Grenze International Journal of Engineering and Technology, June Issue

Robust Multi-Modal Image Fusion under Noisy and Adverse Conditions – A Study

Subash Chandra R[1] and Siddesha S[2]
[1,2]Department of Computer Applications, JSS Science and Technology University, Mysuru, India
Email: subashchandra@jssstuniv.in

***Abstract*—Robust multi-modal Image fusion under noisy and adverse conditions is crucial for enhancing the reliability of real-world systems in fields such as surveil-lance, autonomous vehicles, and medical diagnostics. As the scale and diversity of sensors increase, scalability becomes an issue necessitating efficient fusion algorithms capable of managing high-dimensional data. Real-world environments are dynamic and require adaptive fusion methods for responding dynamically based on changing conditions. Handling sensor-specific noise and artifacts remains complex due to the distinct characteristics of each sensor type. This work highlights a study of major works reported in the literature with key challenges.**

***Keywords*— Image fusion, Multimodalities, Generative Adversarial Net-work, Transformers, Deep Learning, Principal Component Analysis.**

I. INTRODUCTION

The amalgamation of data from multiple modalities, such as images, audio, and textual information enables the development of a comprehensive decision-making application or systems with enhanced information. Image fusion has proved critical in many real-world applications, e.g., medical [1], remote sensing [2], military [3], and computer vision applications [4]. Fusion tasks become more difficult and complex due to incomplete data, nonlinear relationships between input data, misalignment, and heterogeneous data. The presence of noise, incomplete data, and adverse environmental conditions significantly make it more challenging in terms of reliability. A single fused image with supplementary information is generated from various sensors and devices. Images from a visible sensor, for instance, are full of subtle details like contrast, colour, and texture. However, this is unable to differentiate between objects and backgrounds in low light levels. Thermal sensors can identify key features that distinguish objects from the background both during the day and at night, but they cannot identify the colour space and texture of an object. The reason behind this is that thermal sensors operate in the 8–14 μm wavelength range, whereas visible sensors operate in the 300–530 μm range [5]. The fusion method has gained significant devotion in research domains, like medical imaging, remote sensing, robotics, and autonomous driving. Numerous studies that address various facets of multimodal image fusion, including fusion methods, performance assessment, and applications, have been published.

II. RELATED WORK

Thanks to developments in multimodal image fusion technology, humans can now work with the newest intelligent machines, improving their capabilities and quality of life.

Grenze ID: 01.GIJET.11.2.737

A multi-modal image fusion's taxonomy is determined by a number of factors, including input modalities, time and complexity factors, fusion levels, fusion strategies, and more [12].

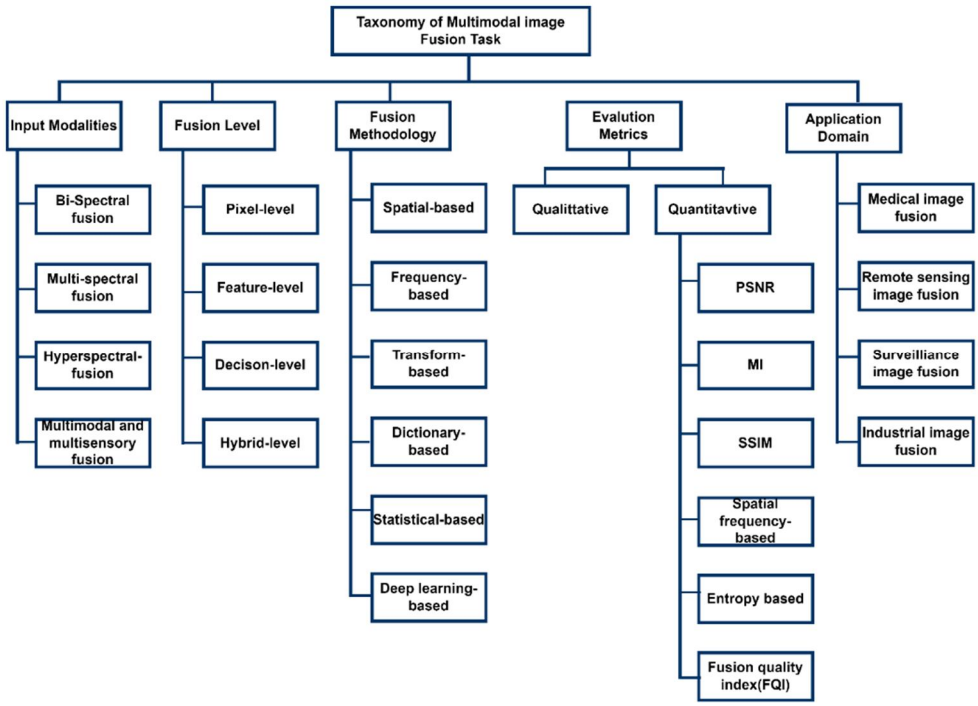

Fig. 1. Multimodal Image Fusion-Taxonomy. [8]

A. Modalities

Bi-spectral fusion provides detailed information about an object by combining two modalities, usually visible and infrared images. The color of the scene affects visible images, whereas the temperature of the scene affects infrared images. When there is little visible light and infrared can offer crucial information, bi-spectral fusion can be especially helpful. More than two modalities involving more than two spectrum ranges typically three to ten distinct bands are combined in multispectral fusion. The objective is to remove unnecessary features from each spectral band while retaining the most relevant and instructive ones [12]. High spectral resolution images are integrated using hyperspectral. It provides a comprehensive spectral signature that can be used to identify and categorize ma-terials based on their image by capturing hundreds or even thousands of narrow, continuous spectral bands over a broad range of wavelengths. Combining data from several sources, such as various sensor types, to increase the precision and dependability of the final image is known as multimodal and multisensory fusion [15].

Challenges

There are a lot of challenges across different input modalities challenges during fusion activity, these are related to sensor resolution mismatch, noise in sensors vary or have a unique pattern, and trade-off between spatial and spectral fidelity means merging infrared with visible image can degrade spectral details, and hav-ing a consistent quality under rough weather conditions. Hyperspectral fusion can be computationally demanding and requires a large number of resources and processing power moreover, it is prone to amplifying noise degrading overall image quality. It captures detailed spectral information, while other input modali-ties provide spatial or structural data. It will be more challenging to perform techniques like pixel-level fusion as this may fail to capture the intricate rela-tionship between data types [8].

Possible Solutions

There are a few possible solutions to address the challenges, but this comes with an additional computational cost with advances in algorithm & sensor technolo-gy.

One of the methods is a coupled non-negative block-term tensor model and it was proposed by authors Guo et al. [13] to estimate the optimal high spatial resolu-tion hyperspectral images. The image's sparsity is described by the L1-norm, and piecewise smoothness is described by total variation (TV). The model is solved using the

alternating multiplier method and the proximal alternating optimization technique. To create a high-resolution hyperspectral image. Another model based on a deep learning technique for combining high-resolution and low-resolution multispectral images was proposed by author Feng et al [14]. An iterative algo-rithm that uses the gradient method to solve the model.
Multimodal and multisensory fusion there have been advancements in terms of learning complex spatial-spectral relationships using deep learning models ex: - CNN, GANs etc..) and transformer-based multiscale fusion [21].

B. Fusion Levels

An overview of various fusion-level approaches, concentrating on the ad-vantages, disadvantages, applicability, challenges, and Improvisation of these strategies.

TABLE I: A SUMMARY OF MULTIMODAL FUSION LEVEL. [8].

Fusion Level	**Description**	**Advantages**	**Disadvantage's**	**applications**	**Challenges**	**Improvisation work**
Pixel-level fusion	Combines raw data(pixels) intensities by using mathematical operations, like an averaging or maximum selection	An intuitive and simple to implement Retains original image data, Simple and computationally efficient	Sensitive to data misalignment and noise	Image sharpening, low-level feature extraction	Accurate image registration is critical. Managing noise and artifacts is challenging.	Advanced multi-resolution techniques to overcome misalignment
Feature level fusion	Extracts feature from edged, texture, or shapes and combine them	Reduces redundancy. Robust to misregistration, preserves spatial and spectral information,	Requires effective feature extraction. Requires careful feature selection as extraction, may not capture all vital data points, and even can be computationally expensive	Object recognition, image segmentation, object tracking,	Robust feature extraction methods. Features alignment from various modalities.	Integration of deep learning for automated feature extraction
Decision-level fusion	A combination of classification outcome from various input images leveraging mathematical operations such as majority voting or weighted averaging	Strong against noise and misregistration, appropriate for complex image processing tasks	Sensitive to the quality and reliability of the classification results from each input image, may not capture all relevant information	Object classification, scene recognition, face recognition, medical diagnosis	Managing conflicting decisions. Combining probabilistic outputs effectively.	Advanced decision fusion methods using ensemble learning.
Hybrid-level fusion	A combination of pixel, feature, and decision-level fusion to achieve better performance and accuracy.	Due to different levels of fusion techniques, it provides greater flexibility and optimization.	More complex and computationally expensive than individual levels of fusion, requires careful consideration of the computational and memory requirements	Remote sensing, medical imaging, surveillance, robotics, autonomous vehicles, quality control in manufacturing and inspection	Optimizing across levels without introducing redundancy	Development of AI-driven adaptive fusion frameworks.

C. Advancement in Multi-Modal Fusion

Image fusion using deep learning (CNN, Autoencoders, GAN, Transformers) is overcoming the drawbacks of conventional techniques and acting as a driving force behind the advancement into image fusion [17].
Image Fusion with CNN.
An increased traction owing to its capability to learn from features automatically from input images across modalities is used to generate fused data, by mining vital from ground truth images. Raw images (e.g., visible and infrared) are direct-ly fed into CNNs to generate fused output. Simple CNN architectures like VGG-Net and

AlexNet are used to learn spatial features and combine modalities. CNNs are used for pixel-level and feature-level to generate informative fused images for tasks like object detection. Typically, we perform tasks like data Preparation, CNN architecture selection, training the network, fusing inputs, and evaluation [17].

The basic CNN with no attention to modality-specific features often fails to pre-serve spectral or granular details. Further enhancements have taken place in this field with multiscale feature fusion (ex: U-Net based architecture), Attention Mechanisms, and Deep Generative Models preserving both high-level and low-level semantics [20]. Encoder-decoder structures will enhance spatial fidelity.

Multimodal image fusion using Auto-encoder.

Another well-liked deep learning model is the autoencoder, which can the mean-ingful, compressed representations of input images. By encoding the input images into a lower-dimensional feature space and then decoding them to produce a fused representation, this can be used to carry out feature-level fusion [19].

Autoencoders, including traditional autoencoders, variational autoencoders (VAEs), and stacked autoencoders, have been applied to multimodal image fusion tasks

Traditional Autoencoders. Learn compact feature representations of input modal-ities (e.g., visible and infrared) for fusion.

Variational Autoencoders (VAEs). VAEs introduce a probabilistic distribution on the latent space to model uncertainties across modalities.

Deep Autoencoders and Stacked Autoencoders. Multiple layers of encoders and decoders are stacked to extract fine-grained, multi-scale features. Enhanced fea-tures are combined for fusion.

Multimodal image fusion using generative adversarial networks (GAN).

GAN has emerged as a powerful model for most of the use cases, it has signifi-cantly improved from its predecessor. It relies on generative capabilities to pro-duce realistic fused outputs by learning data distribution from various modalities [21].

GANs consist of two networks

- Generator GGG. Produces synthetic (fused) images from input data.
- Discriminator DDD. Distinguishes between ground truth and synthetic images. The networks are trained in a min-max game to optimize the adversarial loss

The GAN is trained using an input image dataset and the corresponding fused image data. The generator is trained to produce data that closely resembles the ground truth, while the discriminator is trained to differentiate between ground truth and generated data. To guarantee that the fused image is realistic and retains the most pertinent information from the input images, the loss function used dur-ing training usually combines adversarial and content loss [30].

Deep Convolutional GAN (DCGAN). In the year 2014 the proposal of the first GAN [25], was later improved upon to create the DCGAN [26]. The goal of DCGAN is to replace fully connected (FC) layers with convolutional layers. To stabilize the GAN's training, additional adjustments were proposed in addition to the convolutional layers.

Batch normalization was utilized in both G and D to enhance the diversity of the generated samples and lower noise [27]. The majority of the following GAN models use the innovations that have been suggested in recent research. There are a few other GAN variants like Conditional GAN (cGAN), multi-scale, PROGANs, and Attention-Guided GANs. Use attention mechanisms and multi-scale architectures to prioritize relevant regions/features. Progressive Growing of GANs (ProGAN). Concatenating various training phases is the fundamental con-cept of progressive networks [30].

Multimodal image fusion using transformers.

With the capacity to capture global contextual information and long contextual dependencies, Transformers has revolutionized and taken deep learning models to the next level. Transformers are an innovative variant of the neural networks family and are widely used in Natural Language Processing (NLP), and off-late it has expanded its capability to image fusion tasks [23]. The astounding results from Transformer models on natural language tasks have intrigued the computer vision area to study the application of its related use cases. Among their salient benefits, Transformers enable modeling long dependencies between input se-quence elements and support parallel processing of sequence as compared to recurrent networks [31].

Self-attention. This captures the intra-modality relationships and can compute attention scores for each feature map within a single modality. Self-attention calculates context relevance with other data in a sequence of data (e.g., which words are likely to come together in a sentence). Transformers, which explicitly model the interactions between all entities of a sequence for structured predic-tion tasks, depend on the self-attention mechanism. A self-attention layer aggre-gates global information from the entire input sequence to update each sequence component [23].

III. Real-world Generalization

Various methods have been applied in real-world applications to check the be-havior and efficacy. Table 2 provides a real-world realization under different methods with corresponding evaluation metrics. Quantitative metrics are used to evaluate the model's performance objectively.

TABLE II: Application Realization Under Various Methods

Methods	**Realization**	**Quantitative Evaluation metrics**
CNN	**Medical Imaging**: MRI and PET scans are fused to get an augmented image. MRI provides high-resolution images of the body, such as organs, muscles, etc. PET provides functional information by highlighting areas with high metabolic activity, typically in identifying cancer cells. Augmenting these two provides a granular-level anatomical structure. This will aid doctors in precisely locating and characterizing tumors.	EOG, RMSE, PSNR, SSIM, MI, VIF, QAB.
GAN	**Remote Sensing:** Helps to combine images from different sensors such as visible light and infrared to improve the interpretability and usability of aerial image data. Helps in environmental monitoring.	SD, SSIM, VIF, MI
AutoEncoders	**Image Denoising:** Removing noise from images in applications like surveillance and astronomy	EN, MI, QAB/F,SD
Transformers	**Vision Models:** Useful in microscopy, where images that are sharp across different planes, enhancing both research and diagnostic purpose	AP, MI, EN, SSIM

IV. Future Directions

Even though there are still a few challenges to be addressed like noise, illumina-tion, temperature, multiple modalities, and absence of ground truth, multimodal image fusion advancement brings us one step closer to developing an autono-mous system that can work smoothly and effectively with minimal human loop [7]. While this is still in a nascent state, the development happening in deep learning and other conventional learning approaches looks quite promising. CNN [8], GAN [30] and Transformer based methods [31] have demonstrated signifi-cant promising results.

Traditional learning models have been successfully used, including multi-scale and multi-feature decomposition techniques and methods based on sparse representation [9]. There are still open gaps or challenges that need traction to build a robust model.

Interpretability: One crucial area of research is creating interpretable deep learn-ing techniques to describe the fusion results. Understanding why a specific image fusion result was generated in the manner that it was is typically difficult.

Robustness. Changes to the input data, like occlusions or lighting conditions, will subtly impact the generated images.

Generalization. Training a model that can generate an appropriate result even on a smaller data set. Eliminating overfitting of the model.

Multimodal data fusion. The majority of current techniques concentrate on bi-spectral images. One unresolved issue is creating a model for fusing multiple input modalities.

Real-time processing. Due to intensive resource utilization, adhering to sacro-sanct time constraints will be a challenge to perform a complex computation and develop a lightweight model.

V. Conclusion

In recent years, various deep-learning techniques have been used for multimodal image fusion the widespread use of deep learning is altering the way problems are solved in the real world.

This paper aims to examine and address the recent advancements and methods used for multimodal image fusion as this will serve as the foundation for numerous upcoming computer vision-based research and is extremely promising.

There is much scope for advancement when dealing with multimodal fusion techniques by utilizing sophisticated techniques such as GAN and transformer-based architecture models.

Declaration of Competing Interest

We declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

DATASET READINESS

The data sets utilized to conduct surveys as part of the study

TABLE III. BENCHMARKED DATASETS FOR MULTIMODAL IMAGE FUSION [8].

Dataset	Source
ChikuseiDataset	http://park.itc.u tokyo.ac.jp/sal/hyperdata
IndianPines	https://www.ehu.eus/ccwinzco/ index.php?title=Hyperspectral_Remote_Sensing_Scenes
Harvard	http://vsion.seas.havard.edu/hyperspec/dowload.htm
Cuprite	https://aviris.jpl.nasa.gov/
ATLASwholebraindataset	http://www.med.harvard.edu/ aanlib/home.html
Cancer ImagingArchive (TCIA)	https://www. cancerimagingarchive.net/about the-cancer-imaging-archive-tcia/
FLAME2	https://ieee-dataport.org/open access/flame-2-fire-detection-and modeling-aerial-multi-spectral image-dataset
COCO	

REFERENCES

[1] Shutao Li, Xudong Kang, Leyuan Fang, Jianwen Hu, and Haitao Yin, “Pixel-level image fusion: A survey of the state of the art,” information Fusion, vol. 33, pp. 100–112, 2017

[2] M. Eslami and A. Mohammadzadeh, "Developing a Spectral-Based Strategy for Urban Object Detection From Airborne Hyperspectral TIR and Visible Data," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 9, no. 5, pp. 1808-1816, May 2016, doi: 10.1109/JSTARS.2015.2489838.

[3] Jiayi Ma, Yong Ma, Chang Li, Infrared and visible image fusion methods and applications: A survey, Information Fusion, Volume 45, 2019, Pages 153-178, ISSN 1566-2535, https://doi.org/10.1016/j.inffus.2018.02.004.

[4] C. Lopez-Molina, J. Montero, H. Bustince, B. De Baets, Self-adapting weighted operators for multiscale gradient fusion,Information Fusion,Volume 44,2018,Pages 136-146,ISSN 1566-2535,https://doi.org/10.1016/j.inffus.2018.03.004.

[5] Jian Tian, Yanhua Leng, Zhenhuan Zhao, Yang Xia, Yuanhua Sang, Pin Hao, Jie Zhan, Meicheng Li, Hong Liu, Carbon quantum dots/hydrogenated TiO2 nano-belt heterostructures and their broad spectrum photocatalytic properties under UV, visible, and near-infrared irradiation, Nano Energy, Volume 11, 2015, Pages 419-427, ISSN 2211-2855, https://doi.org/10.1016/j.nanoen.2014.10.025.

[6] Xin Jin, Qian Jiang, Shaowen Yao, Dongming Zhou, Rencan Nie, Jinjin Hai, Kangjian He, A survey of infrared and visual image fusion methods, Infrared Physics & Technology, Volume 85, 2017, Pages 478-501, ISSN 1350-4495, https://doi.org/10.1016/j.infrared.2017.07.010.

[7] Hemanth Kumar, Pratik Pimparkar (2018); Data Fusion for the Internet of Things; Int J Sci Res Publ 8(3) (ISSN: 2250-3153). http://www.ijsrp.org/research-paper-0318.php?rp=P757283.

[8] Huang, B., Yang, F., Yin, M., Mo, X., & Zhong, C. (2020). A Review of Multi-modal Medical Image Fusion Techniques. Computational and mathematical methods in medicine, 2020, 8279342. https://doi.org/10.1155/2020/8279342

[9] Durga Prasad Bavirisetti, Ravindra Dhuli, Multi-focus image fusion using multi-scale image decomposition and saliency detection, Ain Shams Engineer-ing Journal, Volume 9, Issue 4, 2018, Pages 1103-1117,ISSN 2090-4479, https://doi.org/10.1016/j.asej.2016.06.011.

[10] Kelei He and Chen Gan and Zhuoyuan Li and Islem Rekik and Zihao Yin and Wen Ji and Yang Gao and Qian Wang and Junfeng Zhang and Dinggang Shen,Transformers in Medical Image Analysis: A Re-view,2022,https://doi.org/10.48550/arXiv.2202.12165

[11] Huang, J., Le, Z., Ma, Y. et al. A generative adversarial network with adaptive constraints for multi-focus image fusion. Neural Comput & Applic 32, 15119–15129 (2020). https://doi.org/10.1007/s00521-020-04863-1

[12] Shrida Kalamkar, Geetha Mary A., Multimodal image fusion: A systematic review, Decision Analytics Journal, Volume 9,2023,100327, ISSN 2772-6622, https://doi.org/10.1016/j.dajour.2023.100327.

[13] H. Guo, W. Bao, K. Qu, X. Ma, M. Cao, Multispectral and hyperspectral image fu sion based on regularized coupled non-negative block-term tensor decom-position, Remote Sens. (Basel) 14 (21) (2022) http://dx.doi.org/10.3390/rs14215306

[14] X. Feng, L. He, Q. Cheng, X. Long, Y. Yuan, Hyperspectral and multispectral remote sensing image fusion based on endmember spatial information, Remote Sens. (Basel) 12 (6) (2020) http://dx.doi.org/10.3390/rs12061009.

[15] F. Abdullah Al-Wassai, N. Kalyankar, A.A. Al-Zaky, A. Professor, Multisensor Images Fusion Based on Feature-Level.

[16] M.A. Bakr, S. Lee, Distributed multisensor data fusion under unknown correla-tion and data inconsistency, Sensors (Switzerland) 17 (11) (2017) http://dx.doi.org/ 10.3390/s17112472.

[17] M. Nejati, S. Samavi, S. Shirani, Multi-focus image fusion using dictionary-based sparse representation, Inf. Fusion 25 (2015) 72–84, http://dx.doi.org/10.1016/ j.inffus.2014.10.004.

[18] M. Kaur, D. Singh, Multi-modality medical image fusion technique using multi objective differential evolution based deep neural networks, J. Amb. Intell. Humaniz. Comput. 12 (2) (2021) 2483–2493, http://dx.doi.org/10.1007/s12652 020-02386-0.
[19] N. Tawfik, H.A. Elnemr, M. Fakhr, M.I. Dessouky, F.E.A. El-Samie, Multimodal medical image fusion using stacked auto-encoder in NSCT domain, J. Digit. Imag. 35 (5) (2022) 1308–1325, http://dx.doi.org/10.1007/s10278-021-00554-y.
[20] M. Kaur, D. Singh, Fusion of medical images using deep belief networks, Cluster Comput. 23 (2) (2020) 1439–1453, http://dx.doi.org/10.1007/s10586 019-02999-x.
[21] J. Zhang, et al., Transformer based conditional GAN for multimodal image fusion, IEEE Trans. Multimedia (2023) 1–14, http://dx.doi.org/10.1109/TMM. 2023.3243659.
[22] Y. Wang, J. Peng, J. Zhang, R. Yi, Y. Wang, C. Wang, Multimodal industrial anomaly detection via hybrid fusion, 2023, [Online]. Available: http://arxiv.org/ abs/2303.00601
[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in NeurIPS, 2017.
[24] Radford A., Metz L., Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks
[25] Jump-er J., Evans R., Pritzel A., Green T., Figurnov M., Tunyasuvunakool K., Ronneberger O., Bates R., Žídek A., Bridgland A., et al. High accuracy protein structure prediction using deep learning
[26] Li Y., Xiao N., Ouyang W.Improved boundary equilibrium generative adversar-ial networks
[27] Wu S., Li G., Deng L., Liu L., Wu D., Xie Y., Shi L. L1 norm batch normalization for efficient training of deep neural networks
[28] Hubel D.H., Wiesel T.N. Receptive fields of single neurones in the cat's striate cortex J. Physiol., 148 (3) (1959), pp. 574-591
[29] Rusu A.A., Rabinowitz N.C., Desjardins G., Soyer H., Kirkpatrick J., Kavukcuoglu K., Pascanu R., Hadsell R.Progressive neural networks
[30] Karras T., Aila T., Laine S., Lehtinen J.Progressive growing of GANs for im-proved quality, stability and variation(2018)
[31] Transformers in Vision: A Survey 1 Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah
[32] D. Sara, A.K. Mandava, A. Kumar, S. Duela, A. Jude, Hyperspectral and Multi spectral Image Fusion Techniques for High-Resolution Applications: A Review. http://dx.doi.org/10.1007/s12145-021-00621-6/Published.