Text-Guided Image Fusion

Jialing Huang[1], **Yufa Duan**[1], **Lichen Wei**[1], **Changcheng Li**[2]

[1]Artificial Intelligence Research Institute, Xiamen University [2]School of Infomatics, Xiamen University
36920241153213,36920241153202,36920241153254,30920241154552

Abstract

Image fusion aims to integrate multiple modalities of images into a single modality, thereby achieving complementary information and better adapting to downstream tasks such as segmentation and detection. Existing image fusion methods predominantly focus on visual information, often neglecting the crucial role that semantic information at the textual level can play in guiding the deep fusion of images. To address this, we propose a novel framework—Text-Guided Image Fusion (**TGIF**)—that leverages textual guidance for general image fusion tasks. Specifically, TGIF extracts visual information from images and employs large language models to generate corresponding textual descriptions. These descriptions are then used to obtain textual features, which guide the fusion process through cross-attention mechanisms, thereby enhancing information complementarity during the fusion. We will conduct extensive experiments on various image fusion tasks to validate the effectiveness of our framework, including both qualitative and quantitative analyses.

Introduction

Image fusion aims to enhance the visual quality of images and offer more accurate and reliable information for diverse applications (Zhang 2021). Multimodal image fusion is crucial in computer vision (Zhang et al. 2021), like infrared-visible (Zhao et al. 2023b), medical (James and Dasarathy 2014), multi-exposure (Ma et al. 2019b), and multi-focus image fusion (Zhang and Ma 2021). However, current image fusion overly depends on visual features, neglecting deeper semantic layers.While some advanced methods have attempted to incorporate downstream tasks such as semantic segmentation (Tang et al. 2022a) and object detection (Zhao et al. 2023a), these approaches are still mainly confined to superficial semantics derived from visual-level features, failing to tap into the more complex textual semantics that images can convey. Consequently, a key challenge that remains is how to effectively leverage deeper semantic features that extend beyond the visual information present in the images, which is a critical area in need of further research.

Vision Language Models (VLM), trained on image-text pairs, can align visual and language features for multimodal fusion. For example, models like CLIP (Radford et al. 2021) and GPT-4 (Achiam et al. 2023) have strong capabilities. We propose the Text-Guided Image Fusion (TGIF) model, which has four stages: vision features extraction, text features extraction, text-guided vision features fusion, and generation of the fused image. This approach incorporates VLMs into image fusion, leveraging textual semantic understanding to guide visual feature fusion.

Overall, our study addresses the problem that existing image fusion techniques do not fully utilize deep semantic information, and innovatively proposes an image fusion paradigm based on visual language models. Using a large language model to guide the fusion process based on the textual description of the source image, we can achieve a more comprehensive understanding of the image content.

Related work

Multimodal Image Fusion Methods.

Multimodal Image Fusion Methods Image fusion has drawn much attention recently. In the deep learning era, the main image fusion methods can be grouped into four categories: CNN models (Zhang et al. 2020b), GAN models (Ma et al. 2020), AE-based models (Li and Wu 2018), and Transformer-based models (Zhao et al. 2023b). These methods often use simple fusion rules. CNNs have limitations in extracting global features. GANs like FusionGAN (Ma et al. 2019a) and DDcGAN (Ma et al. 2020) show good performance but have training instability and potential texture distortion. AE-based models need a fusion rule. Transformer-based methods like IFT (Vs et al. 2022) and CDDFuse (Zhao et al. 2023b) combine CNNs with Transformer architectures, and SwinFusion (Ma et al. 2022) uses a unique attention mechanism. Recently, diffusion-based image fusion has emerged, such as DDFM (Zhao et al. 2023c).

Vision-Language Model.

Visual language multimodal learning has become a hot research topic. Vision-language models like BLIP (Li et al. 2023), DALL-E (Ramesh et al. 2022), and GPT4 (Achiam et al. 2023) perform well in downstream tasks. BLIP bridges visual and language models, and GPT4 has strong general performance. These models provide external knowledge for image captioning, and we are inspired to introduce a vision-language model into image fusion to enable text to guide the

process effectively.

Method

In this paper, we have a pair of input images, which can be infrared visible, medical, multi-exposure, or multi-focus images. The algorithm ultimately outputs a fused image. In this section, we will provide a comprehensive description of our TFIG algorithm, explaining its workflow and design details.

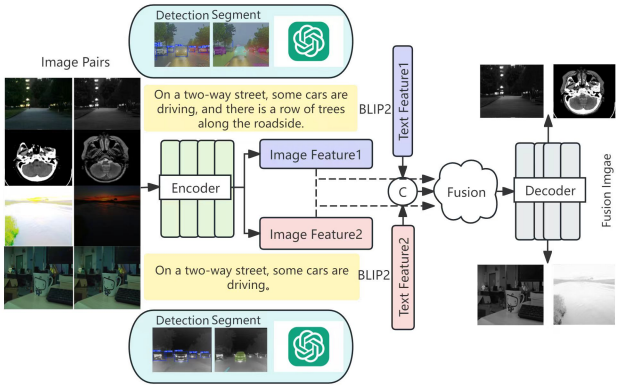

Figure 1: Frame diagram.

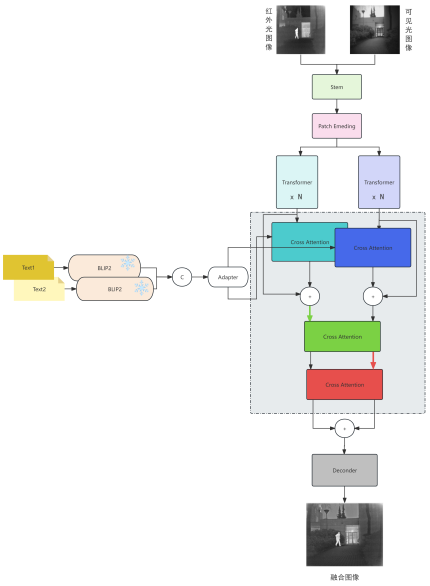

Figure 2: Detailed Procedure.

Workflow Overview

Brief and detailed workflows are illustrated in Figure 1 and Figure 2.The Image Fusion paradigm TGIF (text-guided Image Fusion) aims to enhance the effect of image fusion through Text guidance, ensuring that the final fused image can better retain the key information and semantic features of the source image. TGIF's workflow consists of three closely linked modules: a visual encoder, a text encoder, and a text-guided visual feature fusion module.

Text-Guided Image Fusion

Component I: Vision feature extraction. This module uses an encoder based on the Transformer architecture for visual feature extraction. The encoder is composed of stacked Transformer blocks, and its self-attention mechanism can capture long-distance dependencies between image regions, having an advantage in processing image data. In the workflow, paired source images are input into the encoder. Starting from the pixel level, features are gradually extracted through Transformer blocks. In the Transformer block, the self-attention mechanism calculates the degree of association between positions, highlighting key feature areas. As the feature information is passed through multiple blocks, the image feature representation is gradually refined, and finally, a feature vector containing the core visual features is output.

Component II: Text feature extraction. To obtain accurate and comprehensive text features, this module considers multiple input information, including the original image, its segmentation map, detection map, and necessary artificial prompts. The original image provides the visual scene content, the segmentation map determines the boundaries of objects or regions, the detection map gives the category and position information of objects, and the artificial prompts supplement details. These information help a subsequent large model to generate text describing the image.The process of generating data is shown in Figure 3. The generated text is processed by the pre-trained BLIP2 model. BLIP2 converts the text from the natural language space to the feature space and encodes it into a feature vector form suitable for guiding image fusion. During the training process, the BLIP2 model is frozen to ensure stability and consistency.

Component III: Text-Guided Vision Feature Fusion. This module first preprocesses the text features of the source image pair. Specifically, it splices the text features into a unified vector to integrate the text description information and provide a coherent and comprehensive text guidance for subsequent fusion. Then, an adapter is used to process the spliced text features to achieve the alignment of the text feature space and the image feature space. The adapter can automatically learn and adjust parameters according to the input text features and image features, making them in the same semantic dimension. After the text features are aligned, they perform cross-attention operations with the image features. The text features act as query vectors, and the image features act as key and value vectors. By calculating the attention weights, the text features guide the image feature extraction, highlighting relevant regions and features. This operation is repeated multiple times to make the image features fully absorb the semantic information in the text features and improve the fusion quality. Under the guidance of the cross-attention mechanism, the source image pair respectively performs cross-attention for feature fusion, com-

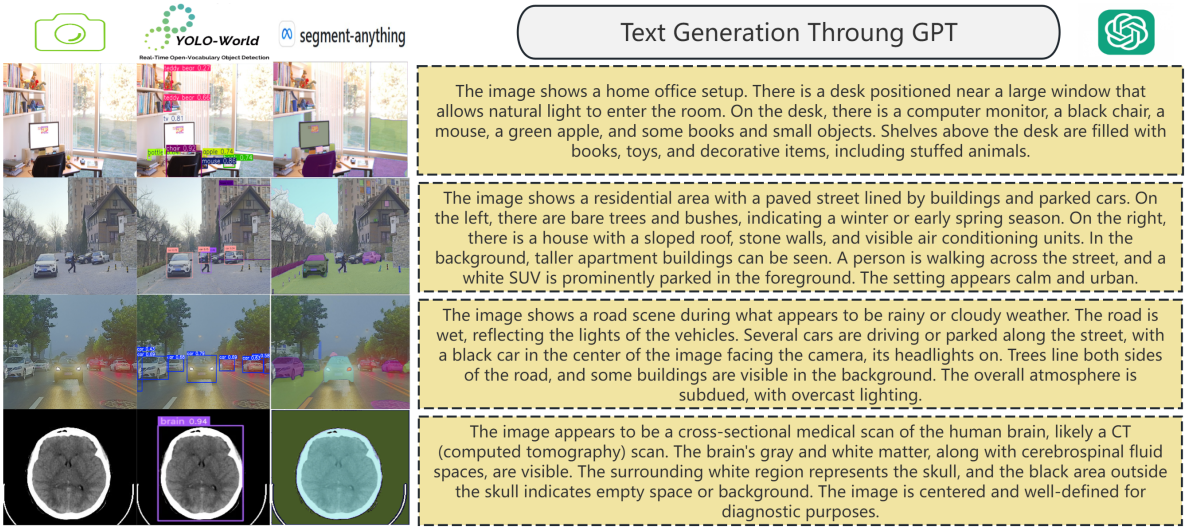

Figure 3: Visualization of the dataset creation process and representative data displays.

bining the advantageous features of the two images.The specific process of cross-attention is shown in Figure 4 .Finally, the fused features are processed by a decoder, which decodes the fused features into the final fused image. The structure and parameters of the decoder are designed according to the specific task and model architecture, and its role is to convert the fused features into an image form conforming to visual perception.

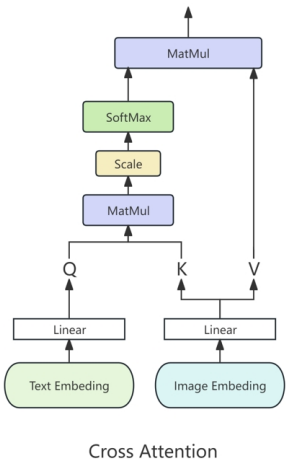

Figure 4: Cross attention.

Experiment

In this section, we will demonstrate the performance of TGIF on various image fusion tasks, showcasing its superiority.

Loss Function. For the total training loss, we set it as:

$$\mathcal{L}_{total} = \mathcal{L}_{int} + \alpha_1 \mathcal{L}_{grad} + \alpha_2 \mathcal{L}_{\text{SSIM}}, \quad (1)$$

where α_1, α_2 are tuning parameters. In the IVF task, following the setting in Zhao et al. (Zhao et al. 2023b), $\mathcal{L}_{int} = \frac{1}{HW} \|\mathcal{I}_F - \max(I_1, I_2)\|_1$, and $\mathcal{L}_{grad} = \frac{1}{HW} \||\nabla \mathcal{I}_F| - \max(|\nabla I_1|, |\nabla I_2|)\|_1$.$\nabla$ indicates the Sobel gradient operator. α_1 and α_2 are set to 20 and 0, respectively. MIF task does not need fine-tuning training, therefore it has no loss function. For MFF and MEF tasks, inspired by Liu et al. (Liu et al. 2023), we set $\mathcal{L}_{int} = \frac{1}{HW} \|\mathcal{I}_F - mean(I_1, I_2)\|_1$,$\mathcal{L}_{grad} = \frac{1}{HW} \||\nabla \mathcal{I}_F| - \max(|\nabla I_1|, |\nabla I_2|)\|_1$, and $\mathcal{L}_{SSIM} = 2 - SSIM(I_1, I_F) - SSIM(I_2, I_F)$.$\{\alpha_1, \alpha_2\}$ are set to {300,1} and {500,1} in MFF and MEF tasks respectively, in order to ensure the magnitude comparable in each term.

Training Details. A machine with a NVIDIA GeForce RTX 4090 GPUs is utilized for our experiments. We train the network for 300 epochs using the Adam optimizer, with an initial learning rate of 1e-4 and decreasing by a factor of 0.5 every 50 epochs. The Adam optimization strategy is employed with the batchsize set as 2. We incorporate Restormer blocks in both languageguided vision encoder $V(\cdot)$ and vision feature decoder $D(\cdot)$, with each block having 8 attention heads and a dimensionality of 64. M and N, representing the number of blocks in $V(\cdot)$ and $D(\cdot)$, and set to 2 and 3, respectively.

Metrics. We employ six quantitative metrics to assess the fusion outcomes: entropy(EN), standard deviation(SD), spatial frequency(SF), average gradient(AG), visual information fidelity(VIF) and $Q^{AB/F}$. Higher metric values indicate superior quality in the fused image.Further information is available in Ma et al. (Ma, Ma, and Li 2019).

Infrared and Visible Image Fusion

Setup. Infrared-visible fusion experiments are conducted on the MSRS (Tang et al. 2022b), 1083 image pairs in MSRS

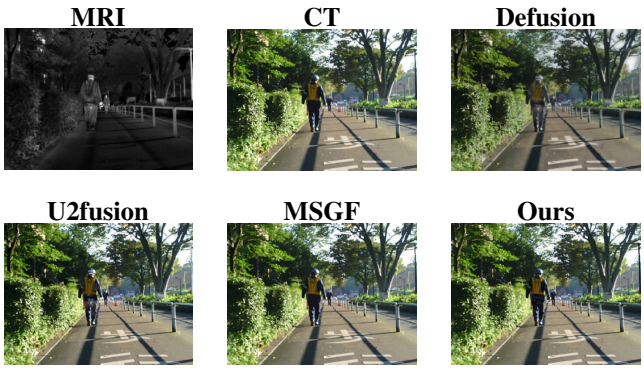

Figure 5: Visualization comparison of the fusion results in the infrared-visible image fusion task.

Table 1: Quantitative results of IVF.This table presents the quantitative performance evaluation of different methods on the MSRS Infrared-Visible Fusion dataset.

	EN	SD	SF	AG	VIF	Qabf
DEF	6.46	37.63	8.60	2.80	0.77	0.54
CDDF	6.70	43.39	11.56	3.74	1.05	0.69
DDFM	6.19	29.26	7.44	2.51	0.73	0.48
Ours	6.62	42.00	11.53	3.64	0.83	0.67

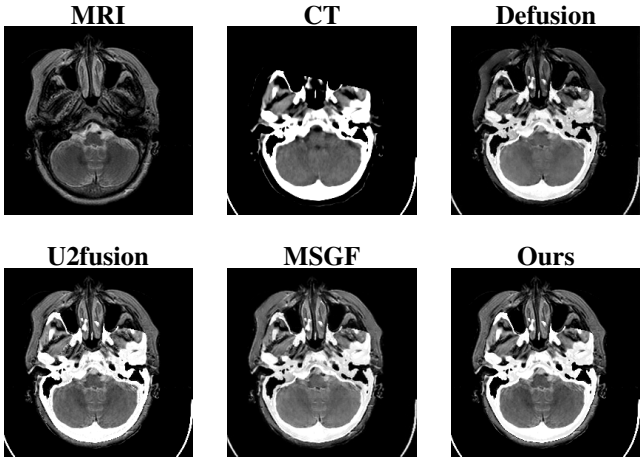

Figure 6: Visualization comparison of the fusion results in the medical image fusion task.

Table 2: Quantitative results of MIF.This table presents the quantitative performance evaluation of different methods on the Harvard Medical Image Fusion dataset.

	EN	SD	SF	AG	VIF	Qabf
DEF	3.90	54.77	16.87	4.30	0.62	0.57
U2F	3.56	49.95	19.70	4.98	0.47	0.53
MSGF	4.06	75.01	20.34	5.09	0.49	0.50
Ours	4.45	65.26	23.35	5.16	0.78	0.76

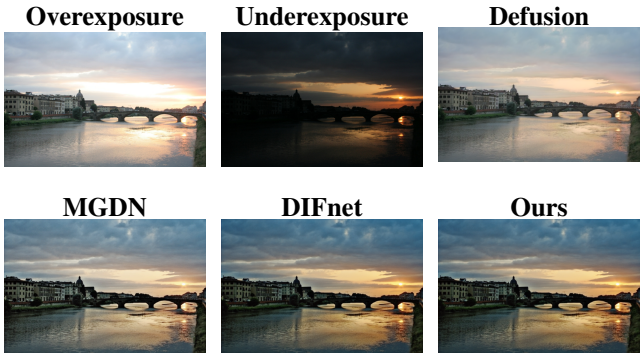

Figure 7: Visualization comparison of the fusion results in the multi-exposure image fusion task.

Table 3: Quantitative results of MEF.This table presents the quantitative performance evaluation of different methods on the SICE Multi-exposure Image Fusion dataset.

	EN	SD	SF	AG	VIF	Qabf
DEF	6.87	44.73	14.28	4.04	0.87	0.57
MGDN	6.94	43.69	15.04	4.59	0.88	0.64
DIFN	6.56	35.76	11.86	3.09	0.46	0.50
Ours	6.82	54.06	19.80	5.24	1.49	0.78

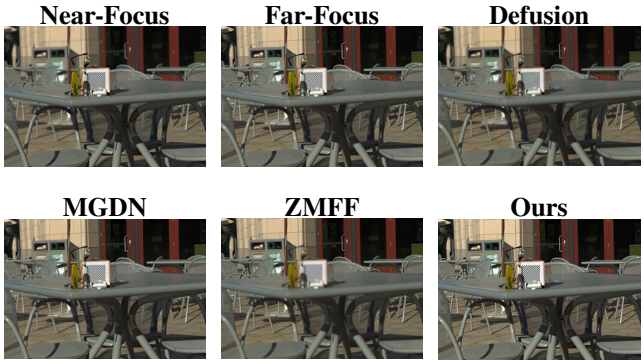

Figure 8: Visualization comparison of the fusion results and error maps in the multi-focus image fusion task.

Table 4: Quantitative results of MFF.This table presents the quantitative performance evaluation of different methods on the RealMFF Multi-focus Image Fusion dataset.

	EN	SD	SF	AG	VIF	Qabf
DEF	7.09	54.42	11.24	4.08	0.98	0.69
MGDN	7.09	54.24	15.15	5.24	1.07	0.75
ZMFF	6.99	51.15	13.93	4.95	0.94	0.70
Ours	7.14	56.64	15.52	5.34	1.54	0.74

Table 5: Ablation experiments results

Configurations				Metrics					
	Detection	Segment	GPT	EN	SD	SF	AG	VIF	Qabf
w/o text				6.36	40.89	11.03	3.52	0.61	0.53
w/o Segment	✓		✓	6.45	41.31	11.19	3.66	0.71	0.61
w/o Detection		✓	✓	6.48	41.23	11.23	3.63	0.73	0.59
Ours	✓	✓	✓	6.62	42.00	11.53	3.64	0.83	0.67

are for training and 361 pairs are for testing. We evaluated TGIF against various state-of-the-art(SOTA) methods including DeFusion, CDDFuse and DDFM.

Comparison with SOTA Methods. In Figure 5, TGIF successfully integrated the thermal radiation information with the detailed texture features. Leveraging textual features and knowledge, the fusion process enhanced the visibility of objects in low-light environments, making textures and contours clearer, and reducing artifacts. For the quantitative results in Table 1, our method showcases exceptional performance in almost all metrics, confirming its adaptability for various environmental scenarios and object categories. Hence, TGIF is proven to well maintain the completeness and richness of the information from source images, and generate results that conform to human visual perception.

Medical Image Fusion

Setup. we engage the Harvard Medical dataset (Johnson and Becker), which consisted of 50 pairs of MRI-CT, MRI-PET, and MRI-SPECT images,to evaluate the generalizability of our model.Notably, we employ the model trained on the IVF experiments and conducted a generalization test on the Harvard Medical dataset without any fine-tuning. The competitors include DeFusion , U2Fusion,and MsgFusion.

Comparison with SOTA Methods. In terms of visual perception and quantitative analysis (Figure 6 and Table 2), TGIF has shown outstanding accuracy in extracting cross-modal structural highlights and detailed texture features, effectively integrating source information into the fused images. These achievements surpass even those of fusion models specifically fine-tuned via medical image pairs.

Multi-exposure Image Fusion

Setup. We conduct MEF experiments on the SICE. We utilized 499 pairs from SICE dataset for training, while 90 pairs from SICE for testing. Our comparison methods encompass DIFNet,U2Fusion and DeFusion.

Comparison with SOTA Methods. Both quantitative and qualitative results in Table 3 and Figure 7 demonstrate the effectiveness of TGIF, which adeptly handles multiple images with varying exposures, expanding the dynamic range while simultaneously improving image quality and enhancing contrast.

Multi-focus Image Fusion

Setup. MFF experiments are conducted using RealMFF (Zhang et al. 2020a). 639 image pairs from RealMFF are employed for training. Comparative methods encompass DeFusion,ZMFF and MGDN.

Comparison with SOTA Methods. As illustrated in Figure 8, benefiting from textual descriptions, TGIF excels in identifying clear regions within multi-focus image pairs, ensuring sharp foreground and background elements. The quantitative results in Table 4 further underscore the excellence of our methodology.

Ablation Studies

To explore the effectiveness of each module in our proposed method, using the infrared-visible fusion task as an example, we conduct ablation studies on the test dataset of MSRS. The results are presented in Table 5.

Textual Guidance. In Exp. I, we removed the guidance through segmentation and detection prompts and only used the initial text for fusion.

Semantic Prompts. Then, in Exp. II-III, the original images, segmentation prompts and detection prompts were input into GPT respectively. Different text prompts were obtained, and these descriptions would be used as the text inputs for image fusion.

In conclusion, ablation experiments demonstrate that relying on the comprehensive information from different grains of captions and the powerful summarization capability of GPT, our experimental setup achieved optimal fusion performance, validating the rationality of our TGIF setting.

Conclusion

Our method provides a new perspective for the image fusion task by combining visual and language models. Future research may further explore how to introduce 3D into the multimodal image fusion process. After all, 3D can bring more perceptual effects and three - dimensional structural information that cannot be shown on a two - dimensional plane, which can further improve the quality of image fusion.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

James, A. P.; and Dasarathy, B. V. 2014. Medical image fusion: A survey of the state of the art. *Information fusion*, 19: 4–19.

Johnson, B. A.; and Becker, J. A. ???? Harvard medical website. http://www.med.harvard.edu/AANLIB/home.html. Accessed: [2022].

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, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Liu, J.; Wu, G.; Luan, J.; Jiang, Z.; Liu, R.; and Fan, X. 2023. HoLoCo: Holistic and local contrastive learning network for multi-exposure image fusion. *Information Fusion*, 95: 237–249.

Ma, J.; Ma, Y.; and Li, C. 2019. Infrared and visible image fusion methods and applications: A survey. *Information fusion*, 45: 153–178.

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.

Ma, J.; Xu, H.; Jiang, J.; Mei, X.; and Zhang, X.-P. 2020. DDcGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion. *IEEE Transactions on Image Processing*, 29: 4980–4995.

Ma, J.; Yu, W.; Liang, P.; Li, C.; and Jiang, J. 2019a. FusionGAN: A generative adversarial network for infrared and visible image fusion. *Information fusion*, 48: 11–26.

Ma, K.; Duanmu, Z.; Zhu, H.; Fang, Y.; and Wang, Z. 2019b. Deep guided learning for fast multi-exposure image fusion. *IEEE Transactions on Image Processing*, 29: 2808–2819.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3.

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.; 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.

Vs, V.; Valanarasu, J. M. J.; Oza, P.; and Patel, V. M. 2022. Image fusion transformer. In *2022 IEEE International conference on image processing (ICIP)*, 3566–3570. IEEE.

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, H.; Xu, H.; Tian, X.; Jiang, J.; and Ma, J. 2021. Image fusion meets deep learning: A survey and perspective. *Information Fusion*, 76: 323–336.

Zhang, J.; Liao, Q.; Liu, S.; Ma, H.; Yang, W.; and Xue, J.-H. 2020a. Real-MFF: A large realistic multi-focus image dataset with ground truth. *Pattern Recognition Letters*, 138: 370–377.

Zhang, X. 2021. Deep learning-based multi-focus image fusion: A survey and a comparative study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9): 4819–4838.

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

Zhao, W.; Xie, S.; Zhao, F.; He, Y.; and Lu, H. 2023a. Metafusion: Infrared and visible image fusion via meta-feature embedding from object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13955–13965.

Zhao, Z.; Bai, H.; Zhang, J.; Zhang, Y.; Xu, S.; Lin, Z.; Timofte, R.; and Van Gool, L. 2023b. 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.; Zhu, Y.; Zhang, J.; Xu, S.; Zhang, Y.; Zhang, K.; Meng, D.; Timofte, R.; and Van Gool, L. 2023c. DDFM: denoising diffusion model for multi-modality image fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8082–8093.