TaGAT: Topology-Aware Graph Attention Network For Multi-modal Retinal Image Fusion

Xin Tian[1](✉), Nantheera Anantrasirichai[1], Lindsay Nicholson[2], and Alin Achim[1]

[1] Visual Information Laboratory, University of Bristol, Bristol, UK
[2] Autoimmune Inflammation Research, University of Bristol, Bristol, UK
{xin.tian, n.anantrasirichai, l.nicholson, alin.achim}@bristol.ac.uk

Abstract. In the realm of medical image fusion, integrating information from various modalities is crucial for improving diagnostics and treatment planning, especially in retinal health, where the important features exhibit differently in different imaging modalities. Existing deep learning-based approaches insufficiently focus on retinal image fusion, and thus fail to preserve enough anatomical structure and fine vessel details in retinal image fusion. To address this, we propose the Topology-Aware Graph Attention Network (TaGAT) for multi-modal retinal image fusion, leveraging a novel Topology-Aware Encoder (TAE) with Graph Attention Networks (GAT) to effectively enhance spatial features with retinal vasculature's graph topology across modalities. The TAE encodes the base and detail features, extracted via a Long-short Range (LSR) encoder from retinal images, into the graph extracted from the retinal vessel. Within the TAE, the GAT-based Graph Information Update block dynamically refines and aggregates the node features to generate topology-aware graph features. The updated graph features with base and detail features are combined and decoded as a fused image. Our model outperforms state-of-the-art methods in Fluorescein Fundus Angiography (FFA) with Color Fundus (CF) and Optical Coherence Tomography (OCT) with confocal microscopy retinal image fusion. The source code can be accessed via https://github.com/xintian-99/TaGAT.

Keywords: Multi-modal Image Fusion · Graph Attention Network · Multi-modal Retinal Image.

1 Introduction

Multi-modal medical image fusion aims to combine the complementary information from various medical imaging modalities, thereby aiding in more comprehensive diagnostics and treatment planning in brain, lungs, eye/retina, and cardiac [1]. In ophthalmology, this can involve the fusion of Color Fundus (CF) images with Fluorescein Fundus Angiography (FFA), Optical Coherence Tomography (OCT) with Fundus images, and OCT with confocal microscopy images [16], among others. An example illustrating the need for fusion arises when

the contrast between the retinal vasculature and the background in CF is limited, thereby complicating the analysis of small retinal vessels. Conversely, FFA images enhance the visibility of the retinal vasculature by employing a fluorescent dye [5]. The fusion of CF and FFA can integrate the high-resolution detail of pathologies in CF images with the enhanced vascular contrast from FFA. This integration furnishes a more detailed and comprehensive representation of the retinal structure [8], which can facilitate the early detection, accurate diagnosis, and effective monitoring of ocular diseases such as Diabetic Retinopathy (DR) [20]. The results of image fusion not only enhance the visualisation and analysis of retinal diseases by clinicians but also potentially support a range of downstream tasks, including vessel segmentation, disease classification, and monitoring of disease progression [1,9,25,26,27].

The current deep learning-based multi-modal image fusion has achieved significant advancements with two primary branches: generation-based methods (e.g. diffusion model [6], generative adversarial networks [4]), and discrimination-based methods (e.g. auto-encoder) [1]. DDFM [26] is a generative method utilising a denoising diffusion-based posterior sampling model to preserve more details for image fusion. SwinFusion [12] used cross-domain long-range learning and the Swin Transformer [10] to efficiently integrate structure, detail, and intensity across modalities. CDDFuse [25] is an auto-encoder-based model using a decomposition loss to modulate between modality-specific and shared features extracted through a dual-branch Transformer-CNN Long-short Range (LSR) encoder to leverage CNN's proficiency in capturing local spatial details and Transformer's capability in modelling long-range dependencies [17,23]. With the advancements in feature representation capabilities of Graph Neural Networks [22,7], IGNet [9] employed a fixed node weights GNN for cross-modality feature interaction. However, their approach to graph construction is solely based on feature space and does not incorporate the spatial structures of the images. Although these methods are effective in Visible-Infrared, MRI-CT, and MRI-PET fusion tasks, our findings reveal that they often fail to capture detailed features of the retinal vasculature and optic disc areas, particularly in abnormal retinas, when applied to retinal image fusion.

To address this gap in retinal image fusion, we introduce the Topology-Aware Graph Attention Network (TaGAT) for multi-modal retinal image fusion as shown in Fig. 1. A Topology-Aware Encoder (TAE) is proposed to bridge the base and detail spatial features in Euclidean space with the underlying graph topology in the non-Euclidean geometric space of retinal vasculature. This leverages the consistent topological properties of vascular structures across different retinal imaging modalities, which enhances feature representation and model generalisation. The TAE utilises base and detail features extracted by a Long-short Range (LSR) Encoder as node features, combined with a graph derived from the retinal vessel structure. With a Graph Attention Network (GAT), the TAE dynamically updates the graph by aggregating and refining node features, thereby connecting long-range structural features and preserving local details. Finally, a decoder is applied to reconstruct the fused image from the base, de-

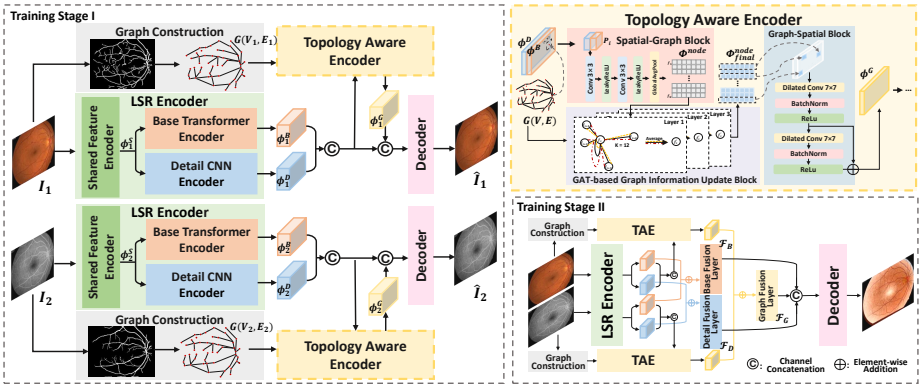

Fig. 1. Illustration of the proposed TaGAT framework and TAE.

tail, and graph features. In conclusion, our contribution can be summarised as follows.

1) We introduce an end-to-end framework of a topology-aware graph attention network for multi-modal retinal image fusion.
2) We propose a GAT-based Topology-Aware Encoder, the first to bridge spatial features with the consistent graph topology of retinal vasculature across modalities. This enhances feature representation and model generalisation and ensures the preservation of important anatomical structures and fine vasculature details in the retinal image fusion.
3) Our method achieves leading performances in retinal image fusion evaluated on both the DRFF(FFA-CF) and OCT2Confocal datasets, with exceptional preservation of fine structures, details, and textures.

2 Methodology

The proposed framework for multi-modal retinal image fusion is shown in Fig. 1, where the image inputs are registered, the significant features of each modality are enhanced and finally fused. We employ an LSR encoder to extract the base and detail features across modalities. The proposed GAT-based TAE encodes and updates these features with the graph topology extracted from the vessel structure. Then, the base, detail, and graph features are fused and decoded to the image domain. We employ a two-stage training strategy [25], where the decoder in Stage I reconstructs original images and in Stage II generates fusion images.

2.1 Graph Construction

The graph is constructed based on a tailored wavelet-based segmentation of blood vessels [14] from retinal images I_1 and I_2 of Modality 1 and Modality 2, respectively. After vessel segmentation, the vascular branching points and

endpoints are identified as graph nodes V, $V = \{v_i\}_{i=0}^{N}$, where v_i is a vertex i of the total N vertices. These nodes V are interconnected through edges E, $E = \{e_j\}_{j=0}^{M}$, where e_j is an edge j of the total M edges. The interconnection is based on the connectivity of V within the vessel network. Consequently, we have the graphs $G_1(V_1, E_1)$ and $G_2(V_2, E_2)$, which capture essential vascular characteristics such as lines, shapes, and topological configurations. By identifying the nodes' locations within the images, the graph links the graph topology (geodimensional space) and spatial domain (Euclidean space within the image domain).

2.2 Long-short Range Encoder

The LSR Encoder [25] is a dual branch encoder with three components:

i) *Shared Feature Encoder* (SFE): Restormer block [24]-based encoder extracts shared shallow features Φ_1^S and Φ_2^S across modalities without increasing computational complexity.

ii) *Base Transformer Encoder* (BTE): Lite Transformer (LT) block [21]-based encoder extracts low-frequency base features Φ_1^B and Φ_2^B from the shared features.

iii) *Detail CNN Encoder* (DCE): Invertible Neural Networks (INN) block [3]-based encoder extracts high-frequency details Φ_1^D and Φ_2^D from the shared features for preserving edge and texture information in both modalities.

2.3 Topology Aware Encoder

The proposed TAE encoder is designed to integrate spatial and topological information from retinal images. It comprises three main blocks:

Spatial-to-Graph Block (S2G) maps spatial features to the graph domain. The concatenated base and detail feature maps are reduced in their number of channels via convolution with a kernel size of 1 to compress features and reduce computation:

$$\Phi_{reduced} = \mathrm{Conv}_{1\times1}\left(\mathrm{Concat}\left(\Phi^B, \Phi^D\right)\right). \tag{1}$$

Subsequently, at each node i of $G(V, E)$, the feature patch P_i with the size of $p \times p$ ($p = 21$) is extracted from $\Phi_{reduced}$ as expressed in Eq.2, where (x_i, y_i) is the spatial location of node i.

$$\mathbf{P}_i = \Phi_{reduced}[:, (y_i - \frac{p}{2}) : (y_i + \frac{p}{2}), (x_i - \frac{p}{2}) : (x_i + \frac{p}{2})]. \tag{2}$$

Then, $\mathbf{P}_i$ are encoded through convolutions with kernel size 3 and LeakyReLU activations.The global average pooling is applied to yield a single feature vector f_i for each node i:

$$f_i = \mathrm{GlobalAvgPool}\left(\mathrm{LeakyReLU}\left(\mathrm{Conv}_{3\times3}\left(\mathrm{LeakyReLU}\left(\mathrm{Conv}_{3\times3}\left(\mathbf{P}_i\right)\right)\right)\right)\right). \tag{3}$$

Subsequently, the final graph node feature matrix $\mathbf{\Phi}^{node} = \{f_1, f_2, ..., f_N\}$. $\mathbf{\Phi}^{node} \in \mathbb{R}^{N \times C}$, where N is the number of nodes and C is the dimensionality of the feature vectors. The $\mathbf{\Phi}^{node}$ implicitly integrates spatial attributes into the graph for further updating by GAT-GIU block.

GAT-based Graph Information Update Block (GAT-GIU) employs a multi-layer, multi-head GAT [18] structure to iteratively refine node features $\mathbf{\Phi}^{node}$ through attention-driven, weighted aggregation of neighbourhood information. The node features first undergo a linear transformation, $\mathbf{H} = \mathbf{\Phi}^{node}W$, where W is a weight matrix. Then an attention mechanism is employed to compute attention coefficients e_{ij} for each node pair (i, j). These coefficients are then normalised across all neighbours to ensure selective attention α_{ij}:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k \in \mathcal{N}_i} \exp(e_{ik})}, \quad e_{ij} = \text{LeakyReLU}(\mathbf{a}^T[\mathbf{H}_i \| \mathbf{H}_j]), \tag{4}$$

and the updated node features $\mathbf{\Phi}^{node}_{updated}$ based on weighted neighborhood feature aggregation is defined as Eq. 5. The outputs of K heads multi-head attention are averaged to produce $\mathbf{\Phi}^{node}_{final}$. ELU is the Exponential Linear Unit [2].

$$\mathbf{\Phi}^{node}_{final} = \frac{1}{K} \sum_{k=1}^{K} \mathbf{\Phi}^{node}_{updated,k}, \;\; \mathbf{\Phi}^{node}_{updated} = \text{ELU}\left(\sum_{j \in \mathcal{N}_i} \alpha_{ij} \mathbf{H}_j \right). \tag{5}$$

Graph-to-Spatial Block (G2S) first maps $\mathbf{\Phi}^{node}_{final}$ onto a feature matrix based on corresponding node coordinates and then diffuse across the spatial domain using convolutions with a larger kernel size (7×7) and dilation to extend the spatial influence of node features. Skip connections are applied to merge the adjusted features for producing the final enhanced feature $\mathbf{\Phi}^G$ with graph topology.

2.4 Decoder

The decoder reduces the channel of concatenated features and uses the Restomer block for decoding. In training stage I, it concatenates encoders extracted features via channel dimension as input and the reconstructed image $\hat{I}_1$ and $\hat{I}_2$ as output. In the training stage II, the decoder takes the concatenation of features processed through feature fusion layers $\mathcal{F}_\mathcal{B}$, $\mathcal{F}_\mathcal{D}$, and $\mathcal{F}_\mathcal{G}$ (base, detail, and graph, respectively) as input and a fused image I_f as the output.

2.5 Loss Function

For Training Stage I, the total loss $\mathcal{L}^I_{total}$ is:

$$\mathcal{L}^I_{total} = \mathcal{L}_1 + \alpha_1 \mathcal{L}_2 + \alpha_2 \mathcal{L}_{decomp} + \alpha_3 \mathcal{L}_{graph}, \tag{6}$$

$\mathcal{L}_1$ and $\mathcal{L}_2$ are the reconstruction losses [25] for Modality 1 and Modality 2, ensuring original image information preservation during encoding and decoding:

$$\mathcal{L}_m = \mathcal{L}_{int}^{I}(I_m, \hat{I_m}) + \mu\mathcal{L}_{SSIM}(I_m, \hat{I_m}), \ \ m \in \{1, 2\} \tag{7}$$

where $\mathcal{L}_{int}^{I} = \|I_m - \hat{I_m}\|_2^2$ is the intensity loss [13], and $\mathcal{L}_{SSIM}(I_m, \hat{I_m}) = 1 - SSIM(I_m, \hat{I}_m)$ [19].

The $\mathcal{L}_{decomp}$ denotes the feature decomposition loss [25] :

$$\mathcal{L}_{decomp} = \frac{\left(\mathcal{L}_{CC}^{D}\right)^2}{\mathcal{L}_{CC}^{B}} = \frac{\left(\mathcal{CC}\left({\Phi_1}^D, {\Phi_2}^D\right)\right)^2}{\mathcal{CC}\left({\Phi_1}^B, {\Phi_2}^B\right) + \epsilon} \tag{8}$$

where $\mathcal{CC}(\cdot, \cdot)$ is the correlation coefficient operator, ϵ is set to 1.01 keeping this term positive. The $\mathcal{L}_{decomp}$ extracts detail and base features by modulating the correlation between low-frequency and high-frequency components accordingly.

$$\mathcal{L}_{graph} = 1 - \frac{\langle {\Phi_1}^G, {\Phi_2}^G \rangle}{|{\Phi_1}^G| \cdot |{\Phi_2}^G|} \tag{9}$$

The $\mathcal{L}_{graph}$ employed cosine similarity emphasises the directional alignment of GAT-encoded features over magnitude to maintain the vascular topology similarity and adjacency information across modalities.

For Training Stage II, the total loss is:

$$\mathcal{L}_{total}^{II} = \mathcal{L}_{int}^{II} + \alpha_3\mathcal{L}_{graph} + \alpha_4\mathcal{L}_{grad} + \alpha_5\mathcal{L}_{decomp}, \tag{10}$$

where $\mathcal{L}_{grad} = \frac{1}{HW}\| \, |\nabla I_f| - \max(|\nabla I_1|, |\nabla I_2|)\|_1$ ensures more fine-grained texture information [13]. $\mathcal{L}_{int}^{II} = \frac{1}{HW}\|I_f - \max(I_1, I_2)\|_1$. ∇ is the Sobel gradient operator. α_{1-5} are the hyperparameters.

3 Experimental Results and Discussion

3.1 Datasets

Two datasets were involved in our experiments. i) **DRFF** [5]: The DRFF dataset comprises 30 abnormal and 29 normal unregistered FFA-Fundus pairs. We applied the segmentation and registration method from [14] and used a subset comprising 20 normal and 20 abnormal pairs for training and 19 pairs for testing. Data augmentation with flipping, rotating by ±8 degrees, and translating by ±20 pixels are applied. ii) **OCT2Confocal** [15]: The dataset has paired grayscale OCT and corresponding coloured confocal microscopy retinal images from 3 mice afflicted with autoimmune uveitis. The registration is through manual registration and confirmed by an ophthalmologist. We use this data to test models trained on the DRFF dataset.

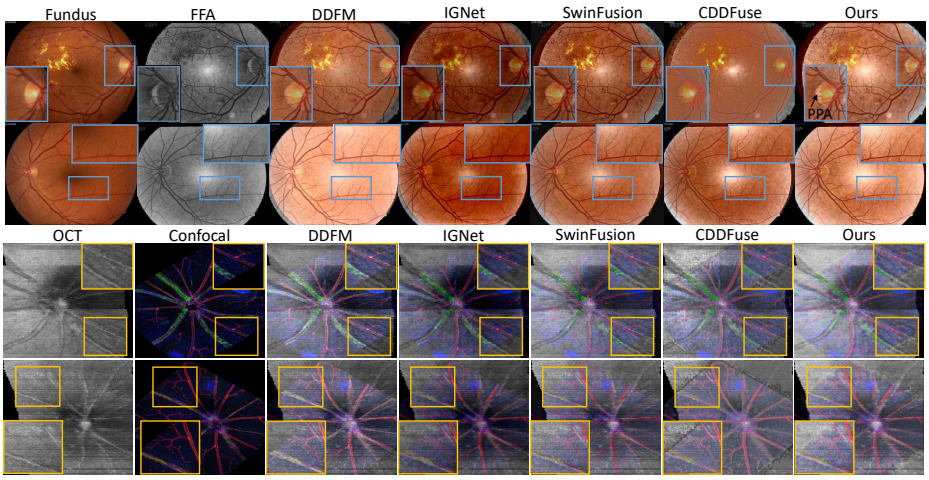

Fig. 2. Visual comparison results in DRFF and OCT2Confocal.

3.2 Experimental Setup and Evaluation Metrics

Our computational experiments were conducted on a high-performance computing environment featuring NVIDIA Tesla V100 GPUs (32 GB). Training is conducted with the first stage 40 epochs and the second stage 80 epochs. The images are resized to 288×360 pixels for training with a batch size of 1 due to memory limitation. The Adam optimiser is employed, with an initial learning rate of 10^{-4} and decay by a factor of 0.5 every 20 epochs.

In the LRS encoder, both the SFE and BTE are configured with 4 Restormer blocks, utilising 4 attention heads (compared to 8 as used in CDDFuse [25]) within a 64-dimensional embedding space. The GIU block is equipped with 12 attention heads, each operating in a 64-dimensional space. The decoder employs 4 Restormer blocks with 4 attention heads. For the loss function, the weighting coefficients α_1 through α_5 are finely tuned to the values of 1, 2, 0.5, 10, and 2.

We use eight metrics to measure the fusion results [11]: entropy (EN), standard deviation (SD), spatial frequency (SF), mutual information (MI), sum of the correlations of differences (SCD), visual information fidelity (VIF), $Q^{AB/F}$ and SSIM. Higher metrics indicate that a fusion image is better.

3.3 Benchmarking Results

We tested our model and compared the fusion results with the existing benchmarks including DDFM [26], IGNet [9], SwinFusion [12], and CDDFuse [25].

For the FFA-Fundus dataset, Table 1 (left) showcases our model's leading performance, particularly highlighted by the highest SD and SF, indicating marked improvements in detail and structure preservation. Despite competitive VIF and SSIM scores from SwinFusion and CDDFuse, our model demonstrates

Table 1. Quantitative results of the DRFF and OCT2Confocal retinal image fusion. **Bold** and Underline show the best and second-best results, respectively.

	DRFF Retinal Images								OCT2Confocal Retinal Images							
	EN	SD	SF	MI	SCD	VIF	$Q^{AB/F}$	SSIM	EN	SD	SF	MI	SCD	VIF	$Q^{AB/F}$	SSIM
DDFM [26]	6.55	58.9	14.07	1.41	1.3	0.22	0.21	0.27	7.01	38.21	15.98	1.08	1.4	0.17	0.19	0.36
IGNet [9]	6.75	39.12	12.28	1.61	0.29	0.66	0.49	0.91	3.45	16.16	5.41	0.6	0.54	0.19	0.17	0.36
SwinFusion [12]	6.86	49.01	16.41	3.15	0.66	1.03	0.65	0.99	**7.12**	41.05	17.2	2.71	1.19	0.73	0.6	0.95
CDDFuse [25]	**7.08**	57.22	17.06	2.88	0.72	0.91	0.64	**1.01**	7.05	41.28	17.87	1.56	1.28	0.48	0.4	0.87
Ours	6.97	**69**	**19.22**	**3.45**	**1.52**	**1.03**	**0.66**	0.96	6.94	**67.59**	**19.18**	**3.36**	**1.54**	**1**	**0.66**	**0.98**

Table 2. Ablation experiments results with DRFF. **Bold** indicates the best value.

	Configurations	SD	MI	VIF	SSIM
I	w/o L_{graph}	67.23	3.4	1.01	0.96
II	w/o G2S	65	2.61	0.66	0.84
III	w/o Φ^B and Φ^D for Decoder	39.26	1.02	0.25	0.41
IV	w/o Φ^G	66.69	3.38	0.64	0.92
V	GAT $\rightarrow$ GCN	68.29	2.94	0.88	0.95
	Ours	**69**	**3.45**	**1.03**	**0.96**

balanced performance across all metrics. Additionally, visual results in Fig. 2 (top 3 rows) highlight our model's ability to clearly delineate the optic disc's shape and maintain fine vasculature and texture details, indicating the TAE effectively encodes vessel topology into features to enhance focus on vasculature. The first row in Fig. 2 also demonstrates our model's ability to reveal the Peripapillary Atrophy (PPA) region, characterised by atrophic changes and irregular retinal pigmentation around the optic disc, which is challenging to discern in standard CF. Identifying PPA is crucial for diagnosing conditions like diabetic retinopathy.

For OCT2Confocal dataset, our method outperforms other methods as shown in Table 1 (right). However, visual results in Fig. 2 (bottom 2 rows) indicate none of the tested methods offers sufficient detail and clarity preservation across both modalities. Compared to CDDFuse, our results show a denoising effect and exhibit fewer border artefacts. This improvement is attributed to our topology-aware graph feature, which emphasises vessel-related information for fusion, effectively minimising irrelevant features such as noise.

Ablation Studies We verified the effectiveness of i) the graph loss L_{graph}, ii) G2S block, iii) Φ^B and Φ^D, iv) Φ^G, and v) GAT. Table 2 shows that without L_{graph} or G2S block leads to a slight decrease in all metrics highlighting their role in refining the feature representations. Removing Φ^G slightly diminishes the performance due to less attention around vessels. When excluding Φ^B and Φ^D from the decoder, the marked reduction across all evaluated metrics suggests that graph-related features are insufficient for reconstructing a full Euclidean

space image primarily due to lack of the detailed pixel-level information. Substituting GAT with normal GCN [7] is to validate the utility of dynamic attention mechanisms. The reduced performance suggests the effectiveness of the dynamic attention mechanisms for GAT in feature aggregation.

4 Conclusion

This paper presents a multimodal retinal image fusion method with a novel GAT-based TAE feature encoder that effectively bridges spatial-temporal and graph topology characteristics across different modalities. Our approach has demonstrated superior performance in enhancing key feature visualisation such as the clarity of optic disc and PPA, the preservation of fine vasculature and texture details in FFA-Fundus fusion with the ablation studies validating the significance of each model component. In future work, we aim to enhance the fusion of low-quality and high-resolution images and extend our approach to other types of medical images with vessel structure, such as brain MRI and CT scans.

Acknowledgments. Xin is supported by the China Scholarship Council.

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

References

1. Azam, M.A., Khan, K.B., Salahuddin, S., Rehman, E., Khan, S.A., Khan, M.A., Kadry, S., Gandomi, A.H.: A review on multimodal medical image fusion: Compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics. Computers in biology and medicine **144**, 105253 (2022)
2. Clevert, D.A., Unterthiner, T., Hochreiter, S.: Fast and accurate deep network learning by exponential linear units (elus). arXiv preprint arXiv:1511.07289 (2015)
3. Dinh, L., Sohl-Dickstein, J., Bengio, S.: Density estimation using real NVP. In: ICLR (2017)
4. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
5. Hajeb Mohammad Alipour, S., Rabbani, H., Akhlaghi, M.R., et al.: Diabetic retinopathy grading by digital curvelet transform. Computational and mathematical methods in medicine **2012** (2012)
6. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
7. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
8. Laliberte, F., Gagnon, L., Sheng, Y.: Registration and fusion of retinal images-an evaluation study. IEEE Transactions on Medical Imaging **22**(5), 661–673 (2003)
9. Li, J., Chen, J., Liu, J., Ma, H.: Learning a graph neural network with cross modality interaction for image fusion. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 4471–4479 (2023)

10. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
11. Ma, J., Ma, Y., Li, C.: Infrared and visible image fusion methods and applications: A survey. Information Fusion **45**, 153–178 (2019)
12. Ma, J., Tang, L., Fan, F., Huang, J., Mei, X., Ma, Y.: Swinfusion: Cross-domain long-range learning for general image fusion via swin transformer. IEEE/CAA Journal of Automatica Sinica **9**(7), 1200–1217 (2022)
13. Tang, L., Yuan, J., Ma, J.: Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network. Inf. Fusion **82**, 28–42 (2022)
14. Tian, X., Anantrasirichai, N., Nicholson, L., Achim, A.: Optimal transport-based graph matching for 3D retinal OCT image registration. In: 2022 IEEE International Conference on Image Processing (ICIP). pp. 2791–2795. IEEE (2022)
15. Tian, X., Anantrasirichai, N., Nicholson, L., Achim, A.: OCT2Confocal: 3D cycle-gan based translation of retinal OCT images to confocal microscopy. arXiv preprint arXiv:2311.10902 (2023)
16. Tian, X., Zheng, R., Chu, C.J., Bell, O.H., Nicholson, L.B., Achim, A.: Multi-modal retinal image registration and fusion based on sparse regularization via a generalized minimax-concave penalty. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1010–1014. IEEE (2019)
17. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
18. Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y., et al.: Graph attention networks. stat **1050**(20), 10–48550 (2017)
19. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., et al.: Image quality assessment: from error visibility to structural similarity. IEEE TIP **13**(4), 600–612 (2004)
20. Wong, T.Y., Cheung, N., Tay, W.T., Wang, J.J., Aung, T., Saw, S.M., Lim, S.C., Tai, E.S., Mitchell, P.: Prevalence and risk factors for diabetic retinopathy: the singapore malay eye study. Ophthalmology **115**(11), 1869–1875 (2008)
21. Wu, Z., Liu, Z., Lin, J., Lin, Y., Han, S.: Lite transformer with long-short range attention. In: ICLR (2020)
22. Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., Philip, S.Y.: A comprehensive survey on graph neural networks. IEEE transactions on neural networks and learning systems **32**(1), 4–24 (2020)
23. Xie, H., Huang, Z., Leung, F.H., Ju, Y., Zheng, Y.P., Ling, S.H.: A structure-affinity dual attention-based network to segment spine for scoliosis assessment. In: 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 1567–1574. IEEE (2023)
24. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.: Restormer: Efficient transformer for high-resolution image restoration. In: CVPR. pp. 5718–5729 (2022)
25. 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 (CVPR). pp. 5906–5916 (June 2023)

26. Zhao, Z., Bai, H., Zhu, Y., Zhang, J., Xu, S., Zhang, Y., Zhang, K., Meng, D., Timofte, R., Van Gool, L.: DDFM: denoising diffusion model for multi-modality image fusion. arXiv preprint arXiv:2303.06840 (2023)
27. Zhou, T., Ruan, S., Canu, S.: A review: Deep learning for medical image segmentation using multi-modality fusion. Array **3**, 100004 (2019)