A Light Gradient Residual Encoder-Decoder Network for Multimodal Image Fusion

Muhammad Ishfaq Hussain*†, Zubia Naz*, Linh Van Ma*, Jeonghwan Gwak‡ and Moongu Jeon*
*School of Electrical Engineering and Computer Sciene, Gwangju Institute of Science and Technology (GIST), South Korea,
†Department of Software, Korea National University of Transportation, South Korea
‡Large-scale AI Research Group, Korea Institute of Science and Technology Information (KISTI), South Korea
Email: (ishfaqhussain, mgjeon)@gist.ac.kr, (linh.mavan, zubianaz)@gm.gist.ac.kr, jgwak@ut.ac.kr

***Abstract*—Image fusion combines the complementary traits of source images into a single output, enhancing both human visual observation and machine vision perception. The existing fusion algorithms typically prioritize visual enhancement, often overlooking the real-time needs for critical surveillance applications. To address these real-time deployment needs, we present a compact fusion network for combining infrared and visible image representations, named Light-weight Fusion (LightFusion). This network employs incremental semantic integration and scene recognition accuracy constraints by incorporating three different bands of images (IR, RGB, and Grayscale) to fuse the data. Our approach includes a sparse semantic perception branch that captures critical semantic features, which are then integrated into the fusion network through a semantic injection module. This ensures that high-level vision tasks are adequately addressed. The scene fidelity path ensures that fusion features preserve all details required to reconstruct the original images. The importance and applicability of the proposed network are enhanced by employing an extra input in the form of a grayscale image, obtained by converting the RGB image for improved contrast, along with prominent target masks to enhance the visual quality of the fusion results. Our extensive analysis shows that the lightweight LightFusion network outperforms existing methods in both visual quality and semantic integrity, even under challenging conditions. The source code will be released at https://github.com/MI-Hussain/LightFusion.**

***Index Terms*—RGB, IR, Fusion, Neural Network**

I. Introduction

With the evolution of diverse sensors, multimodal images have become increasingly common across various application domains. Visible (RGB) and infrared (IR) sensors, in particular, are extensively used due to their complementary imaging properties [1], [2], [3]. Visible images capture rich texture and color information but are heavily affected by low-light environments [4]. Conversely, IR images convey thermal information and can clearly delineate objects in poor lighting conditions, albeit lacking in detailed texture [5], [6]. Recent studies have concentrated on the feature-level fusion of visible and IR images to improve performance in downstream tasks such as object detection [2]. Conventional RGB-IR object detection techniques often involve the addition or concatenation of modality-specific features from RGB and IR images [7]. However, this "Late fusion" strategy is limited in effectively combining complementary information, leading to subpar performance. Alternative "Halfway fusion" strategies incorporate interaction modules between different modality features to enhance fusion, yet they still face challenges with modality noise and fail to achieve fully complementary fusion [7], [8], [9]. The cognitive theories such as Treisman's "Attenuation Theory," which describes a coarse-to-fine process of filtering out extraneous information [10], authors developed an approach starting with the Redundant Spectrum Removal (RSR) module to filter irrelevant information in the frequency domain, followed by a fine selection of features for fusion [10]. [5] introduced a semantic fusion framework called SeAFusion [5], which incorporates a segmentation model to enrich semantic information in fused images as a possible solution. However, such methods may limit the applicability of fused images to other models and may not perform well under extreme conditions. Feature-level fusion, which directly processes multi-modal fusion features without generating a fused image, has emerged as a prominent solution for advanced vision tasks. This approach utilizes feature extraction networks to capture semantic features from source images, followed by dedicated fusion modules to integrate complementary representations. Nonetheless, existing feature-level fusion methods [11] are often designed for specific tasks and require significant redesign when applied to new backbones like Transformer and ConvNeXt. Furthermore, these methods do not fully explore the potential of image-level fusion compared to feature-level fusion [11], creating a gap in the development of effective fusion strategies. To address these limitations, we propose a lightweight neural network sensor fusion network that efficiently combines data, leveraging the complementary strengths of visible and IR imaging modalities. This approach aims to enhance performance in high-level vision tasks while maintaining efficient computational requirements.

The major points of the work are highlighted below:

- We propose a lightweight network that utilizes a gradient residual encoder-decoder network to fuse sensor data from three separate modalities.
- The model was trained on the M3FD dataset [1] and tested on the TNO [12], MSRS [2], and Camel [13] datasets for a fair comparison to validate the authenticity of the proposed work.

979-8-3503-5469-0/24/$31.00 ©2024 IEEE

II. Related Work

Traditional image fusion methods emphasize feature extraction and merging through techniques like multi-scale transforms (e.g., Laplacian pyramid, discrete wavelet), sparse representation, and subspace-based methods such as principal component analysis (PCA) [14]. Optimization-based approaches and hybrid models combining various frameworks are also explored. Auto-encoder-based methods train networks to extract and reconstruct features using convolutional layers and dense blocks, employing different strategies to merge high-level features. Convolutional neural networks (CNNs) perform implicit feature extraction, aggregation, and image reconstruction [15], integrating fusion layers within the training process to optimize hand-crafted fusion procedures. Generative adversarial networks (GANs) [16] use adversarial loss to create fused images with rich textures, aiming to enhance detail information and sharpen edges, though they may face challenges like mode collapse. The fusion module is essential for detecting objects using multi-modality sensors. This section reviews previous learning-based IR and visible image fusion (IVIF) approaches [1], [2] and relevant benchmarks needed for learning and empirical evaluation. Deep learning has made significant progress in low-level vision tasks due to the powerful nonlinear fitting abilities of multi-layer neural networks. Initial efforts integrated deep networks into the IVIF process for feature extraction or weight generation, such as Liu et al. [1], who cascaded two pre-trained CNNs for feature and weight learning. End-to-end architectures have also been developed to produce fused images in one step, like the residual fusion network by [17], which learns enhanced features in a common space for structure-consistent results. Recently, IVIF approaches using GANs have shown promising results by transforming different distributions into the desired one. For instance, [18] introduced an adversarial game between fused and visible images to enhance texture details, although this method can miss critical IR information. Despite these advances, current methods struggle to capture the unique characteristics of different imaging types, highlighting the need for further research. Several benchmarks, such as the TNO Image Fusion, INO Videos Analytics, OSU Color-Thermal, RoadScene, and Multispectral datasets [1], [12], [13], have been developed to support IVIF research, each offering unique scenarios and challenges for evaluating fusion and detection tasks.

III. Proposed Methodology

In this work, we propose a novel approach called LightFusion, a simple and lightweight encoder-decoder network designed to enhance in-depth feature extraction by incorporating extra attention through triple-band input feeding. The unique and salient feature of LightFusion that differentiates it from previous work is its ability to independently process three different bands IR, RGB, and grayscale. Each band is fed separately into distinct encoder modules within the network, allowing for independent and detailed feature extraction from each input source. The proposed network diagram can be seen in Fig. 1.

A. Network Architecture

1) Triple-Band Input Feeding: The network architecture is composed of triple band input feeding from IR, RGB, and Grayscale from RGB. Each of these bands is processed by separate encoder modules. This independent processing enables the network to extract unique and in-depth features from each band independently. The encoders leverage gradient-based recurrent neural networks (RNNs) to enhance the feature extraction capabilities further.

2) Encoder Module: Each encoder processes its respective input band and extracts salient features independently. The use of RNNs (Light-GRLB) within the encoders helps in capturing temporal dependencies and enhances feature extraction. The light gradient residual block is composed of two blocks as shown in Fig. 2.

3) Fusion Layer: Once the encoding process is complete, the extracted features from the three bands are concatenated in the fusion layer. This fusion step combines the rich and diverse information from the IR, RGB, and grayscale bands, providing a comprehensive feature set for subsequent processing.

4) Decoder Module: The fused features are then processed by the decoder module. The decoder reconstructs the image by retaining the original information while integrating the additional insights provided by the IR band. This reconstruction ensures that the final output is more informative and readable compared to standalone sensor data.

B. Advantages of LightFusion

1) Late Fusion Technique: The independent processing of each input band (late fusion) allows the network to learn more effectively compared to early fusion techniques, where inputs are combined before feature extraction. This method ensures that the unique features of each band are preserved and fully utilized.

2) Enhanced Feature Extraction: The use of gradient-based RNNs within the encoder modules enables the network to capture and utilize temporal dependencies, enhancing the depth and quality of the extracted features.

3) Improved Readability and Information Content: By integrating IR data, the final output image retains more information and is more readable, providing a significant advantage over using standalone sensor data. In comparison with Early Fusion, we also experimented with an early fusion approach, where the inputs were combined before being fed into a single encoder. However, the results were inferior compared to the late fusion approach utilized in LightFusion. The independent processing in the late fusion technique proved to be more effective in learning and preserving the unique features of each band.

IV. Experimental Setup / Experimentation

The experiments were performed with PyTorch 2.0.1+cu117 and Python 3.10.13, running on Ubuntu 20.04.6 LTS x86-64.

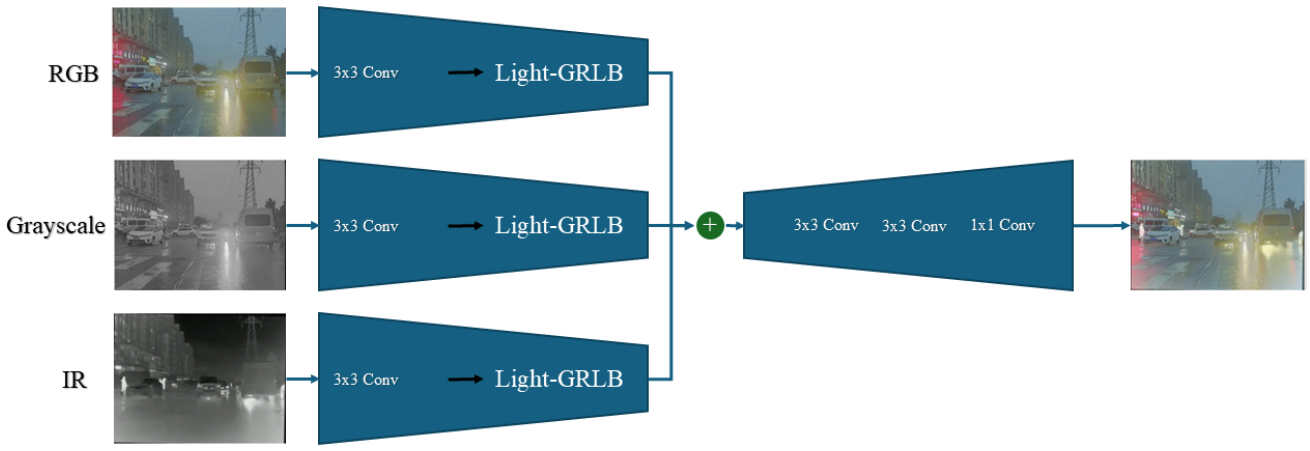

Fig. 1. The overall diagram of our proposed light gradient residual-based Encoder-decoder network. Input images used for fusion are RGB, GrayScale, and IR images. These images are passed through light gradient residuals based on an encoder-decoder network to output a fusion image.

It features an NVIDIA RTX 3090 GPU and an AMD Ryzen 5 5600x 6-Core Processor, providing substantial computational power for handling complex tasks. The image sizes processed by the system are 640x480 pixels, with a processing speed of 10.67 images per second. This setup ensures efficient and robust performance for various image processing and fusion tasks.

A. Datasets

The M3FD dataset [1] comprises images collected from three primary locations: the campus of Dalian University of Technology, the State Tourism Holiday Resort at Golden Stone Beach in Dalian, China, and the main roads in Jinzhou District, Dalian, China. It includes a total of 8,400 images intended for fusion, detection, and fusion-based detection, and 600 images from independent scenes for fusion. This translates to 4,200 image pairs for the first group of tasks and 300 pairs for the latter. The images are provided in two formats: 24-bit grayscale bitmaps for IR images and 24-bit color bitmaps for visible images, with most images sized at 1024 x 768 pixels. All image pairs are registered, with visible images calibrated using internal system parameters and IR images artificially distorted using a homography matrix. The dataset contains 34,407 manually labeled instances across six categories: People, Car, Bus, Motorcycle, Lamp, and Truck. It should be noted that some labels may be incorrect or missing due to manual labeling constraints, and feedback to improve the dataset is welcomed. For additional testing, we also utilized the TNO and Camel datasets as well for a fair comparison. The Multi-Spectral Road Scene (MSRS) dataset [2] comprises images captured in road scene environments, including both visible and IR spectral bands. This dataset is designed to support the development and evaluation of algorithms for autonomous driving and advanced driver-assistance systems (ADAS). The MSRS dataset includes a wide range of road conditions and environments, such as urban, rural, and high-way scenes, providing a comprehensive testbed for evaluating fusion techniques in automotive applications. The CAMEL (Composite Attention-based Multispectral and Hyperspectral Enhanced Learning) dataset [13] is designed for complex scene analysis, featuring both multispectral and hyperspectral images. This dataset contains images from various scenarios, including urban, rural, and natural environments, captured under different lighting and weather conditions. The CAMEL dataset is particularly valuable for research in advanced image processing techniques, such as hyperspectral image analysis, and for tasks that require detailed spectral information for accurate scene interpretation.

B. Pre-processing

The M3FD dataset, consisting of 4200 images, was used for the fusion task. The dataset was split into training and validation sets in an 80:20 ratio. Prior to feeding the images into the network, the image resolution was resized to the required dimensions (640x480). Additionally, RGB images were converted to gray-scale to serve as independent inputs to the network. The network architecture employed in this work utilizes a late fusion technique, therefore it accepts three distinct inputs separately. Besides that, before feeding to the network we need to make sure all the images have the same resolution.

C. Experimentation's

The proposed network was trained from scratch using 4200 images over 20 epochs. The performance of the network was subsequently evaluated on 300 fusion images from the M3FD dataset [1], TNO dataset [12], MSRS dataset [2], and Camel dataset [13]. We just trained on M3FD dataset where as we tested on the other dataset using the pre-trained weights for M3FD dataset. This rigorous training and testing protocol ensures a robust comparison of the efficacy and capabilities of the proposed networks in handling image fusion tasks. For better understanding and strengthening, the network is trained with early and late-fusion. During the network training, we utilized the Adam optimizer whereas the $tanh$ and $Leke{-}relu$ are used as activation functions. For the evaluation metrics, we

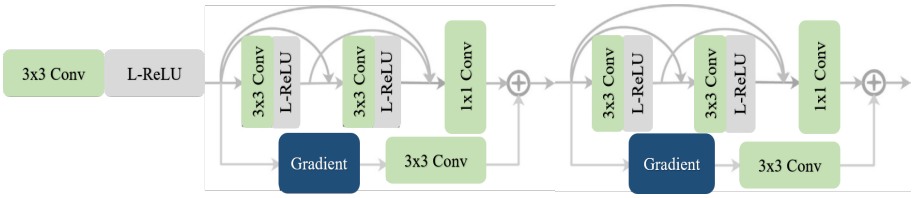

Fig. 2. An in-depth explanation of Gradient Residual-based Convolutional Layers (Light-GRLB), illustrating the integration of pointwise (1x1) convolutions to reduce computational complexity and make the network lightweight while preserving essential gradient features.

utilized concepts such as Mutual Information (MI), Entropy, Spatial Frequency (SF), Structural Similarity Index Measure (SSIM), Standard Deviation (SD), Qabf, and VIF. These metrics are essential in information theory and are used to quantify and transfer information between images, which is crucial for tasks such as image denoising, deblurring, and compression.

$$\text{Entropy} = -\sum_{i=1}^{n} p_i \log_2 p_i \tag{1}$$

$$MI(X;Y) = \sum_{x\in X, y\in Y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)} \tag{2}$$

$$F_{MI}(X,Y;F) = \sum_{x,y,f} p(x,y,f) \log_2 \frac{p(x,y)}{p(x)p(y)} \tag{3}$$

These equations define entropy as a measure of uncertainty or randomness in an image, mutual information as a measure of the amount of information transferred from one random variable to another, and an expression that relates these concepts within the context of fused images. Two statistical measures used in image processing are Standard Deviation (SD) and Spatial Frequency (SF). These measures are important for analyzing the brightness, contrast, and texture details of images.

$$SD = \sqrt{\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}(I(i,j)-\mu)^2} \tag{4}$$

This equation calculates the SD of an image, reflecting its brightness and contrast variations.

$$SF = \sqrt{SF_{row}^2 + SF_{col}^2} \tag{5}$$

$$SF_{row} = \sqrt{\frac{1}{M(N-1)}\sum_{i=1}^{M}\sum_{j=1}^{N-1}(I(i,j+1)-I(i,j))^2} \quad and \tag{6}$$

$$SF_{col} = \sqrt{\frac{1}{(M-1)N}\sum_{i=1}^{M-1}\sum_{j=1}^{N}(I(i+1,j)-I(i,j))^2}$$

The Structural Similarity Index Measure (SSIM) is a method used to assess the quality of images by comparing them to a reference image in terms of luminance, contrast, and structure. SSIM is particularly useful in image analysis for measuring the similarity between two images, which is important for tasks like image compression, transmission, and denoising. The following equation calculates the SSIM between two image patches (X_F) and (Y_F), considering the mean values (μ), standard deviations (σ), and constants (C_1) and (C_2) to stabilize the division with weak denominators.

$$\text{SSIM}XF, YF = \frac{(2\mu XF\mu_{YF} + C_1)(2\sigma_{XFYF} + C_2)}{(\mu_{XF}^2 + \mu_{YF}^2 + C_1)(\sigma_{XF}^2 + \sigma_{YF}^2 + C_2)} \quad and$$

$$\text{SSIM}X, VISUAL + IR = \frac{\sum^{N} n = 1\text{SSIM}_{XFn,YFn}}{N} \tag{7}$$

We utilize the semantic loss function in order to train the network. The semantic loss function for IR and RGB sensor fusion can be defined as follows:

$$\begin{aligned}\mathcal{L}_{\text{semantic}} = \alpha \sum_i \text{CrossEntropy}(L_{\text{RGB}}(x_i), y_i) \\ +\beta \sum_i \text{CrossEntropy}(L_{\text{IR}}(x_i), y_i) \\ +\gamma \sum_i \|L_{\text{RGB}}(x_i) - L_{\text{IR}}(x_i)\|^2\end{aligned} \tag{8}$$

where: $\mathcal{L}_{\text{semantic}}$ is the overall semantic loss. α, β, and γ are weights that balance the contributions of each loss term.$CrossEntropy(L_{\text{RGB}}(x_i), y_i)$ is the cross-entropy loss for the RGB image at sample i. CrossEntropy$(L_{\text{IR}}(x_i), y_i)$ is the cross-entropy loss for the IR image at sample i. $L_{\text{RGB}}(x_i)$ and $L_{\text{IR}}(x_i)$ are the predicted labels for the RGB and IR images at sample i, respectively. y_i is the ground truth label for sample i. $\|L_{\text{RGB}}(x_i) - L_{\text{IR}}(x_i)\|^2$ is the consistency loss ensuring that the predictions from the RGB and IR images are similar.

D. Experimental Results

The experimental results are calculated on all the datasets and compared with the existing methods for a fair comparison. The qualitative results are shown in Fig. 3. The quantitative results are compared with different datasets as well. The Table. I , II, III, IV-D given below highlighted the results on the M3FD datasets, MSRS Camel and TNO datasets. As we plan to deploy this fusion technique for surveillance, it is

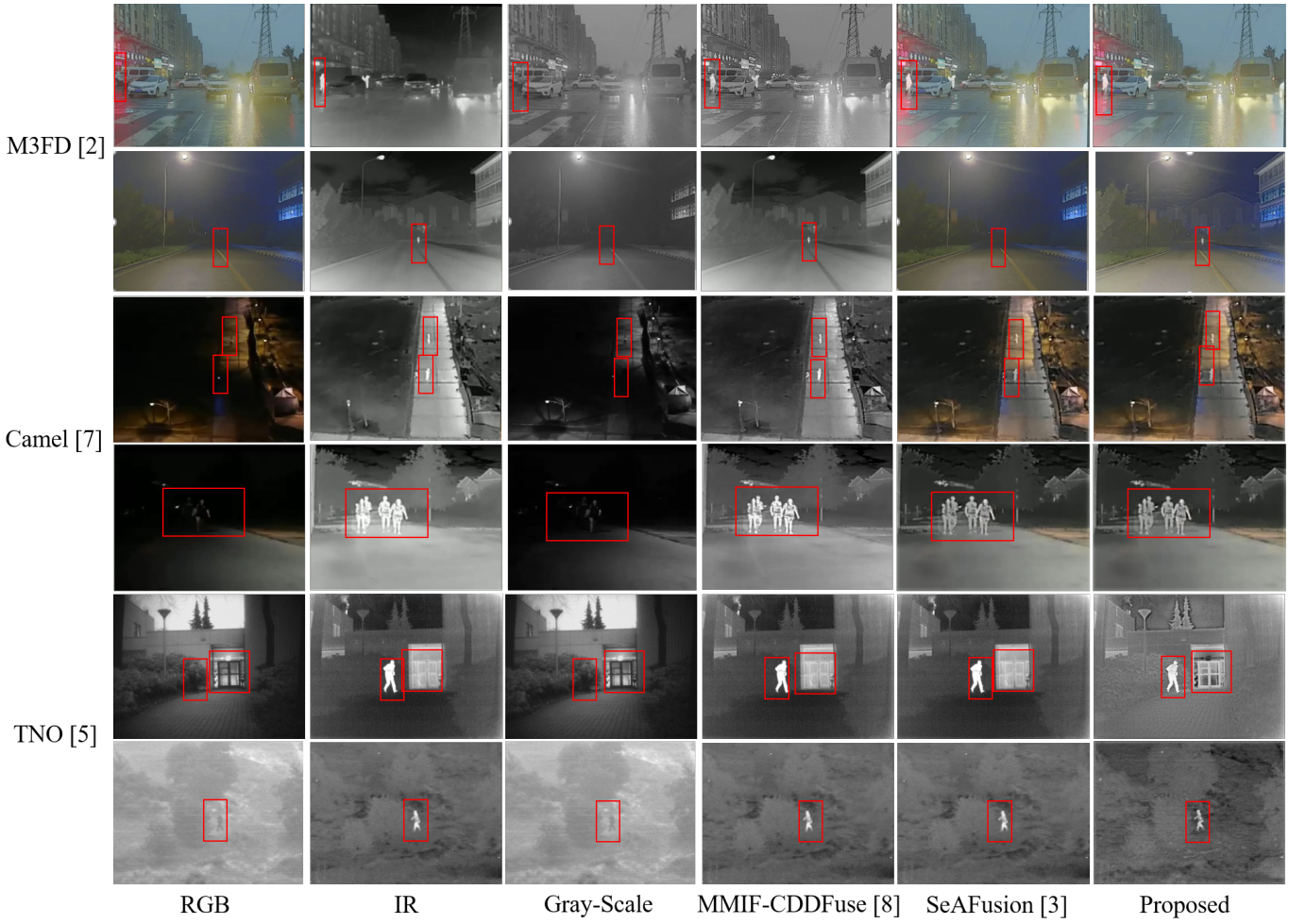

Fig. 3. The results of the proposed network on the Camel dataset for three different sequences are shown in this image. As illustrated by the TNO dataset, the proposed algorithm's results within the bounding box clearly reveal the window structure compared to other algorithms. Similarly, bounding boxes are drawn in other sequences to highlight the comparative analysis.

noteworthy that the FPS score is approximately 48 FPS on the M3FD dataset, which is significantly better than existing algorithms. Additionally, the algorithm was tested on the Camel dataset using three sequences (13th, 15th, and 30th). On this test set, we achieved approximately 90 FPS due to the lower resolution of the images in the dataset. As per the qualitative and quantitative results, the results are closer and even better than the existing algorithms.

TABLE I
QUANTITATIVE RESULTS ON M3FD DATASETS

Dataset: M3FD Infrared-Visible Fusion Dataset [1]							
Method	EN	SD	SF	MI	VIF	Qabf	SSIM
TarDal [1]	6.98	39.36	-	2.84	-	-	-
SeAFusion [5]	6.86	35.91	17.01	2.44	0.65	0.58	0.94
LightFusion	**7.01**	41.08	**18.61**	2.28	**1.79**	0.65	0.91

V. CONCLUSION

In conclusion, the Lightweight Fusion (LightFusion) network effectively addresses the shortcomings of existing fusion algorithms by integrating incremental semantic embeddings and scene recognition requirements, utilizing three different bands (IR, RGB, and Grayscale) images. Our novel approach, which includes a sparse contextual awareness branch and a semantic injection module, ensures that high-level vision

TABLE II
QUANTITATIVE RESULTS ON CAMEL DATASETS

Dataset: Camel dataset with sequence (13th / 15th & 30th) Infrared-Visible Fusion Dataset [13]							
Method	EN	SD	SF	MI	VIF	Qabf	SSIM
CDDFuse [19]	7.29	39.41	15.59	3.12	0.77	0.63	0.92
LightFusion	**7.41**	**48.45**	**19.92**	2.97	**1.51**	0.59	0.75

TABLE III
QUANTITATIVE RESULTS ON MSRS DATASET

Dataset: MSRS Infrared-Visible Fusion Dataset [2]							
Method	EN	SD	SF	MI	VIF	Qabf	SSIM
LightFusion	7.21	44.8	12.81	1.88	1.14	0.62	0.65

Dataset: TNO Infrared-Visible Fusion Dataset [12]							
Method	EN	SD	SF	MI	VIF	Qabf	SSIM
TarDal [1]	6.94	38.34	11.56	1.49	0.51	0.36	0.89
CDDFuse [19]	7.12	**46.00**	13.15	2.19	0.77	0.54	**1.03**
DeF [20]	6.95	38.41	8.21	1.78	0.60	0.42	0.97
DID [21]	6.97	45.12	12.59	1.70	0.60	0.40	0.81
SDN [22]	6.64	32.66	12.05	1.52	0.56	0.44	1.00
ReC [23]	7.10	40.83	13.15	2.19	0.77	0.54	1.03
RFN [24]	6.83	34.50	15.71	1.20	0.51	0.39	0.92
U2F [25]	6.83	34.55	10.57	1.37	0.47	0.31	0.81
LightFusion	**7.32**	44.54	**16.10**	**2.25**	0.77	**0.69**	0.94

tasks are adequately addressed while preserving all required information for the reconstruction of the original images. The introduction of an extra grayscale input, obtained by converting the RGB image, enhances contrast and salient target masks, further improving the visual quality of the fusion results. The LightFusion network was rigorously tested on four different datasets, and both qualitative and quantitative results demonstrate superior visual quality and semantic integrity compared to existing methods, even under challenging conditions. This significant advancement underscores the potential of LightFusion for a wide range of applications. In the future, Short-wave infrared(SWIR) and Mid-wave infrared (MWIR) modalities will be added to enhance the solution's robustness in challenging weather conditions.

ACKNOWLEDGEMENTS

This work was supported by Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Korea government (MCST) in 2024 (R2022060001), and GIST-MIT Research Collaboration Project.

REFERENCES

[1] J. Liu, X. Fan, Z. Huang, G. Wu, R. Liu, W. Zhong, and Z. Luo, "Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 5802–5811.

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

[3] H. Zhang, H. Xu, X. Tian, J. Jiang, and J. Ma, "Image fusion meets deep learning: A survey and perspective," *Information Fusion*, vol. 76, pp. 323–336, 2021.

[4] M. I. Hussain, S. Azam, M. A. Rafique, A. M. Sheri, and M. Jeon, "Drivable region estimation for self-driving vehicles using radar," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 6, pp. 5971–5982, 2022.

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

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

[7] J. Liu and Q. Zhang, "Multi-level modality-specific and modality-common features fusion network for RGB-IR person re-identification," *Neurocomputing*, vol. 600, p. 128183, 2024.

[8] M. I. Hussain, M. A. Rafique, S. Khurbaev, and M. Jeon, "Exploring data variance challenges in fusion of radar and camera for robotics and autonomous driving," in *2022 10th International Conference on Control, Mechatronics and Automation (ICCMA)*. IEEE, 2022, pp. 7–12.

[9] M. I. Hussain, M. A. Rafique, J. Kim, M. Jeon, and W. Pedrycz, "Artificial proprioceptive reflex warning using EMG in advanced driving assistance system," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 1635–1644, 2023.

[10] T. Zhao, M. Yuan, and X. Wei, "Removal and selection: Improving RGB-infrared object detection via coarse-to-fine fusion," *arXiv preprint arXiv:2401.10731*, 2024.

[11] K. R. Ranipa, W.-P. Zhu, and M. Swamy, "A novel feature-level fusion scheme with multimodal attention cnn for heart sound classification," *Computer Methods and Programs in Biomedicine*, vol. 248, p. 108122, 2024.

[12] A. Toet, "The TNO multiband image data collection," *Data in brief*, vol. 15, pp. 249–251, 2017.

[13] E. Gebhardt and M. Wolf, "CAMEL dataset for visual and thermal infrared multiple object detection and tracking," in *2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS)*. IEEE, 2018, pp. 1–6.

[14] H.-T. Hu and T.-T. Lee, "Robust complementary dual image watermarking in subbands derived from the laplacian pyramid, discrete wavelet transform, and directional filter bank," *Circuits, Systems, and Signal Processing*, vol. 41, no. 7, pp. 4090–4116, 2022.

[15] J. Schlemper, J. Caballero, J. V. Hajnal, A. N. Price, and D. Rueckert, "A deep cascade of convolutional neural networks for dynamic MR image reconstruction," *IEEE Transactions on Medical Imaging*, vol. 37, no. 2, pp. 491–503, 2017.

[16] Y. Fu, X. Wu, and T. S. Durrani, "Image fusion based on generative adversarial network consistent with perception," *Information Fusion*, vol. 72, pp. 110–125, 2021.

[17] P. Li, J. Chen, B. Lin, and X. Xu, "Residual spatial fusion network for RGB-thermal semantic segmentation," *Neurocomputing*, vol. 595, p. 127913, 2024.

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

[19] Z. Zhao, H. Bai, J. Zhang, Y. Zhang, S. Xu, Z. Lin, R. Timofte, and L. V. Gool, "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*, 2023, pp. 5906–5916.

[20] A. Liu, X. Liu, J. Fan, Y. Ma, A. Zhang, H. Xie, and D. Tao, "Perceptual-sensitive gan for generating adversarial patches," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 1028–1035.

[21] Z. Zhao, S. Xu, C. Zhang, J. Liu, P. Li, and J. Zhang, "DIDFuse: Deep image decomposition for infrared and visible image fusion," in *International Joint Conference on Artificial Intelligence*, 2020.

[22] H. Zhang and J. Ma, "SDNet: A versatile squeeze-and-decomposition network for real-time image fusion," *International Journal of Computer Vision*, vol. 129, no. 10, pp. 2761–2785, 2021.

[23] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, and W. Gao, "Pre-trained image processing transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 12 299–12 310.

[24] H. Xu, J. Ma, J. Yuan, Z. Le, and W. Liu, "RFNet: Unsupervised network for mutually reinforcing multi-modal image registration and fusion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 19 679–19 688.

[25] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2Fusion: A unified unsupervised image fusion network," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 1, pp. 502–518, 2020.