International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 12 Issue: 08 | Aug 2025 www.irjet.net p-ISSN: 2395-0072

A REVIEW ON DEEP LEARNING MODELS FOR IMAGE ENHANCEMENT THROUGH VISIBLE AND INFRARED IMAGE FUSION

Richa Sukhdev Ambapkar[1], Dr. A. S. Yadav[2]

[1]Department of Computer Science and Engineering, D.Y. Patil College of Engineering & Technology,Kolhapur, Maharashtra, 416006, India

[2] Associate Professor, Department of Computer Science and Engineering, D.Y. Patil College of Engineering & Technology,Kolhapur, Maharashtra, 416006, India

---

Abstract - *Combining visible and infrared (IR) images has emerged as an integral technique in image enhancement, allowing for improved perception, identification, and decision- making in a range of industrial and real-world situations. By combining spatial detail from visible light with thermal information from infrared imaging, fused outputs offer significantly improved clarity and utility, particularly in low-light or obscured environments. This review presents a comprehensive study of recent deep learning models developed for visible and IR image fusion. It explores a variety of approaches including convolutional neural networks (CNNs), encoder-decoder architectures, and attention mechanisms. The paper also discusses key application areas such as surveillance, autonomous vehicles, medical diagnostics, and environmental monitoring. In addition, major challenges such as heterogeneous data alignment, high computational cost, and the scarcity of labeled datasets are examined. In an attempt to facilitate the creation of more resilient and intelligent image fusion systems, future research directions are finally noted.*

Key Words: **Image Fusion, Deep learning, Convolutional Neural Networks (CNN), Image enhancement, feature extraction,Visible Images, Infrared Images**

1.INTRODUCTION

An important field of study in computer vision and image processing is picture fusion, specifically the combination of visible and infrared (IR) images. Visible images capture high-resolution spatial details and color textures under optimal lighting conditions, whereas infrared images record thermal radiation, allowing visibility even in darkness, smoke, or fog. Combining these complementary modalities through fusion techniques results in enhanced images with greater detail, contrast, and semantic richness.

The demand for robust image fusion techniques is growing across several critical domains. In surveillance and security, fused images can improve the detection of intruders and suspicious activities in low-light or night-time conditions. In autonomous driving, fusing visible and IR inputs enables better recognition of pedestrians and obstacles under varying environmental conditions. Medical imaging benefits from the fusion of thermal and visible data to detect conditions such as inflammation or tumors with higher accuracy. Additionally, environmental monitoring applications— such as wildfire detection or crop health assessment— can be improved through the integration of spatial and thermal information.

Traditionally, image fusion methods relied on manual feature extraction or statistical models, which often struggled with robustness, adaptability, and computational efficiency. The advent of deep learning has introduced a paradigm shift, offering data-driven models that learn rich feature representations automatically and can adapt across multiple fusion scenarios. Convolutional Neural Networks (CNNs), encoder-decoder models, residual learning, and attention mechanisms have been increasingly applied to enhance merged pictures' efficacy and quality.

Recent developments in deep learning models for visible and infrared image fusion are thoroughly examined in this article. It explores a range of techniques, compares their strengths and limitations, and identifies challenges and opportunities for future research. The objective is to guide researchers and practitioners in selecting and designing effective image fusion solutions for real-world applications.

2. Methodology

The proposed methodology begins with collecting dual-modality image inputs—visible and IR images—from relevant datasets. These inputs undergo preprocessing and normalization steps including data cleaning, formatting, and scaling to ensure consistency across modalities. The core feature extraction is performed using a dual-branch convolutional neural network (CNN), where each branch independently processes one image type. The extracted features are then fused using an attention-based fusion layer to preserve and emphasize significant image details. The fused features are passed through a classification layer and tested using standard model

training practices. The system's performance is evaluated using metrics such as precision, recall, F1-score, accuracy, and ROC-AUC curve, providing a comprehensive assessment of its effectiveness in image enhancement and classification tasks.

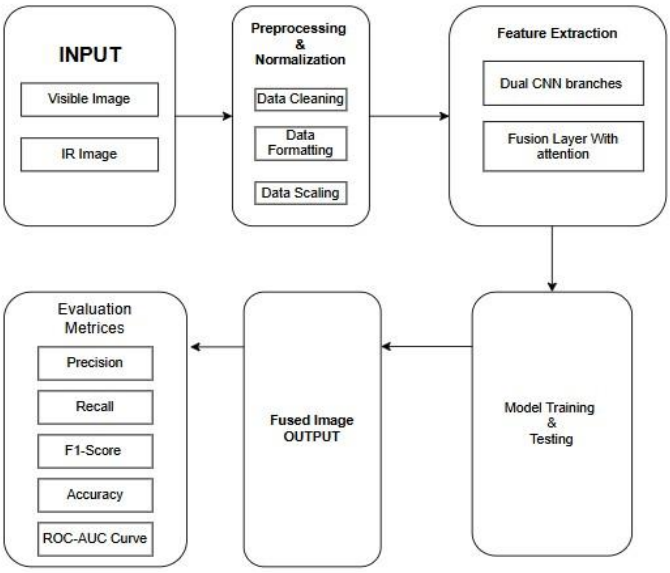

Figure 1: Conceptual Framework of Proposed Work

The proposed methodology divides into following modules as-

Module 1: Data Collection

This module focuses on acquiring dual-modality image datasets, including both visible spectrum images and infrared (IR) images, from publicly available sources. Suitable datasets are collected from platforms such as FLIR Thermal Dataset, KAIST Multispectral Pedestrian Dataset, and other benchmark repositories for multi-modal image fusion and classification. These datasets typically include day and night-time images, urban and natural scenes, and varying environmental conditions to ensure diversity. The visible images capture color and texture, while the IR images provide thermal information, enabling robust analysis. This dual-source approach allows the system to fuse complementary information for better accuracy in downstream tasks such as object detection, surveillance, and scene understanding.

Module 2: Data Preprocessing

The preprocessing module standardizes the input data to improve compatibility and performance during training. Raw visible and IR images often come with varying resolutions, color spaces, and noise levels. Therefore, preprocessing includes resizing the images to a common resolution, converting them into a consistent format (e.g., grayscale or RGB as needed), and applying noise reduction techniques. The data is then normalized to scale pixel values within a fixed range, improving model convergence. Furthermore, image alignment and registration techniques may be used to ensure spatial correspondence between visible and IR images. This preprocessing ensures clean, uniform, and well-structured input that facilitates efficient feature extraction and fusion in subsequent stages.

Module 3: Feature Extraction and Representation

This module is responsible for extracting meaningful features from both visible and infrared (IR) images to aid in image fusion and classification. The system employs dual CNN branches, where one convolutional neural network processes visible images and the other processes IR images. Each branch learns modality-specific features such as edges, textures, and temperature gradients. These features are then passed through a fusion layer with an attention mechanism, which intelligently combines relevant information from both sources. The attention mechanism ensures that the most informative features from each modality are emphasized, leading to better fused representations. This dual-stream approach enhances the ability to capture complementary characteristics of the two modalities, making the fused image more informative for downstream tasks like object recognition, surveillance, or thermal analysis.

Module 4: Model Training and Testing

In this module, the extracted and fused features are used to train a deep learning model for classification or detection tasks. The fused output from the attention-based layer is flattened and passed through fully connected layers, where the model learns high-level semantic patterns. During training, the model is optimized using loss functions such as categorical cross-entropy or binary cross-entropy, depending on the application. Training is performed using backpropagation and gradient-based optimizers like Adam. A validation set is used to fine-tune hyperparameters and avoid overfitting. Once trained, the model is tested on unseen data to evaluate its generalization ability. This module also supports inference on real-time data, allowing for live deployment in intelligent imaging systems. The output is a fused image with associated labels or decisions that demonstrate improved performance due to the synergy between visible and IR modalities.

Module 5: Fused Image Output

The result of the fusion process is a single image or decision output that retains important features from both the visible and IR sources. This output is used either for visualization or as an input to downstream tasks (e.g., object detection, classification).

Module 6: Evaluation Metrics

This module is dedicated to evaluating the performance of the fused image classification model using standard quantitative metrics. Once the model is trained and tested, its output is assessed using precision, recall, F1-score, accuracy, and the ROC-AUC curve.

- Precision measures how many of the predicted positive results are actually correct, ensuring the model avoids false positives.
- Recall evaluates how well the model identifies all relevant instances, minimizing false negatives.
- F1-Score provides a balanced measure between precision and recall, especially useful for imbalanced datasets.
- Accuracy gives an overall percentage of correctly classified instances.
- The ROC-AUC Curve plots the true positive rate against the false positive rate, helping visualize the model's discriminative ability across different thresholds.

3. Literature Review

Deep learning has revolutionized by making automatic feature extraction possible in the field of image fusion, adaptive fusion strategies, and end-to-end learning architectures. Numerous methods have been proposed to enhance the fusion of visible and infrared (IR) images using Convolutional Neural Networks (CNNs), encoder-decoder models, attention mechanisms, and generative adversarial networks. This section summarizes and compares key contributions from the recent literature.

- **Liu et al.** suggested an algorithm for extracting multilayer features from input photos that is based on Convolutional Sparse Representation (CSR). While effective in combining key elements from both modalities, the model's robustness to diverse input conditions remains limited.
- **Liu et al.** later introduced a **CNN-based fusion approach** using image patches and decision graphs. Although the technique improved feature-level integration, it lacked generalizability across multiple fusion tasks.
- **Li et al.** provided a framework supporting deep learning leveraging **VGG-19** to split source image information into low-frequency and texture components. Multilayer fusion helped extract deep features, though the method suffered from high computational cost and reliance on manually engineered features.
- **Prabhakar et al.** recommended an encoder-fusion-decoder based on CNN architecture that did not require parameter adjustment for input changes. By optimizing loss functions, the model achieved reliable fusion, though generalizability across different input types was not fully explored.
- **Ma et al.**provided a method for eliminating total variation that maintained the gradients in visible images and the pixel intensity in infrared images. While this approach maintained essential information, it was prone to visual artifacts in complex scenes.
- **Li et al.** used a **ResNet-based framework** combined with Zero-Phase Component Analysis (ZCA) and L1-normalization to extract deep features and reconstruct fused images. Although effective in retaining detail, the model lacked flexibility for varied fusion tasks.
- **Xu et al.** presented an **unsupervised, densely connected CNN** trained across multiple fusion tasks using **Elastic Weight Consolidation (EWC)** to retain knowledge across domains. This method showed promise for adaptive multi-task fusion but required further optimization for real-time and high-resolution applications.
- **Zhang et al.** presented an end-to-end CNN with distinct modules for feature extraction, combination, and reconstruction. The model demonstrated good fusion accuracy but lacked analysis of robustness under multimodal input variability.
- **Chen et al.** designed a **Multilayer Fusion CNN (MLF-CNN)** for pedestrian detection in low-light conditions using multispectral inputs. A multiscale region proposal network enabled the fusion of IR and visible features, although scalability for real-time deployment remained a concern.
- **Hou et al.** introduced a method that utilized a **mixed loss function** to adaptively merge IR and visible inputs while suppressing noise. The model demonstrated strong preservation of texture and salient features but needed validation on larger datasets.
- **Zhang et al.** conducted a **comprehensive survey** of visible-infrared fusion (VIF) methods, datasets,

and evaluation techniques. While this work serves as a valuable reference, it did not propose new fusion models.

- **Meher et al.** introduced a **region-based fusion technique**, called Adaptive Transition Region Extraction (ATRE), which selectively extracts bright thermal features and preserves visual clarity from visible images. The method proved effective for military and surveillance tasks.

- **Sulthana et al.** presented an adversarial network-based method that enhances source images before fusion. This **GAN-based model** operated without the need for supervised training data and delivered strong fusion quality by improving contrast and texture—offering a new benchmark in image fusion performance.

4. Discussion and Research Gaps

Deep learning-based visible and infrared (IR) image fusion is getting more popular, which demonstrates its potential to transform imaging applications in complex environments. While the current literature presents a wide range of architectures and fusion strategies, a closer examination reveals several gaps that still need to be addressed for more robust, efficient, and general-purpose solutions.

4.1 Fragmented Focus Across Applications

Many research efforts are tailored to specific use cases—such as surveillance, pedestrian detection, or medical diagnostics—resulting in models that are highly task-specific. While this can yield strong performance in a particular domain, it limits the broader applicability of these models. There's a noticeable lack of **unified frameworks** that can adapt across different domains and data distributions.

4.2 Lack of Benchmark Datasets

A major limitation across the studies reviewed is the **absence of standard, large-scale, and diverse datasets** that include paired visible and IR images. Many models are evaluated on privately curated or small-scale datasets, hindering reproducibility and comparative analysis. The research community needs to invest in the creation and open dissemination of benchmark datasets that reflect varied real-world conditions—night/day, fog, rain, moving objects, etc.

4.3 Inadequate Evaluation Practices

Current evaluation approaches rely heavily on visual inspection or a limited set of quantitative metrics. Few studies perform thorough evaluations across both **low-level (image quality)** and **high-level (task performance)** metrics. Additionally, **human perception-driven evaluations**, such as user studies, are rarely conducted even in critical domains like healthcare or defense.

4.4 Generalization and Overfitting

Several models demonstrate good performance on specific datasets but **fail to generalize** when applied to different scenes, objects, or lighting conditions. This suggests potential overfitting, particularly in architectures trained with limited or highly curated data. Future models should be trained and validated using **cross-dataset evaluations** to ensure robustness.

4.5 Limited Real-Time Implementation

While some models show theoretical promise, **real-time performance remains underexplored**. Many methods involve deep encoder-decoder structures, attention mechanisms, or adversarial training—all of which increase inference time and computational load. There is a need to balance model complexity with speed and develop **lightweight architectures** optimized for edge deployment.

4.6 Absence of Interpretability Mechanisms

As deep models become more opaque, **explainability is largely ignored** in fusion research. For high-stakes domains like healthcare, military, and autonomous driving, black-box models can reduce trust and hinder deployment. Techniques such as feature importance visualization, attention maps, and saliency detection need to be integrated into future fusion pipelines.

4.7 Underutilized Advances in Other Domains

Recent advances in **transformers, diffusion models, and self-supervised learning** have yet to be fully exploited for image fusion. Most existing work still relies on CNNs. Exploring these newer paradigms could lead to breakthroughs in model accuracy, adaptability, and interpretability.

5. Future Scope

With new techniques and applications appearing annually, the area of visible and infrared (IR) image fusion utilizing deep learning is developing quickly. However, in order to attain wider acceptance and usefulness, future studies need to tackle current constraints and investigate novel avenues. This section lists promising directions for further research and development.

5.1 Development of Lightweight and Real-Time Models

The enormous computational complexity of current models is one of their main limitations. The development of lightweight architectures that can be instantly deployed on resource-constrained devices like edge AI platforms, embedded systems, and drones should be the main focus of future research. To cut down on model size and inference time, methods including knowledge distillation, quantization, and model pruning could be investigated.

5.2 Adoption of Transformer and Diffusion Architectures

While CNNs dominate current fusion architectures, **transformers**—widely successful in image classification and natural language processing—have demonstrated promise in simulating contextual linkages and long-range dependencies. Similarly, **diffusion models** have recently emerged as powerful tools for image generation and could be adapted for fusion tasks to produce higher-quality outputs with better control over the generation process.

5.3 Self-Supervised and Unsupervised Learning

Lack of labeled training data is a recurring challenge in this field. To overcome this, future research should explore **self-supervised learning frameworks** that can learn from unlabeled or weakly labeled data. Unsupervised learning techniques can also be improved to handle domain shifts, ensuring models perform consistently across varied environmental conditions.

5.4 Multi-Modal and Multi-Sensor Fusion

In addition to visible and IR data, other modalities such as depth, LiDAR, or hyperspectral images can be integrated for richer perception. Future models should investigate **multi-modal fusion pipelines**, allowing systems to adaptively choose relevant information from diverse sensor inputs.

5.5 Explainable and Trustworthy AI

As deep learning models are increasingly deployed in critical applications like defense and healthcare, **explainability and trust** become essential. Future research must incorporate interpretability tools such as attention visualization, decision reasoning modules, and confidence scoring. This will help stakeholders better understand and trust the fusion decisions made by AI systems.

5.6 Standardization of Datasets and Benchmarks

The community would greatly benefit from the creation of **open-source, standardized datasets** with paired visible and IR images across varied environments. Moreover, unified **benchmarking protocols** should be established to compare models based on accuracy, speed, robustness, and visual quality.

5.7 Application-Specific Optimizations

There is potential to build **domain-optimized fusion models**, fine-tuned for specific applications like smart surveillance, autonomous driving, medical diagnostics, or remote sensing. This includes optimizing for task-specific objectives such as detection accuracy, thermal signature clarity, or anomaly localization.

6. Conclusion

Using deep learning approaches to fuse visible and infrared (IR) images has shown itself to be a potent tool for improving scene understanding, situational awareness, and image quality in a variety of demanding contexts. By leveraging the complementary strengths of visible and thermal imaging, fused outputs enable more accurate perception and decision-making in critical applications such as surveillance, autonomous driving, medical diagnostics, and environmental monitoring.

Numerous new research initiatives have been reviewed in this study, emphasizing the transition from conventional techniques to complex deep learning systems. Techniques such as convolutional neural networks (CNNs), encoder-decoder models, attention mechanisms, and generative adversarial networks (GANs) have significantly improved the quality and adaptability of fusion results. Nonetheless, issues like lack of interpretability, processing demands, model generalization, and data scarcity continue to exist.

Through a detailed analysis, we have identified key research gaps, including the need for standardized datasets, better evaluation metrics, lightweight models for real-time processing, and the integration of newer learning paradigms such as transformers and self-supervised learning. Additionally, there is a pressing demand for explainable and trustworthy fusion systems that can be confidently used in high-stakes domains.

As the field continues to grow, addressing these gaps will be essential for the development of robust, scalable, and intelligent image fusion solutions. Deep learning offers a promising foundation—but its success will depend on cross-disciplinary collaboration, ethical deployment, and continuous innovation in both model design and application strategy.

7. References

1. S. Kalamkar and A. G. Mary, “Multimodal image fusion: A systematic review,” *Decision Analytics Journal*, vol. 9, p. 100327, Dec. 2023, doi: 10.1016/j.dajour.2023.100327.

2. C. Brunner, T. Peynot, T. Vidal-Calleja, and J. Underwood, "Selective Combination of Visual and Thermal Imaging for Resilient Localization in Adverse Conditions: Day and Night, Smoke and Fire," *Journal of Field Robotics*, vol. 30, no. 4, pp. 641–666, Jul. 2013, doi: 10.1002/rob.21464.

3. J. Ma, Y. Ma, and C. Li, "Infrared and visible image fusion methods and applications: A survey," *Information Fusion*, vol. 45, pp. 153–178, Jan. 2019, doi: 10.1016/j.inffus.2018.02.004.

4. J. Qi, D. E. Abera, M. N. Fanose, L. Wang, and J. Cheng, "A deep learning and image enhancement based pipeline for infrared and visible image fusion," *Neurocomputing*, vol. 578, p. 127353, Apr. 2024, doi: 10.1016/j.neucom.2024.127353.

5. T. L. Narayana et al., "Advances in real time smart monitoring of environmental parameters using IoT and sensors," *Heliyon*, vol. 10, no. 7, p. e28195, Apr. 2024, doi: 10.1016/j.heliyon.2024.e28195.

6. Y. Liu, X. Chen, R. K. Ward, and J. Wang, "Image Fusion with Convolutional Sparse Representation," *IEEE Signal Processing Letters*, vol. 23, no. 12, pp. 1882–1886, Dec. 2016, doi: 10.1109/LSP.2016.2618776.

7. Y. Liu, X. Chen, H. Peng, and Z. Wang, "Multi-focus image fusion with a deep convolutional neural network," *Information Fusion*, vol. 36, pp. 191–207, Jul. 2017, doi: 10.1016/j.inffus.2016.12.001.

8. H. Li, X. J. Wu, and J. Kittler, "Infrared and Visible Image Fusion using a Deep Learning Framework," *Proc. International Conference on Pattern Recognition (ICPR)*, 2018, pp. 2705–2710, doi: 10.1109/ICPR.2018.8546006.

9. K. R. Prabhakar, V. S. Srikar, and R. V. Babu, "DeepFuse: A Deep Unsupervised Approach for Exposure Fusion with Extreme Exposure Image Pairs," *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 4724–4732, doi: 10.1109/ICCV.2017.505.

10. Y. Ma, J. Chen, C. Chen, F. Fan, and J. Ma, "Infrared and visible image fusion using total variation model," *Neurocomputing*, vol. 202, pp. 12–19, Aug. 2016, doi: 10.1016/j.neucom.2016.03.009.

11. H. Li, X. J. Wu, and T. S. Durrani, "Infrared and visible image fusion with ResNet and zero-phase component analysis," *Infrared Physics & Technology*, vol. 102, p. 103039, Nov. 2019, doi: 10.1016/j.infrared.2019.103039.

12. H. Xu, J. Ma, Z. Le, J. Jiang, and X. Guo, "FusionDN: A Unified Densely Connected Network for Image Fusion," *Proc. AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, pp. 12484–12491, 2020, doi: 10.1609/aaai.v34i07.6936.

13. Y. Zhang et al., "IFCNN: A general image fusion framework based on convolutional neural network," *Information Fusion*, vol. 54, pp. 99–118, Feb. 2020, doi: 10.1016/j.inffus.2019.07.011.

14. Y. Chen, H. Xie, and H. Shin, "Multi-layer fusion techniques using a CNN for multispectral pedestrian detection," *IET Computer Vision*, vol. 12, no. 8, pp. 1179–1187, Dec. 2018, doi: 10.1049/iet-cvi.2018.5315.

15. R. Hou et al., "VIF-Net: An Unsupervised Framework for Infrared and Visible Image Fusion," *IEEE Transactions on Computational Imaging*, vol. 6, pp. 640–651, Jan. 2020, doi: 10.1109/TCI.2020.2965304.

16. X. Zhang and Y. Demiris, "Visible and Infrared Image Fusion Using Deep Learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 10535–10554, Aug. 2023, doi: 10.1109/TPAMI.2023.3261282.

17. B. Meher et al., "Visible and infrared image fusion using an efficient adaptive transition region extraction technique," *Engineering Science and Technology, an International Journal*, vol. 29, p. 101037, May 2022, doi: 10.1016/j.jestch.2021.06.017.

18. N. T. N. Sulthana and S. Joseph, "Infrared and visible image: Enhancement and fusion using adversarial network," *AIP Conference Proceedings*, vol. 3037, no. 1, Apr. 2024, doi: 10.1063/5.0196355.