Neural Style Transfer based Infrared-Visible Fusion

Lokesh Kumar KM
Department of ECE
Amrita School of Engineering
Amrita Vishwa Vidyapeetham
Chennai, India
ch.en.u4cce21029@ch.students.amrita.edu

Aishwarya N*
Department of ECE
Amrita School of Engineering
Amrita Vishwa Vidyapeetham
Chennai, India
n_aishwarya@ch.amrita.edu

Ashwin B
Department of ECE
Amrita School of Engineering
Amrita Vishwa Vidyapeetham
Chennai, India
ch.en.u4cce21004@ch.students.amrita.edu

***Abstract*- Infrared-Visible image fusion is a method of integrating information from images captured in separate spectra, including infrared (IR) and visible light. This technique aims to create a single composite image that takes advantage of the strengths of the original image while minimizing their inherent limitations. This paper introduces a Neural Style Transfer based non-End-to-End framework for seamlessly merging infrared and visible images. Our approach entails an optimization process where fused features interplay with the initial composite image. Then, the vital features are extracted from input images using the first four layers of the ResNet50 network. These features subsequently unite through an appropriate fusion rule. The original images are blended using the average rule to formulate the initial composite image. By employing backpropagation, the final synthesized image emerges as the initial composite image is fine-tuned with the imbibed features. We have validated the efficacy of our fusion framework by conducting experiments on the TNO Image Fusion dataset. The outcomes of these experiments clearly demonstrate that our approach outperforms currently approaches, as evident from improvements in both subjective and objective assessments.**

***Keywords*—Deep Learning, Sensor Fusion, Optimization, Resnet50, Feature Extraction, Neural Style transfer**

I. INTRODUCTION

In recent times, research on multi-sensory images has seen a notable increase, driven by their diverse range of applications, including Military surveillance, Restricted area monitoring, concealed weapon detection, and Forest Fire detection. The objective of integrating multi-sensory images is to amalgamate crucial details from these input images into a unified composite image [1, 2]. It is worth highlighting that visible images contribute insights into color gradients, while Infrared images provide texture-related information to the resulting composite image. Visible images, being sensitive to environmental changes, primarily focus on capturing the background. Conversely, Infrared images excel at detecting thermal radiations and distinguishing targets from the background. Consequently, the final image retains essential details about both the subject and its surroundings.

Over the past few decades, fusion algorithms predominantly utilized two main strategies – multi-scale techniques and sparse as well as low-rank representation methods. In the former technique, the input images are broken down into various scales to extract features at different levels. These features are then combined using appropriate fusion strategies such as maximum, minimum, or average rules. The final fused image is reconstructed using an inverse process [3]. Conversely, in the latter approaches, sparse representation techniques have been employed to extract similar and dissimilar features are extracted from the source images [4-6]. Low-rank representation methods, on the other hand, divide source images into smaller patches and utilize histogram of gradients (HOG) features to classify these image patches [7-9]. Even though, SR, LRR and other traditional fusion algorithms yield high accuracy, they suffer from some critical challenges. Initially, the accuracy of the traditional fusion algorithms degrades with the increase in complexity of the source images. Additionally, the runtime of these algorithms is significantly impacted by the operator chosen for implementation. It's important to highlight that both feature extraction and fusion strategies are carried out manually, contributing to increased complexity and time-consuming nature of fusion methods.

With the advent of Automatic Feature Extraction powered by Deep Learning techniques, a multitude of image fusion algorithms have been introduced [10, 11]. These algorithms can be categorized into two groups: those where both feature extraction and fusion are constructed using deep learning techniques, and those where only the feature extraction process is implemented through deep learning techniques. In the latter case, the algorithm continues to rely on traditional fusion strategies.

In this paper, a multi-sensory fusion algorithm that leverages a deep learning technique known as Neural Style Transfer (NST) is proposed. Within NST, features are extracted from the source images employing a pre-trained model, ResNet50. The max fusion rule is then used to fuse these extracted features. To create the initial synthesized image, the source images are also merged with the average rule. An optimization method that combines the fused features with the original synthesized image is necessary for the construction of the final synthesized image. The Backpropagation Algorithm is employed in this procedure to iteratively change the pixels of the initial fused image in accordance with the fused features. Notably, this proposed approach does away with the use of traditional fusion techniques.

II. RELATED WORK

In this section, we will introduce several image fusion algorithms that are based on CNNs, GANs, Autoencoders, and Transformers.

a) CNN based algorithms

An algorithm based on CNN was introduced in 2017 by authors in [12] for multi-focus image fusion. Their algorithm's fundamental idea relies around creating a score map that collects the focus data from the input images. The synthesized image is then created using conventional fusion techniques on this score map. They also unveiled a CNN-based image fusion approach designed specifically for medical picture fusion the same year [13]. In this method, CNN is employed to extract features from the source images, and those features are then utilized to create a weight map. Once again, the algorithm utilizes traditional fusion strategies applied to the obtained weight map. A significant challenge encountered when employing CNNs for image reconstruction is pan sharpening.

*Corresponding author

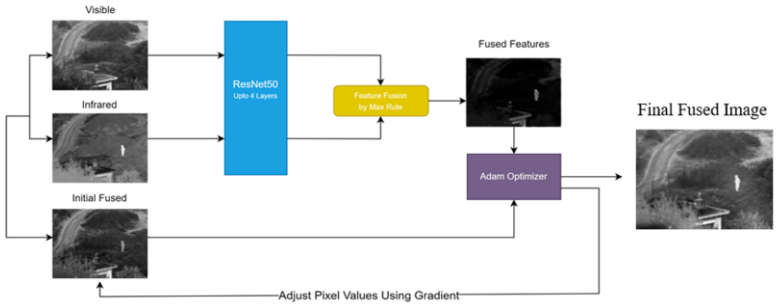

Fig. 1 Block Diagram of Proposed Algorithm

To address this issue, the authors in [14] put forth a three-level CNN architecture that effectively resolves the pan sharpening problem. Additionally, in [15], the authors proposed the DeepFuse technique, which employs CNNs with tied weights to enable the model to learn the same features from the source images. Furthermore, this technique employs CNN blocks to fuse the learned features.

b) GAN based algorithms

GAN's based image fusion is an adversarial process between the visible and infrared source images and the fused image, where the discriminator forces the Generator to generate fused image which cannot be differentiated. The first GAN-based image fusion technique, called FusionGAN, was presented by the authors of [16] in 2018, for the integration of multi-sensory images. To strengthen the distribution of gradient information in the synthesized image, this method applied adversarial learning. In order to amplify both gradient and non-gradient information within the synthesized image, they added an extra Discriminator to their prior work in 2019 [17]. In 2020, GANFuse was put forth in [18], introducing gradient loss and pixel intensity loss to the DDcGAN architecture. These additions further narrow down the probability distribution between the source images and the fused image.

c) Autoencoder based algorithms

An image fusion technique for multi-sensory pictures using NSCT (Non-Subsampled Contourlet Transform) and stacked sparse autoencoders was published by the authors in [19]. Moving to 2019, a convolutional autoencoder-based multi-spectral image fusion algorithm was introduced. This architecture, composed of an encoder-decoder model, aimed to extract high-frequency information from the source images and then reconstruct these images, which were eventually fused using simple concatenation [20]. In 2020, the DIDFuse technique was proposed [21], featuring an Autoencoder model that served the dual purpose of both feature extraction and feature fusion. Transitioning to 2022, FusionVAE was introduced in [22] for RGB image fusion. This deep hierarchical Variational Encoder performs image-based data fusion. It's worth noting that the model requires prior knowledge about the dataset being used. Within the same year, an end-to-end autoencoder-based algorithm for mult-sensor image fusion was introduced by Yan et al. [23]. This approach incorporates a residual fusion module designed for the multiscale extraction and fusion of features from the source images. It is obvious to infer that Deep Learning-based image fusion algorithms are widely categorized into two categories: Deep learning for feature extraction and fusion and Deep learning for feature extraction and classical fusion procedures. The first group suffers from high computational time and power requirements with ground truth requirements, whilst the second category fails to overcome the dependence of classic image fusion techniques. The suggested strategy is a Non - End to End Framework that does not require any training or ground truth and is thus independent of typical fusion methodologies.

III. PROPOSED FUSION APPROACH

In this section, the design and architecture of the feature extraction model are provided as depicted in Fig. 1, the formulation of fusion strategy and the definition of Loss function. The proposed algorithm is designed based on a deep learning technique called Neural style transfer (NST). NST is a type of Generative AI which tends to incorporate the style of an image into another. In our proposed algorithm we tend to incorporate the style/features of the fused features into the initial synthesized image. The initial synthesized image $f_{initial}$ is generated by elementwise average between the source images f_{IR} and f_{VIS}.

$$f_{initial}(x, y) = mean(f_{IR}, f_{VIS}) \quad (1)$$

The task of transferring the features of the fused features to the initial synthesized image is realized as an Optimization problem which involved backpropagation algorithm powered by ADAM optimizer.

A. Feature Extraction

In the proposed method, ResNet50, a pretrained neural network with predetermined weights, is used to extract key characteristics from the input multi-sensory images. ResNet50 is a 50-layer convolutional neural network (CNN) used for feature extraction [11]. In our algorithm, we exclusively utilize the first four layers to capture the significant features from the source images. These source images are resized to (256, 256) RGB images, serving as input for the feature extraction model. Specifically, the first four layers of the ResNet50 model are employed, collectively referred to as the CONV1 block. This block comprises an Input layer with an input shape of (256, 256, 3), a zero-padding layer, a Conv2D layer housing 64 filters of a 7x7 kernel size, a Batch Normalization layer, and a ReLU Activation Layer. The output of this feature extraction block

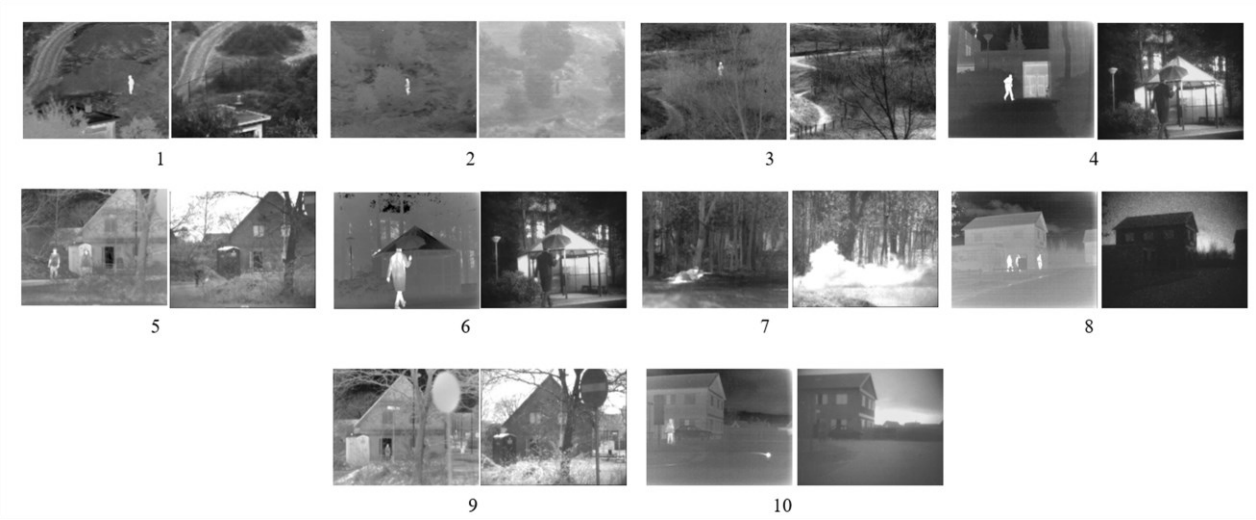

Fig. 2 Infrared-Visible Image Pairs

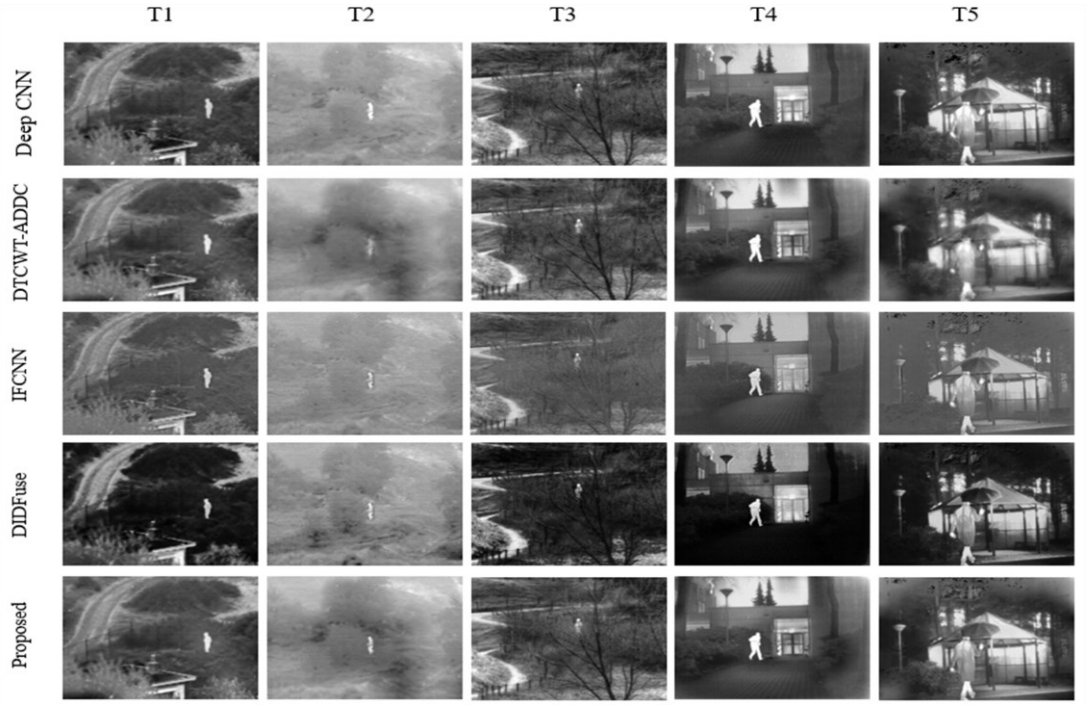

Fig. 3 Qualitative Synthesized Images from TNO Dataset. Fusion results of (a) Deep CNN (b)DTCWT-ACCD (c) IFCNN (d) DIDFuse (e) Proposed NST method

Table 1 Average statistical measurements of the conventional methods and NST method

Methods		$Q^{AB/F}$	Q_{CB}	Q_w	Q	Q_E	FF	H	MG	SD	SF
DTCWT-ACCD	[25]	0.5112	0.5664	0.8103	0.7839	0.6518	3.5258	**7.1753**	7.9118	40.7679	11.9829
IFCNN	[26]	0.46	0.3952	0.5286	0.6197	0.2994	1.4664	6.7142	8.3856	32.7332	12.4299
Deep CNN	[10]	0.5518	0.5395	0.8005	0.8039	0.6715	1.9577	7.0819	7.484	44.1813	11.3715
DIDFuse	[21]	0.3035	0.3436	0.4474	0.5292	0.2933	1.9957	6.4544	7.897	**48.6806**	12.4472
NST		**0.5686**	**0.5739**	**0.8107**	**0.8121**	**0.6873**	**4.0574**	**7.1461**	**8.7118**	**48.6745**	**12.8363**

assumes a shape of (128, 128, 64). Consequently, this block generates 64 significant feature maps, which are subsequently fused using a simple elementwise sum operation. By the end of this process, two feature maps are obtained, each with a shape of (128, 128), corresponding to the Infrared and Visible images, respectively.

$$F_{IR} = \sum_{64} conv(f_{IR}) \quad (2)$$

$$F_{VIS} = \sum_{64} conv(f_{VIS}) \quad (3)$$

Where F_{IR} and F_{VIS} are the features maps of multi-sensory images respectively.

B. Feature Fusion

The characteristics present in multi-sensor images are integrated through the utilization of the maximum fusion technique. The selection of the max fusion strategy is aimed at safeguarding the retention of essential target information during the fusion process. It is given by,

$$F_{fused} = \max(\ F_{IR}(x,y), F_{VIS}(x,y)\ ) \quad (4)$$

Where F_{fused} is the fused features to be incorporated into initial synthesized image.

C. Loss Function and Backpropagation

The loss between the corresponding pixels of the fused features and the initial synthesized image is calculated using the NST method. The Squared Error is the loss function that is being used. Gradients are calculated to change the pixel values of the initial synthesized image in a way that minimizes the loss using this computed loss. Notably, in this context, backpropagation is applied to modify the pixel values, replacing the traditional weights adjustment in a neural network. To facilitate the backpropagation process, the Adaptive Moment Estimation optimizer (ADAM) is utilized. This optimizer aids in achieving faster convergence and exhibits adaptability with respect to the momentum of the gradient descent.

$$Loss = \sum_x \sum_y (F_{fused}(x,y) - f_{initial}(x,y))^2 \quad (5)$$

D. ADAM optimizer

Adam optimizer [24] is used in our method to adjust the pixel values of $f_{initial}$ using backpropagation. Adam provides us dual advantage of adaptive learning and momentum-based learning which helps in less memory usage and convergence time. In the NST algorithm, the pixel values of the images are updated by the rule given in Eq. (6).

$$p_{updated} = p_c - m_c^h \left[\frac{\alpha}{(v_c^h + \varepsilon)}\right] \quad (6)$$

Where $p_{updated}$ and p_c are the updated and current pixel values respectively, m^h_c and v^h_c are the bias corrected weights parameters. In the Adam optimizer, m^h_c provides momentum to the learning, wherea v^h_c provides adaptive learning to the model.

IV. Experiments and Results

A. Experiment Setup

As shown in Fig. 2, a set of ten pairs of multi-sensory images is used to test the NST approach. These pictures are collected in both daylight and nighttime settings, guaranteeing the algorithm's adaptability to a range of luminosities. To facilitate feature extraction, all images are resized to a uniform shape of 256x256 with 3 channels. For a comprehensive qualitative analysis, the multi-sensory image pairs are tested against four image fusion algorithms which include Deep CNN [10], DTCWT-ACCD [25], IFCNN [26], and DIDFuse [21]. These algorithms are executed with their default parameters. In the context of our proposed algorithm, we employ a pretrained ResNet50 model with default hyperparameters. The optimization of hyperparameters, such as the learning rate (α) and exponential moving average (ema) of the ADAM optimizer, is performed. Specifically, the learning rate is set to 0.01, and the ema parameter is fixed as True. These optimized hyperparameters contribute to efficient loss function minimization with reduced time and memory consumption.

B. Evaluation Metrics

Ten statistically evaluation metrics such as Chen-blum metric Q_{CB} [27], edge based fusion metric $Q^{AB/F}$ [28], Piella's metrics (Q_w, Q, Q_e)[29], fusion factor *FF*, entropy *H,* mean gradient *MG*, standard deviation *SD*, spatial frequency *SF* [30] are evaluated to test the efficacy of the proposed method.

C. Results

Fig. 3 displays the fused images generated by Deep-CNN, DTCWT-ACCD, IFCNN, DIDGAN, and IVFNST. When compared to NST, the Deep-CNN in Fig. 3 (T1 - T5) collects both gradient and texture information but fails to enhance the target, as seen in Fig. 3(T1). In DTCWT-ACCD, the target region experiences blurring due to visible artifacts, noticeable in Fig. 3(T2). Moreover, visible artifacts are observed in the background of Fig. 3 (T1-T5 of DTCWT-ACCD). IFCNN struggles to capture gradient information, resulting in fused images influenced by texture information from the infrared image. While the fused images from DIDGAN offer clear target information, they struggle to accurately extract proper gradient information, as visible in Fig. 3(T1, T3, T4 of DIDGAN). In contrast, the fusion results from NST exhibit enhanced target regions with a well-balanced mix of gradient and texture information, as depicted in Fig. 3 (T1-T5). The statistical evaluations of the conventional methods and the NST approach are presented in Table 1, with the optimal values highlighted in blue. As observed from Table 1, the NST method outperforms in terms of most metrics, excluding *H* and *SD*. DTCWT-ACCD achieves the highest *H* value, followed by the proposed algorithm, whereas DIDGAN exhibits the highest *SD* value, trailed by the proposed algorithm. It's important to note that the disparities in *H* and *SD* metrics are not substantial, and the proposed method excels in effectively capturing gradient and texture information from both visible and infrared images. In conclusion, based on both qualitative and

objective comparisons, it's evident that the proposed NST algorithm outperforms other state-of-the-art fusion methods.

V. CONCLUSION

This paper presents an innovative deep learning-oriented algorithm, referred to as NST, designed for fusing infrared and visible images. In contrast to traditional deep learning-based image fusion algorithms, which often grapple with challenges such as the requirement for ground truth, intensive computational resources, and reliance on conventional fusion strategies, the proposed algorithm adeptly addresses these obstacles. Through comprehensive statistical and objective analyses involving ten evaluation metrics and assessments with different fusion algorithms, our research showcases the efficacy of the NST method in generating fused images that adeptly preserve essential features from both the multi-sensory images. While the proposed algorithm demonstrates superior performance when compared to other state-of-the-art algorithms in both subjective and objective evaluations, there remains room for enhancement in terms of increasing the prominence of the visible image within the fused image. This represents a potential avenue for future research in this domain.

REFERENCES

[1] Ren, Long, Zhibin Pan, Jianzhong Cao, Jiawen Liao, and Yang Wang. "Infrared and visible image fusion based on weighted variance guided filter and image contrast enhancement." *Infrared Physics & Technology* 114 (2021): 103662.

[2] Dogra, Ayush, Bhawna Goyal, and Sunil Agrawal. "From multi-scale decomposition to non-multi-scale decomposition methods: a comprehensive survey of image fusion techniques and its applications." *IEEE access* 5 (2017): 16040-16067.

[3] Aishwarya, N., Y. Asnath Victy Phamila, and R. Amutha. "Multi-focus image fusion using multi-structure top-hat transform and image variance." *In 2013 International Conference on Communication and Signal Processing*, pp. 352-356. IEEE, 2013.

[4] Yang, Bin, and Shutao Li. "Multifocus image fusion and restoration with sparse representation." *IEEE transactions on Instrumentation and Measurement* 59, no. 4 (2009): 884-892.

[5] N. Yu, T. Qiu, F. Bi and A. Wang, "Image Features Extraction and Fusion Based on Joint Sparse Representation," in *IEEE Journal of Selected Topics in Signal Processing,* vol. 5, no. 5, pp. 1074-1082, Sept. 2011, doi: 10.1109/JSTSP.2011.2112332.

[6] Yin, H., Li, Y., Chai, Y., Liu, Z. and Zhu, Z., 2016. A novel sparse-representation-based multi-focus image fusion approach. *Neurocomputing*, *216*, pp.216-229.

[7] Li, H. and Wu, X.J., 2017. Multi-focus image fusion using dictionary learning and low-rank representation. *In Image and Graphics: 9th International Conference, ICIG 2017*, Shanghai, China, September 13-15, 2017, Revised Selected Papers, Part I 9 (pp. 675-686). Springer International Publishing.

[8] Li, Hui, and Xiao-Jun Wu. "Infrared and visible image fusion using latent low-rank representation." arXiv preprint arXiv:1804.08992 (2018).

[9] S. Yu and X. Chen, "Infrared and Visible Image Fusion Based on a Latent Low-Rank Representation Nested With Multiscale Geometric Transform," in *IEEE Access*, vol. 8, pp. 110214-110226, 2020.

[10] Yu Liu, Xun Chen, Juan Cheng, Hu Peng, Zengfu Wang, "Infrared and visible image fusion with convolutional neural networks," *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 16, no. 3, 1850018: 1-20, 2018.

[11] He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

[12] Liu, Y., Chen, X., Peng, H. and Wang, Z., 2017. Multi-focus image fusion with a deep convolutional neural network. *Information Fusion*, *36*, pp.191-207.

[13] Y. Liu, X. Chen, J. Cheng and H. Peng, "A medical image fusion method based on convolutional neural networks," *2017 20th International Conference on Information Fusion (Fusion)*, Xi'an, China, 2017.

[14] Masi, G., Cozzolino, D., Verdoliva, L. and Scarpa, G., 2016. Pansharpening by convolutional neural networks. *Remote Sensing*, *8*(7), p.594.

[15] 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* (pp. 4714-4722).

[16] Ma, J., Yu, W., Liang, P., Li, C. and Jiang, J., 2019. FusionGAN: A generative adversarial network for infrared and visible image fusion. *Information fusion*, *48*, pp.11-26.

[17] J. Ma, H. Xu, J. Jiang, X. Mei and X. -P. Zhang, "DDcGAN: A Dual-Discriminator Conditional Generative Adversarial Network for Multi-Resolution Image Fusion," in *IEEE Transactions on Image Processing*, vol. 29, pp. 4980-4995, 2020, doi: 10.1109/TIP.2020.2977573.

[18] Yang, Z., Chen, Y., Le, Z. and Ma, Y., 2021. GANFuse: a novel multi-exposure image fusion method based on generative adversarial networks. *Neural Computing and Applications*, *33*, pp.6133-6145.

[19] Luo, X., Li, X., Wang, P., Qi, S., Guan, J. and Zhang, Z., 2018. Infrared and visible image fusion based on NSCT and stacked sparse autoencoders. *Multimedia Tools and Applications*, *77*, pp.22407-22431.

[20] A. Azarang, H. E. Manoochehri and N. Kehtarnavaz, "Convolutional Autoencoder-Based Multispectral Image Fusion," in *IEEE Access*, vol. 7, pp. 35673-35683, 2019, doi: 10.1109/ACCESS.2019.2905511.

[21] Zhao, Z., Xu, S., Zhang, C., Liu, J., Li, P. and Zhang, J., 2020. DIDFuse: Deep image decomposition for infrared and visible image fusion. *arXiv preprint arXiv:2003.09210*.

[22] Duffhauss, F., Vien, N.A., Ziesche, H. and Neumann, G., 2022, October. FusionVAE: A Deep Hierarchical Variational Autoencoder for RGB Image Fusion. In *European Conference on Computer Vision* (pp. 674-691). Cham: Springer Nature Switzerland.

[23] Liu, H. and Yan, H., 2023. An end-to-end multi-scale network based on autoencoder for infrared and visible image fusion. *Multimedia Tools and Applications*, *82*(13), pp.20139-20156.

[24] Kingma, D.P. and Ba, J., 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[25] Aishwarya, N. and Thangammal, C.B., 2018. Visible and infrared image fusion using DTCWT and adaptive combined clustered dictionary. *Infrared Physics & Technology*, *93*, pp.300-309.

[26] 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*, pp.99-118.

[27] Qu, G., Zhang, D. and Yan, P., 2002. Information measure for performance of image fusion. *Electronics letters*, *38*(7), p.1.

[28] Yu Liu, Xun Chen, Juan Cheng, Hu Peng, Zengfu Wang, "Infrared and visible image fusion with convolutional neural networks," *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 16, no. 3, 1850018: 1-20, 2018.

[29] Piella, G. and Heijmans, H., 2003, September. A new quality metric for image fusion. In *Proceedings 2003 international conference on image processing (Cat. No. 03CH37429)* (Vol. 3, pp. III-173). IEEE.

[30] Aishwarya, N., C. Bennila Thangammal, and Nalamani G. Praveena. "NSCT and focus measure optimization based multi-focus image fusion." *Journal of Intelligent & Fuzzy Systems* 41, no. 1 (2021): 903-915.