorescein Angiography.

1 Introduction

Ocular diseases are a serious threat to the visual function of patients, for example, diabetic retinopathy (DR) and age-related macular degeneration can cause irreversible vision loss if left untreated. Fundus images play a key role in the prevention, diagnosis and treatment of ocular diseases [8]. In clinical practice, various imaging modalities have been developed to capture detailed information about the retina. Color fundus photography (CFP) is a simple and safe imaging technique widely used for fundus examination. It provides high-resolution color images of retinal structures. However, CFP often suffers from low contrast and artifacts, which may obscure subtle pathological changes critical for accurate diagnosis [9]. Fluorescein angiography (FA) is the standard imaging technique for evaluating the retinal vasculature. By injecting sodium fluorescein, FA produces high-contrast images that show microaneurysms, neovascularization, and vascular leakage. Despite its diagnostic value, FA is invasive and carries risks such as nausea, vomiting, and in rare cases, serious complications such as cardiac arrest [2], which makes FA data difficult to obtain and scarce.

Multimodal medical image fusion (MMIF) combines medical images from different modalities to create composite images, enhancing diagnostic accuracy by integrating complementary information. For fundus imaging, MMIF allows structural information from CFP to be combined with functional information from FA, enabling more comprehensive lesion analysis. While MMIF has been extensively explored in CT (computerized tomography)-MRI (magnetic resonance imaging) [10,3], and PET (positron emission tomography)-MRI [13,5], SPECT (single-photon emission computed tomography)-MRI [15,19], several challenges and shortcomings arise due to the unique characteristics of CFP and FA imaging modalities: 1) **Limited FA data availability:** Publicly available FA images are scarce due to the invasive nature of the examination, while CFP images are abundant but lack the functional insights provided by FA, leading to a modal imbalance. 2) **Modal differences:** CFP and FA images exhibit significant disparities in resolution, contrast, and noise, complicating the fusion process. 3) **Preservation of fine details:** Fusion strategies must ensure structure and pathology details (e.g. retinal vessels, macula, hemorrhage) are retained and enhanced while balancing global and local information for effective clinical interpretation.

To address the above challenges, we propose a multimodal retinal image fusion framework for ocular diseases, called RIFNet. Specifically, first, to compensate for the missing FA modality data and obtain complementary information for CFP, a pre-trained bi-stream generative subnetwork is proposed to generate pseudo FA images. At the same time, we design to use viridis colormap to encode multi-modal images and unify image representation. Secondly, generative adversarial network (GAN) is designed as a multimodal discriminator structure to capture multimodal information globally and locally. Finally, the output of the fusion network serves as a weight matrix to achieve the reconstruction of high-quality color CFP images, thereby effectively enhancing the performance of downstream tasks and providing support for further clinical applications. The contributions of RIFNet can be summarized as: 1) We propose a multimodal image fusion method that integrates CFP and FA imaging modalities. 2) By employing the viridis colormap for color encoding, our approach enhances pixel-level details while significantly improving the efficiency and effectiveness of the fusion process. 3) A pre-trained bi-stream generative subnetwork is introduced to synthesize pseudo-FA images, supplementing the functional information for CFP images. 4) Our strategy demonstrates significant performance improvements across multiple evaluation metrics, confirming its ability to precisely integrate information from different modalities and enhance the model's capacity to identify and analyze fundus diseases. The code is available at `https://github.com/Liyuyu666/RIFNet`.

2 Methodology

As shown in Figure 1, we propose RIFNet, a multimodal fusion framework that integrates complementary multimodal information and generates high-quality fused images. This fusion enhances both lesion visibility and structural details, improving ocular disease diagnosis.

2.1 Pre-trained Bi-stream Generative Subnetwork

In real clinical scenarios, obtaining paired multi-modal fundus images is often challenging. To tackle the issue of missing modalities during the multi-modal fusion process, we introduced a pre-trained bi-stream generative subnetwork [21]. It comprises two key branches: the high-resolution branch focuses on enhancing local details in the modality generation, while the low-resolution branch is responsible for regulating global information in modality generation. This pre-training process primarily involves two key aspects:

1) Multi-modal fundus image registration provides essential paired data for the supervised training of the generative network. We extract vascular structures [12] from both CFP and FA images to serve as matching feature points. Using these vascular maps, we perform keypoint detection [1] to facilitate feature matching [17] between the multi-modal fundus images. Finally, the RANSAC algorithm [4] is applied to estimate the homography matrix, followed by a transformation process, resulting in accurately aligned CFP and FA image pairs.

2) Color fundus images are transformed to generate complementary modalities. Based on registered multi-modal fundus images, we train a bi-stream generative subnetwork through supervised learning to achieve pixel-level modality conversion. This process provides high-quality paired (e.g., CFP & pseudo-FA) training data for the subsequent multi-modal fusion subnetwork.

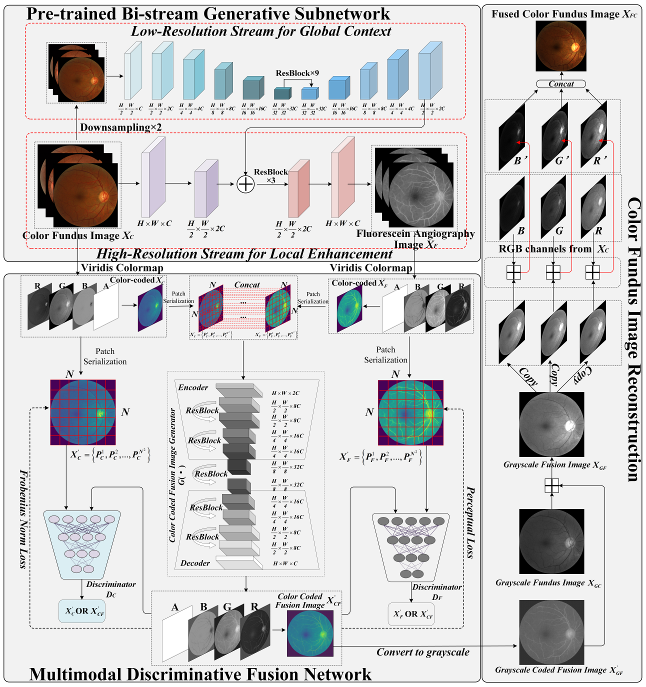

Pre-trained Bi-stream Generative Subnetwork
Low-Resolution Stream for Global Context
ResBlock×9
$\frac{H}{2}\times\frac{W}{2}\times C$
$\frac{H}{2}\times\frac{W}{2}\times 2C$
$\frac{H}{4}\times\frac{W}{4}\times 4C$
$\frac{H}{8}\times\frac{W}{8}\times 8C$
$\frac{H}{16}\times\frac{W}{16}\times 16C$
$\frac{H}{32}\times\frac{W}{32}\times 32C$
$\frac{H}{32}\times\frac{W}{32}\times 32C$
$\frac{H}{16}\times\frac{W}{16}\times 16C$
$\frac{H}{8}\times\frac{W}{8}\times 8C$
$\frac{H}{4}\times\frac{W}{4}\times 4C$
$\frac{H}{2}\times\frac{W}{2}\times 2C$
Downsampling×2
ResBlock ×3
$H\times W\times C$
$\frac{H}{2}\times\frac{W}{2}\times 2C$
$\frac{H}{2}\times\frac{W}{2}\times 2C$
$H\times W\times C$
Color Fundus Image X_C
Fluorescein Angiography Image X_F
High-Resolution Stream for Local Enhancement
Viridis Colormap
Viridis Colormap
R G B A
Color-coded X_C'
Patch Serialization
N
Concat
N
Patch Serialization
Color-coded X_F'
A B G R
$X_C' = \{P_C^1, P_C^2, ..., P_C^{N^2}\}$
$X_F' = \{P_F^1, P_F^2, ..., P_F^{N^2}\}$
Patch Serialization
Patch Serialization
N
N
N
N
Encoder
ResBlock
ResBlock
ResBlock
ResBlock
ResBlock
Decoder
Color Coded Fusion Image Generator G(•)
$H\times W\times 2C$
$\frac{H}{2}\times\frac{W}{2}\times 8C$
$\frac{H}{2}\times\frac{W}{2}\times 8C$
$\frac{H}{4}\times\frac{W}{4}\times 16C$
$\frac{H}{4}\times\frac{W}{4}\times 16C$
$\frac{H}{8}\times\frac{W}{8}\times 32C$
$\frac{H}{8}\times\frac{W}{8}\times 32C$
$\frac{H}{4}\times\frac{W}{4}\times 16C$
$\frac{H}{4}\times\frac{W}{4}\times 16C$
$\frac{H}{2}\times\frac{W}{2}\times 8C$
$\frac{H}{2}\times\frac{W}{2}\times 8C$
$H\times W\times C$
$X_C' = \{P_C^1, P_C^2, ..., P_C^{N^2}\}$
$X_F' = \{P_F^1, P_F^2, ..., P_F^{N^2}\}$
Frobenius Norm Loss
Perceptual Loss
Discriminator D_C
Discriminator D_F
X_C' OR X_{CF}'
X_F' OR X_{CF}'
A B G R
Color Coded Fusion Image X_{CF}'
Convert to grayscale
Multimodal Discriminative Fusion Network
Fused Color Fundus Image X_{FC}
Concat
B' G' R'
B G R
RGB channels from X_C
Copy
Copy
Copy
Grayscale Fusion Image X_{GF}
Grayscale Fundus Image X_{GC}
Grayscale Coded Fusion Image X_{GF}'
Color Fundus Image Reconstruction

Fig. 1: Illustration of the RIFNet framework. **Stage 1:** A bi-stream generative subnetwork processes CFP images X_C: one branch captures global features at low resolution, while the other enhances local details (vessels and lesions) at high resolution. **Stage 2:** Real X_C and synthetic X_F are color-coded via viridis colormap to X_C' and X_F', then fused through a discriminative fusion network to maintain image quality and structure. **Stage 3:** The fusion network output X_{CF}' serves as pixel weights to enhance regions of interest in X_C through both gray-scale and RGB-scale processing, producing the fused image X_{FC}.

2.2 Multimodal Discriminative Fusion Network

CFP and FA Images Encoding CFP and FA exhibit distinct differences in image representation. Specifically, CFP is an RGB image that depicts the structural and color information of the retina, while FA is a high-contrast grayscale image rich in vascular details. These differences provide complementary information but also pose significant challenges to the design of the fusion network. To effectively integrate the complementary information from CFP and FA images, we encode them into a unified color space. To unify feature representation and emphasize the lesion area, the viridis colormap scheme is employed, wherein a mapping function C is defined to linearly transform unstructured data into a uniform structured representation. The viridis scheme, renowned for its superior luminance and tonal contrast, effectively accentuates subtle image variations and reveals potential lesions. The encoding process is outlined as follows:

$X'_C = C_1(f_1(X_C)); X'_F = C_2(f_2(X_F))$, where C_1 and C_2 represent the mapping functions for CFP and FA, and f_1 and f_2 are the normalization functions.

Fusion Network Architecture The proposed fusion network architecture comprises a generator G and dual discriminators D_C and D_F. The generator consists of an encoder network and a decoder network. The color-coded images are segmented into corresponding patches $X'_C = \{P_C^1, P_C^2, ..., P_C^{N^2}\}$ and $X'_F = \{P_F^1, P_F^2, ..., P_F^{N^2}\}$, with their channels concatenated $Concat(X'_C, X'_F)$. The decoder is responsible for upsampling the encoded low-dimensional features and transforming them into a high-resolution image. Throughout this process, the information from X'_C and X'_F images is integrated into the generated image X'_{CF}. To obtain fused images with complementary information, we employ the Frobenius norm loss L_C^G to enforce global constraints between the generated image X'_{CF} and the CFP images X'_C, and applying perceptual loss L_F^G to enforce local constraints on the detailed features between the generated image X'_{CF} and the FA images X'_F. The losses L_C^G and L_F^G are defined as follows:

$$L_C^G = \|G(X'_C, X'_F) - X'_C\|_F^2 \quad ; \quad L_F^G = \|V(G(X'_C, X'_F)) - V(X'_F)\|_2 \tag{1}$$

where $\|\cdot\|_F^2$ indicate Frobenius norm. $V(\cdot)$ is the pre-trained VGG network used to extract the high level features of the image. $\|\cdot\|_2$ denotes the $L2$ norm, which measures the Euclidean distance between the feature representations of the generated image $G(X'_C, X'_F)$ and the FA image X'_F in the feature space extracted by the pre-trained VGG network. Additionally, the generator total loss also incorporates adversarial losses, denoted as (2), to enhance the perceptual quality of generated images and improve the generation robustness.

$$L_{G\_adv} = \mathbb{E}[log(1 - D_C(G(X'_C, X'_F))] + \mathbb{E}[log(1 - D_F(G(X'_C, X'_F)))] \tag{2}$$

Both discriminators (D_C and D_F) share an identical network architecture. The color-coded images X'_C/X'_F and the generated images X'_{CF} are input into the discriminators. The training process incorporates adversarial loss for discriminator optimization, complemented by label smoothing techniques, to reduce overfitting of the discriminators and enhance the training stability of the generator. The discriminator losses L_C^D and L_F^D are specified as follows:

$$L_C^D = \mathbb{E}[-log(D_C(X'_C))] + \mathbb{E}[-log(1 - D_C(G(X'_C, X'_F)))] \tag{3}$$

$$L_F^D = \mathbb{E}[-log(D_F(X'_F))] + \mathbb{E}[-log(1 - D_F(G(X'_C, X'_F)))] \tag{4}$$

Through iterative adversarial training between the generator and dual discriminators, salient features from both color-encoded images are systematically integrated to generate the final fused representation X'_{CF}.

2.3 Color Fundus Image Reconstruction

The purpose of fusing CFP and FA is to fully integrate multimodal information and maximize its utilization. CFP is more general in tasks such as disease diagnosis, vessel segmentation, and lesion segmentation, whereas FA has a relatively limited range of applications. Therefore, the reconstruction of CFP image is designed as the final goal of fusion, and the complementary information provided by FA image is used to enhance CFP, thereby improving image quality and diagnostic performance. First, the fused image X'_{CF} is converted to a grayscale image X'_{GF} and used as a weight matrix to enhance the important structures in the original CFP image X_C at the pixel level. Specifically, the fusion process can be divided into two stages: 1) the reconstructed X_{GF} is obtained from X'_{GF} and the grayscale CFP image X_{GC}, which can be expressed as: $X_{GF} = X_{GC} \times (1 + \alpha \times \frac{X'_{GF}}{max(X'_{GF})})$, where α is a scalar parameter to control the degree of enhancement, $max(\cdot)$ indicates to find the maximum of all pixel values in the X'_{GF}. 2) the three channels of X_C are represented as R, G and B, enhancement using X_{GF} for each channel to obtain new channel values R', G' and B', which are then recombined into a fused color fundus image X_{FC}.

3 Experiments

3.1 Datasets and Implementation Details

We evaluate RIFNet on two benchmark datasets: (1) The Isfahan MISP dataset [6], comprising 60 paired CFP and FA images (30 healthy and 30 DR cases); (2) The DRIVE dataset [18], including 40 CFP images with pixel-wise vascular annotations. RIFNet is developed in PyTorch and executes on a single NVIDIA A100 GPU for both training and testing. The preprocessing of the Isfahan MISP dataset involves three steps: cropping redundant black backgrounds from images, standardizing the fundus region, and performing coarse global registration. We reorganize the dataset into 54 CFP-FA image pairs for training the bi-stream and multimodal discriminative networks, reserving 6 pairs for testing. The bi-stream network receives 1024×1024 resized inputs across 600 epochs (batch size=4, learning rate=0.0002). Subsequently, training images are coded using viridis colormap, resized to 588×588, and divided into 21×21 patches, expanding the training set to 42,336 CFP-FA patch pairs and fed into the multimodal discriminative fusion network, the network is trained with a learning rate of 0.0001, a decay rate of 0.95, and a batch size of 48 for 10 epochs. Furthermore, several no-reference metrics (entropy (EN), standard deviation (SD) and full-reference metrics (mutual information (MI), structural similarity index (SSIM), visual information fidelity (VIF) [7], quality assessment of blended features ($Q^{AB/F}$) [23], learned perceptual image patch similarity(LPIPS) [26] and mutual gradient (MG) [16]) are employed for evaluation of fused results.

3.2 Comparisons With State-of-the-Art Methods

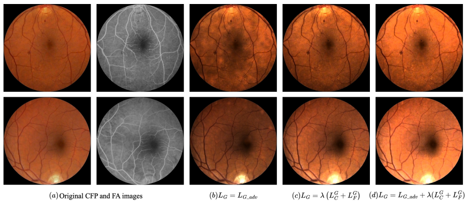
To validate the effectiveness of RIFNet, we conduct quantitative comparisons with other comparable MMIF methods, including DDcGAN [14], DenseFuse [11], MATR [20], PMGI [25], SDNet [24], CDDFuse [27], and EMFusion [22]. As quantitatively demonstrated in Table 1 (columns 2-7), RIFNet achieves state-of-the-art performance in critical evaluation dimensions. In terms of information preservation, RIFNet establishes new benchmarks with MI = 3.06 (0.36 higher than MART's 2.7) and SD = 78.82 (surpassing EMFusion by 3.42), demonstrating superior capability in retaining both statistical information and intensity variations from multi-model images (CFA & FA). In image quality assessment, RIFNet shows significant quality enhancement evidenced by VIF = 1.51 (53% improvement over MATR's 0.98). These quantitative gains visually translate to enhanced detail perception and improved structural fidelity. It should be noted that EN, $Q^{AB/F}$ and SSIM have relatively low values, which can be attributed to our fusion strategy that prioritizes diagnostic relevance and structural clarity rather than maximizing these metrics, this can be further elucidated by analyzing the fused images in comparison with individual source images. As shown in columns 8 to 13 of the Table 1, RIFNet significantly outperforms other methods in the metrics of MI, SSIM, VIF, $Q^{AB/F}$, LPIPS and MG, suggesting that the fused images outperform in terms of perceptual similarity, gradient information retention, and visual informativeness. RIFNet emphasizes clinically significant features in CFP while reducing redundant information in FA.

Table 1: The performance comparison between comparable methods and RIFNet is quantitatively evaluated through multiple metrics. Full-reference metrics are computed using two configurations: (1) multi-modal inputs (CFP and FA) for Columns 2-7, and (2) CFP-only inputs for Columns 8-13.

Methods	Multi-model						CFP					
	EN ↑	SD ↑	MI ↑	SSIM ↑	VIF ↑	$Q^{AB/F}$ ↑	MI ↑	SSIM ↑	VIF ↑	$Q^{AB/F}$ ↑	LPIPS ↓	MG ↑
SDNet [24]	6.73	74.97	2.61	1.18	0.75	0.49	1.57	0.47	0.26	0.22	0.31	0.72
PMGI [25]	**6.79**	61.75	2.68	1.21	0.81	0.61	1.52	0.53	0.15	0.25	0.3	0.57
DenseFuse [11]	5.8	54.21	2.56	**1.43**	0.56	0.25	1.76	0.82	0.29	0.33	0.21	0.87
MATR [20]	6.7	63.75	2.7	1.12	0.98	**0.67**	1.3	0.44	0.13	0.25	0.35	0.56
EMFusion [22]	6.56	75.4	2.66	1.5	0.89	0.63	1.43	0.64	0.19	0.26	0.32	0.67
CDDFuse [27]	6.63	69.44	2.4	1.42	0.93	0.53	1.26	0.62	0.23	0.26	0.31	0.58
DDcGAN [14]	6.62	62.34	2.16	1.28	0.48	0.38	1.5	0.58	0.16	0.18	0.32	0.44
RIFNet(Ours)	6.5	**78.82**	**3.06**	1.33	**1.51**	0.29	**2.43**	**0.86**	**1.44**	**0.45**	**0.15**	**0.91**

3.3 Ablation Study

To evaluate the effectiveness of multimodal fusion networks, this section explores two key components: adversarial loss L_{G_adv} and content loss. Specifically, the content loss consists of Frobenius norm loss L_C^G and perceptual loss L_F^G. We systematically analyze the effects of various loss combinations on the integration of cross-modal image information. Specifically, the ablation study includes: 1) Using adversarial loss only: $L_G = L_{G\_adv}$; 2) Using content loss only: $L_G = \lambda(L_C^G + L_F^G)$; and 3) Using both adversarial and content losses: $L_G = L_{G\_adv} + \lambda(L_C^G + L_F^G)$. As shown in Figure 2, we qualitatively demonstrate the impact of these different loss function combinations on image fusion results. When using only L_{G_adv} in Figure 2 (b), the fusion results exhibit significant degradation in image quality. The retinal vasculature appears blurry and poorly defined, with undesired artifacts that obscure important anatomical details. With content loss alone (L_C^G & L_F^G), as demonstrated in Figure 2 (c), the fusion results show indiscriminate enhancement across the entire image. While the overall visibility is improved compared to using only adversarial loss, RIFNet fails to selectively emphasize clinically relevant features. The vessels and retinal structures are enhanced with similar intensity as the background, resulting in suboptimal feature distinction. In contrast, the optimal combination of both adversarial and content losses in Figure 2 (d) achieves superior fusion quality. This loss combination enhances the vasculature and lesions from the original CFP while suppressing background noise. The background demonstrates improved uniformity without artifacts, resulting in images that are both visually appealing and clinically informative. Besides, the quantitative results (Table 2) show that when combining all loss functions (L_{G_adv}, L_C^G, and L_F^G), our method achieves the best performance across all comparable metrics, including EN=6.5, SD=78.82, VIF=1.51, and MG=1.28, with significant improvements compared to using single loss functions alone.



(a) Original CFP and FA images

(b)$L_G = L_{G\_adv}$

(c)$L_G = \lambda\left(L_C^G + L_F^G\right)$

(d)$L_G = L_{G\_adv} + \lambda(L_C^G + L_F^G)$

Fig. 2: Qualitative results of multimodal fusion network ablation.

Losses			EN ↑	SD ↑	VIF ↑	MG ↑
L_{G_adv}	L_C^G	L_F^G				
✓	×	×	6.23	53.5	1.01	1.21
×	✓	✓	6.28	62.2	1.15	1.16
✓	✓	✓	**6.5**	**78.82**	**1.51**	**1.28**

Table 2: Quantitative results of multimodal fusion network ablation.

3.4 Performance Evaluation of Vessel Segmentation Task

We evaluate RIFNet's effectiveness via a vascular segmentation task. Based on the DRIVE dataset, we conduct comparative experiments across the results of multiple fusion frameworks for both U-Net training and testing. As illustrated in Figure 3, the qualitative results demonstrate the superior performance of RIFNet. The RIFNet-generated results exhibit several advantages over existing methods: 1) The segmentation results reveal more precise delineation of vessels, particularly evident in the enlarged views (red and yellow boxes). 2) Our method preserves the continuity of vascular structures better than comparative methods, avoiding the fragmentation issues in MATR and DenseFuse. 3) The Grad-CAM visualizations indicate that our method focuses more accurately on the vascular network, with heatmap highlighting the clinically relevant vessel structures.

These improvements lead to better segmentation accuracy, showing RIFNet's ability to generate more clinically valuable fusion results.

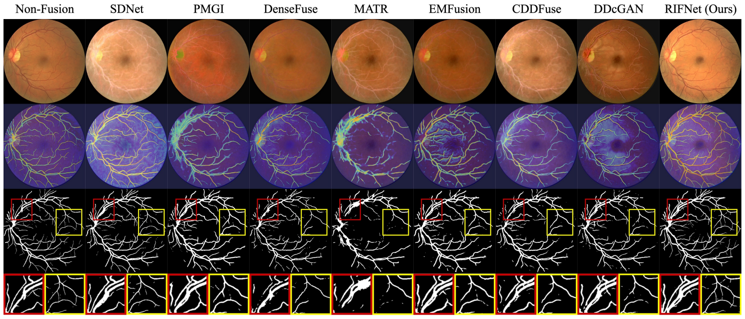


Fig. 3: The segmentation results and Grad-CAM results *(2nd row)* on fused images obtained by different methods.

4 Conclusion

In this study, we propose a multi-stage retinal image fusion framework, RIFNet, to address the limitations of single-modal fundus imaging by integrating complementary features from CFP and FA. The proposed framework employs a bi-stream generative subnetwork to synthesize pseudo-FA images, effectively compensating for scarce FA data caused by its invasive nature. By unifying multimodal color representations with the viridis colormap, we mitigate perceptual bias and enhance pixel-level pathological details. A GAN-based fusion network further optimizes global and local feature integration, while a multi-scale reconstruction strategy ensures high-resolution, high-contrast outputs. Experimental results demonstrate that RIFNet significantly improves image quality and the performance of downstream vascular segmentation tasks.

Acknowledgments. This research was supported by the National Natural Science Foundation of China (No.62076059), the Science and Technology Joint Project of Liaoning province (2023JH2/101700367, ZX20240193). Osmar Zaiane acknowledges the funding from NSERC and the Canada CIFAR AI Chairs Program.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Alcantarilla, P.F., Solutions, T.: Fast explicit diffusion for accelerated features in nonlinear scale spaces. IEEE Trans. Patt. Anal. Mach. Intell **34**(7), 1281–1298 (2011)

2. Bennett, T.J., Quillen, D.A., Coronica, R.: Fundamentals of fluorescein angiography. Curr Concepts Ophthalmology **9**(3), 43–9 (2001)
3. Bhutto, J.A., Tian, L., Du, Q., Sun, Z., Yu, L., Tahir, M.F.: Ct and mri medical image fusion using noise-removal and contrast enhancement scheme with convolutional neural network. Entropy **24**(3), 393 (2022)
4. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM **24**(6), 381–395 (1981)
5. Haddadpour, M., Daneshvar, S., Seyedarabi, H.: Pet and mri image fusion based on combination of 2-d hilbert transform and ihs method. Biomedical journal **40**(4), 219–225 (2017)
6. Hajeb Mohammad Alipour, S., Rabbani, H., Akhlaghi, M.R.: Diabetic retinopathy grading by digital curvelet transform. Computational and mathematical methods in medicine **2012**(1), 761901 (2012)
7. Han, Y., Cai, Y., Cao, Y., Xu, X.: A new image fusion performance metric based on visual information fidelity. Information fusion **14**(2), 127–135 (2013)
8. Hou, Q., Cheng, S., Cao, P., Yang, J., Liu, X., Tham, Y.C., Zaiane, O.R.: A clinical-oriented multi-level contrastive learning method for disease diagnosis in low-quality medical images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 13–23. Springer (2024)
9. Hou, Q., Wang, Y., Cao, P., Cheng, S., Lan, L., Yang, J., Liu, X., Zaiane, O.R.: A collaborative self-supervised domain adaptation for low-quality medical image enhancement. IEEE Transactions on Medical Imaging (2024)
10. Hou, R., Zhou, D., Nie, R., Liu, D., Ruan, X.: Brain ct and mri medical image fusion using convolutional neural networks and a dual-channel spiking cortical model. Medical & biological engineering & computing **57**, 887–900 (2019)
11. Li, H., Wu, X.J.: Densefuse: A fusion approach to infrared and visible images. IEEE Transactions on Image Processing **28**(5), 2614–2623 (2018)
12. Liu, W., Yang, H., Tian, T., Cao, Z., Pan, X., Xu, W., Jin, Y., Gao, F.: Full-resolution network and dual-threshold iteration for retinal vessel and coronary angiograph segmentation. IEEE journal of biomedical and health informatics **26**(9), 4623–4634 (2022)
13. Liu, Z., Song, Y., Sheng, V.S., Xu, C., Maere, C., Xue, K., Yang, K.: Mri and pet image fusion using the nonparametric density model and the theory of variable-weight. Computer Methods and Programs in Biomedicine **175**, 73–82 (2019)
14. Ma, J., Xu, H., Jiang, J., Mei, X., Zhang, X.P.: Ddcgan: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion. IEEE Transactions on Image Processing **29**, 4980–4995 (2020)
15. Panigrahy, C., Seal, A., Mahato, N.K.: Mri and spect image fusion using a weighted parameter adaptive dual channel pcnn. IEEE Signal Processing Letters **27**, 690–694 (2020)
16. Petrovic, V., Cootes, T.: Information representation for image fusion evaluation. In: 2006 9th international conference on information fusion. pp. 1–7. IEEE (2006)
17. Shi, D., Zhang, W., He, S., Chen, Y., Song, F., Liu, S., Wang, R., Zheng, Y., He, M.: Translation of color fundus photography into fluorescein angiography using deep learning for enhanced diabetic retinopathy screening. Ophthalmology science **3**(4), 100401 (2023)
18. Staal, J., Abràmoff, M.D., Niemeijer, M., Viergever, M.A., Van Ginneken, B.: Ridge-based vessel segmentation in color images of the retina. IEEE transactions on medical imaging **23**(4), 501–509 (2004)

19. Tan, W., Thitøn, W., Xiang, P., Zhou, H.: Multi-modal brain image fusion based on multi-level edge-preserving filtering. Biomedical Signal Processing and Control **64**, 102280 (2021)
20. Tang, W., He, F., Liu, Y., Duan, Y.: Matr: Multimodal medical image fusion via multiscale adaptive transformer. IEEE Transactions on Image Processing **31**, 5134–5149 (2022)
21. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8798–8807 (2018)
22. Xu, H., Ma, J.: Emfusion: An unsupervised enhanced medical image fusion network. Information Fusion **76**, 177–186 (2021)
23. Xydeas, C.S., Petrovic, V., et al.: Objective image fusion performance measure. Electronics letters **36**(4), 308–309 (2000)
24. Zhang, H., Ma, J.: Sdnet: A versatile squeeze-and-decomposition network for real-time image fusion. International Journal of Computer Vision **129**(10), 2761–2785 (2021)
25. Zhang, H., Xu, H., Xiao, Y., Guo, X., Ma, J.: Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 12797–12804 (2020)
26. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
27. Zhao, Z., Bai, H., Zhang, J., Zhang, Y., Xu, S., Lin, Z., Timofte, R., Van Gool, L.: 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. pp. 5906–5916 (2023)