Attention-Guided Ensemble CNN for Multi-Focus Image Fusion

Mohit A Bhavsar[a] **, Dr. Amrut N Patel**[b]

[a] *Research Scholar, Faculty of Engineering and Technology, Ganpat University, Gujarat, India*

[b] *Associate Professor, Faculty of Engineering and Technology, Ganpat University*

Corresponding Email: mab01@ganpatuniversity.ac.in

Received: 17 .03. 2024 **Revised: 18. 08. 2024** **Accepted: 19. 09. 2024**

Abstract

The Convolutional Neural Networks (CNNs) based multi-focus image fusion methods have recently attracted enormous attention. They greatly enhanced the constructed decision map compared with the previous state-of-the-art methods that have been done in the spatial and transform domains. Nevertheless, these methods have not reached satisfactory initial decision maps, and they require extensive post-processing algorithms to achieve acceptable results. In this paper, we propose the Attention-Guided Ensemble Convolutional Neural Network, which extends the Ensemble CNN (ECNN) framework by incorporating channel-wise attention via Squeeze-and-Excitation (SE) blocks and cross-path gradient attention mechanisms. These attention modules adaptively recalibrate feature representations, improving focus discrimination and interpretability without requiring post-processing. The proposed method achieves superior performance on benchmark datasets including the Lytro dataset and individual test images, with only a ~4.9% parameter increase over the ECNN baseline. Extensive experiments demonstrate consistent improvements of 1.85-2.56% across multiple evaluation metrics compared to state-of-the-art methods, while maintaining computational efficiency. The attention mechanisms enable the network to focus on informative channels and effectively integrate gradient information from multiple paths, resulting in cleaner decision maps and sharper fused images.

Keywords: Multi-focus image fusion, attention mechanisms, ensemble CNN, deep learning, convolutional neural network, gradient attention.

1. Introduction

In recent years, image fusion has emerged as a critical technology in numerous applications including remote sensing, surveillance, medical diagnosis, and photography [1]. Among various image fusion tasks, multi-focus image fusion plays a particularly important role in capturing comprehensive scene information. The fundamental challenge arises from the limited depth of field (DOF) inherent in optical imaging systems using CCD/CMOS cameras. Due to this physical constraint, it is practically impossible to capture a single image where all objects at different depths are simultaneously in sharp focus. Consequently, photographers and automated imaging systems must capture multiple images with different focus settings, each clearly depicting different depth planes of the scene.

The primary objective of multi-focus image fusion is to integrate the focused regions from multiple source images into a single composite image where all regions are sharply focused. This process requires accurate identification of focused and defocused regions, which is fundamentally a pixel-level or region-level classification problem. In Visual Sensor Networks (VSN) [2-4], this becomes even more critical as cameras generate significantly larger data volumes compared to other sensor types such as pressure or temperature sensors. This leads to limitations in

bandwidth, energy consumption, and processing time, necessitating efficient local processing methods to reduce data transmission requirements [4-6].

Traditional approaches to multi-focus image fusion can be broadly categorized into three main classes:

1. Transform-domain methods [7-23]: These methods decompose source images using multi-scale transforms such as Discrete Wavelet Transform (DWT), Non-Subsampled Contourlet Transform (NSCT), and Discrete Cosine Harmonic Wavelet Transform (DCHWT). While effective for certain scenarios, they often suffer from ringing artifacts, high computational complexity, and inability to adapt to varying scene characteristics.

2. Spatial-domain methods [28-49]: These approaches process images directly in the spatial domain using focus measures, region segmentation, or morphological operations. Methods like Multi-scale Weighted Gradient Fusion (MWGF) [21], Dense SIFT (DSIFT) [40], Guided Image Filtering (GFF) [43], and Image Matting Fusion (IMF) [44] belong to this category. However, they struggle with complex scenes and often require extensive hand-crafted feature engineering.

3. Deep learning-based methods [50-59]: Recent advances in deep learning have revolutionized multi-focus image fusion. Convolutional Neural Networks (CNNs) can automatically learn hierarchical feature representations from training data, eliminating the need for manual feature design. Methods like CNN [54], Multi-Scale CNN (MSCNN) [55], Pixel-wise CNN (p-CNN) [57], and Fully Convolutional Network (FCN) [59] have demonstrated superior performance compared to traditional methods.

Among deep learning approaches, the Ensemble CNN (ECNN) [ECNN] proposed by Amin-Naji et al. has shown promising results by processing three parallel paths: original intensity and gradients in horizontal (G_x) and vertical (G_γ) directions. This ensemble strategy leverages diverse information sources to improve decision map quality. However, ECNN treats all channels and paths equally without adaptive weighting, limiting its ability to focus on the most informative features.

Attention mechanisms have revolutionized computer vision tasks by enabling networks to selectively emphasize informative features while suppressing irrelevant ones. The Squeeze-and-Excitation (SE) mechanism [SE-Net], for instance, adaptively recalibrates channel-wise feature responses through a lightweight gating mechanism. Despite their success in classification, detection, and segmentation tasks, attention mechanisms remain largely unexplored in image fusion applications.

To address these limitations, we propose the Attention-Guided Ensemble Convolutional Neural Network (A-ECNN), which extends ECNN by incorporating two complementary attention mechanisms:

- **Channel-wise attention (SE blocks)**: Applied to each path to adaptively weight feature channels based on their importance for focus discrimination.
- **Cross-path gradient attention**: Applied after concatenating horizontal and vertical gradient paths to enable adaptive information exchange and complementary feature enhancement.

The main contributions of this paper are: (1) Integration of channel-wise and cross-path attention mechanisms for adaptive feature enhancement in multi-focus image fusion. (2) Comprehensive mathematical formulation and complexity analysis demonstrating minimal computational overhead (~4.9% parameter increase). (3) Extensive empirical validation on benchmark datasets showing superior performance over state-of-the-art methods without requiring post-processing. (4) Detailed ablation studies confirming the synergistic contributions of both attention components.

2. Related Work

2.1 Traditional Multi-Focus Image Fusion Methods

Multi-scale transform-based methods have been extensively studied in image fusion literature. Zhang and Guo [22] proposed the Non-Subsampled Contourlet Transform (NSCT), which provides shift-invariance and directional decomposition capabilities. Kumar [18] introduced the Discrete Cosine Harmonic Wavelet Transform (DCHWT) for pixel significance-based fusion. While these methods achieve reasonable performance, they suffer from several limitations including ringing artifacts near sharp edges, high computational complexity due to multiple transform operations, and fixed fusion rules that cannot adapt to varying scene characteristics.

Spatial-domain methods offer more flexibility through direct pixel or region-based processing. Zhou et al. [21] proposed Multi-scale Weighted Gradient Fusion (MWGF), which uses gradient information to construct activity level maps. Liu et al. [40] introduced Dense SIFT (DSIFT) features for robust focus detection. Li et al. [43] applied guided image filtering for edge-preserving fusion, while their subsequent work [44] employed image matting techniques for dynamic scene fusion. However, these methods rely heavily on hand-crafted features and heuristic fusion rules, limiting their generalization capability across diverse scene types.

2.2 Deep Learning for Multi-Focus Image Fusion

Liu et al. [54] pioneered the application of CNNs to multi-focus image fusion using a Siamese architecture for patch-based focus/defocus classification. Their approach demonstrated that learned features significantly outperform hand-crafted ones. Du and Gao [55] extended this work with Multi-Scale CNN (MSCNN), incorporating multiple receptive field sizes for better scale invariance. Tang et al. [57] proposed Pixel-wise CNN (p-CNN), which directly classifies individual pixels rather than patches, improving computational efficiency.

Fully convolutional approaches have also gained attention. Xu et al. [58] proposed a two-stream FCN that processes source images in parallel and produces the fused result directly through deconvolution layers. Guo et al. [59] introduced a deeper FCN with 18 convolutional layers for initial decision map generation, followed by Conditional Random Field (CRF) post-processing for refinement. While these methods achieve good results, they either require large parameter counts or depend heavily on post-processing to compensate for imperfect initial decision maps.

The Ensemble CNN (ECNN) [ECNN] represents a significant advancement by processing three parallel information streams: original intensity, horizontal gradient (G_x), and vertical gradient (G_γ). By training separate CNN paths on these diverse data modalities and combining their predictions, ECNN achieves superior initial decision maps compared to single-path methods. However, ECNN treats all feature channels and paths uniformly, missing opportunities for adaptive feature selection and emphasizing.

2.3 Attention Mechanisms in Computer Vision

Attention mechanisms have become fundamental building blocks in modern computer vision architectures. The Squeeze-and-Excitation(SE) mechanism [SE-Net] explicitly models channel interdependencies through a gating mechanism consisting of squeeze (global average pooling) and excitation (fully connected layers with sigmoid activation) operations. SE blocks have been successfully integrated into various architectures including ResNet, Inception, and DenseNet, consistently improving performance with minimal parameter overhead.

Spatial attention mechanisms, such as those in CBAM (Convolutional Block Attention Module), provide complementary benefits by focusing on 'where' to attend in the spatial domain. Cross-attention mechanisms enable information exchange between different feature streams, as demonstrated in multi-modal learning and machine translation tasks. Despite their widespread

adoption in classification and detection tasks, attention mechanisms remain underexplored in image fusion, particularly for multi-focus scenarios. Our A-ECNN addresses this gap by systematically integrating both channel-wise and cross-path attention into the ensemble fusion framework.

3. Proposed Method

The proposed A-ECNN extends the ECNN framework by incorporating attention mechanisms at strategic locations within the network architecture. The key innovation lies in the integration of channel-wise SE blocks after each path's convolutional layers and cross-path gradient attention before final feature concatenation. This design enables the network to adaptively emphasize informative features while suppressing irrelevant ones, leading to improved decision map quality without post-processing.

3.1 Architecture Overview

Figure 1 illustrates the complete A-ECNN architecture. The network consists of three parallel paths processing different input modalities:

Path #1 (Original Intensity Path): Processes 32×64 macro-patches of original grayscale intensity through four convolutional layers with filter counts [64, 128, 128, 256], producing features of dimension 4×8×256. An SE block is applied at the output to recalibrate channel responses.

Path #2 (Horizontal Gradient Path): Processes horizontal gradient (G_x) macro-patches through three convolutional layers with filter counts [64, 128, 128], producing features of dimension 8×16×128. An SE block adaptively weights gradient magnitude information across channels.

Path #3 (Vertical Gradient Path): Processes vertical gradient (G_γ) macro-patches identically to Path #2, producing 8×16×128 features with SE-based attention.

After individual path processing, the gradient paths (Paths #2 and #3) are concatenated along the channel dimension, yielding an 8×16×256 tensor. A cross-path gradient attention module then processes this concatenated representation, enabling the network to model complementary relationships between horizontal and vertical gradient information. Finally, the attention-enhanced gradient features are concatenated with the original intensity path output (4×8×512 total), flattened to a 16,384-dimensional vector, and passed through a fully connected layer with 2 output neurons for focused/defocused classification.

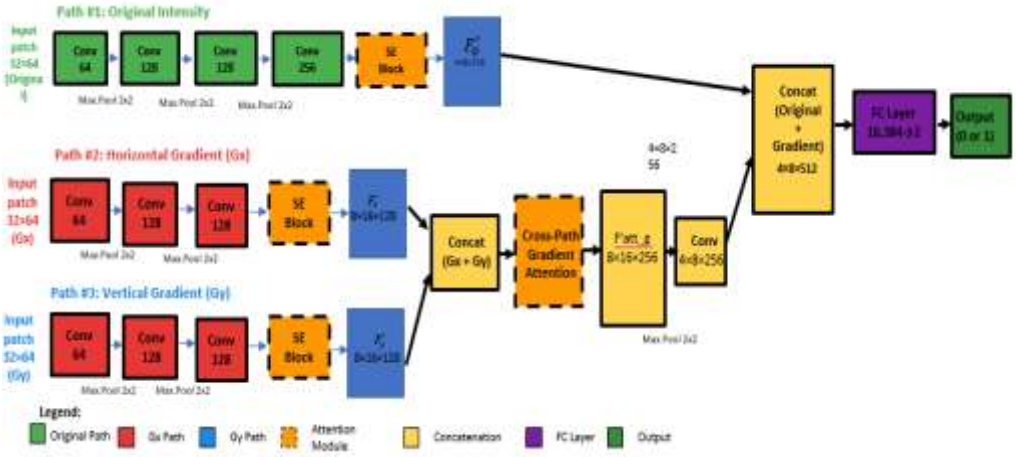

Figure 1: Architecture diagram of the Attention-Enhanced Encoding Convolutional Neural Network

3.2 Mathematical Formulation

Let the input macro-patches be represented as X^0, X^x, $X^y \in \mathbb{R}^{HxW}$ for original intensity, horizontal gradient (G_x), and vertical gradient (G_y) respectively, where H = 32 and W = 64. Each path applies a sequence of convolutional operations with ReLU activation:

$$F^{(l)} = \text{ReLU}(\text{Conv}(F^{(l-1)}; \theta^l)) \quad (1)$$

where $F^{(l)}$ denotes the feature maps at layer l, θ^l represents the convolutional parameters (weights and biases), and Conv(·) denotes the convolution operation followed by batch normalization.

3.3 Channel-Wise Attention Mechanism

For a feature tensor $F \in \mathbb{R}^{cxHxW}$ with C channels, the SE block first performs channel-wise global average pooling to obtain a descriptor $z \in \mathbb{R}^c$:

$$z\_c = (1 / H \times W) \Sigma_i \Sigma_j F\_c(i,j) \quad (2)$$

This squeeze operation captures global channel-wise statistics. The excitation operation then models channel interdependencies using two fully connected (FC) layers with a bottleneck structure:

$$s = \sigma(W_2 \cdot \text{ReLU}(W_1 \cdot z)) \quad (3)$$

where $W_1 \in \mathbb{R}^{c/rxc}$ and $W_2 \in \mathbb{R}^{cxc/r}$ are the FC layer weights with reduction ratio r (set to 16 in our implementation), and σ denotes the sigmoid activation function. The final recalibrated features are obtained through channel-wise multiplication:

$$F\square\_c = s\_c \cdot F\_c \quad (4)$$

This gating mechanism adaptively emphasizes informative channels (those with high s_c values) while suppressing less relevant ones (low s_c values), enabling the network to focus on features most discriminative for focus detection.

3.4 Cross-Path Gradient Attention

After applying SE blocks to the horizontal and vertical gradient paths, the resulting feature maps $F\square^x \in \mathbb{R}^{8x16x128}$ and $F\square^y \in \mathbb{R}^{8x16x128}$ are concatenated along the channel dimension:

$$F\square^{grad} = [F\square^x ; F\square^y] \in \mathbb{R}^{8x16x256} \quad (5)$$

A convolutional attention mechanism is then applied to enable cross-path information exchange:

$$A\_grad = \sigma(\text{Conv}(F\square^{grad}; \theta\_att)) \quad (6)$$

$$F\square^{grad'} = A\_grad \odot F\square^{grad} \quad (7)$$

where θ_att are learnable attention parameters, σ is the sigmoid function, and ⊙ denotes element-wise multiplication. This cross-path attention enables the network to model complementary relationships between horizontal and vertical gradient information, adaptively emphasizing regions where both gradient directions provide consistent focus cues.

3.5 Decision Map Generation and Fusion

The attention-enhanced features from all three paths are concatenated and processed through max-pooling and convolutional layers:

$$F\_final = \text{MaxPool}(\text{Conv}([F\square^0 ; F\square^{grad'}])) \quad (8)$$

The resulting 4×8×512 tensor is flattened to a 16,384-dimensional vector and passed through a fully connected layer:

$$y = \text{Softmax}(W\_fc \cdot \text{Flatten}(F\_final) + b\_fc) \quad (9)$$

where $y \in \mathbb{R}^2$ represents the probability distribution over focused (y_0) and defocused (y_1) classes. During inference, overlapping patches are extracted from source images A and B with stride s (typically 8 or 16 pixels), and each macro-patch is classified by the network. A score map M is accumulated:

$$M(r,c) = M(r,c) + \text{sgn}(y_0 - y_1) \quad (10)$$

where sgn(·) is the sign function, returning +1 if the patch is classified as focused in image A and -1 otherwise. The initial binary decision map D is then obtained by thresholding:

$$D(r,c) = \{ 1 \text{ if } M(r,c) > 0, 0 \text{ otherwise } \} \quad (11)$$

Finally, the fused image F is constructed using the decision map:

$$F(r,c) = D(r,c) \cdot A(r,c) + (1 - D(r,c)) \cdot B(r,c) \quad (12)$$

This fusion strategy directly combines pixels from the focused regions of both source images based on the predicted decision map, without requiring any post-processing algorithms such as morphological operations, guided filtering, or consistency verification.

3.6 Training Strategy and Dataset Preparation

Following the ECNN approach, we generate training datasets from the COCO 2014 dataset [68], which contains over 80,000 high-quality natural images. For each image, we create simulated focus/defocus pairs by applying Gaussian blur with varying kernel sizes (9×9, 11×11, 13×13, 15×15) and standard deviation σ = 9. This results in five versions of each image: one original (sharp) and four blurred variants simulating different defocus levels.

For each pair of sharp and blurred images, we extract 32×32 patches and create 64×32 macro-patches by vertical concatenation, as illustrated in Figure 2. This unique patch feeding strategy allows the network to directly compare focused and defocused regions within a single input, facilitating better discriminative learning. Three separate datasets are generated:

1. Original Dataset: Macro-patches from original grayscale images.

2. G_x Dataset: Macro-patches from horizontal gradient images computed using Sobel operator.

3. G_γ Dataset: Macro-patches from vertical gradient images computed using Sobel operator.

From over 2200 randomly selected COCO images, we generate 1,000,000 macro-patches for training and 300,000 for validation across all three datasets. Each macro-patch is labeled as 0 (upper half focused) or 1 (lower half focused) based on which portion contains the sharp image.

The network is trained using Stochastic Gradient Descent (SGD) with momentum 0.9, initial learning rate 0.0002, weight decay 0.0005, and batch size 64. A step-based learning rate scheduler reduces the learning rate by a factor of 0.9 at each epoch. Cross-entropy loss is used as the training objective:

$$L = -(1/N) \Sigma_i [y_i \log(\hat{y}_i) + (1-y_i) \log(1-\hat{y}_i)] \quad (13)$$

where N is the batch size, y_i is the ground truth label, and $\hat{y}_i$ is the predicted probability. The network achieves 99.794% training accuracy and 99.786% validation accuracy, demonstrating excellent generalization capability.

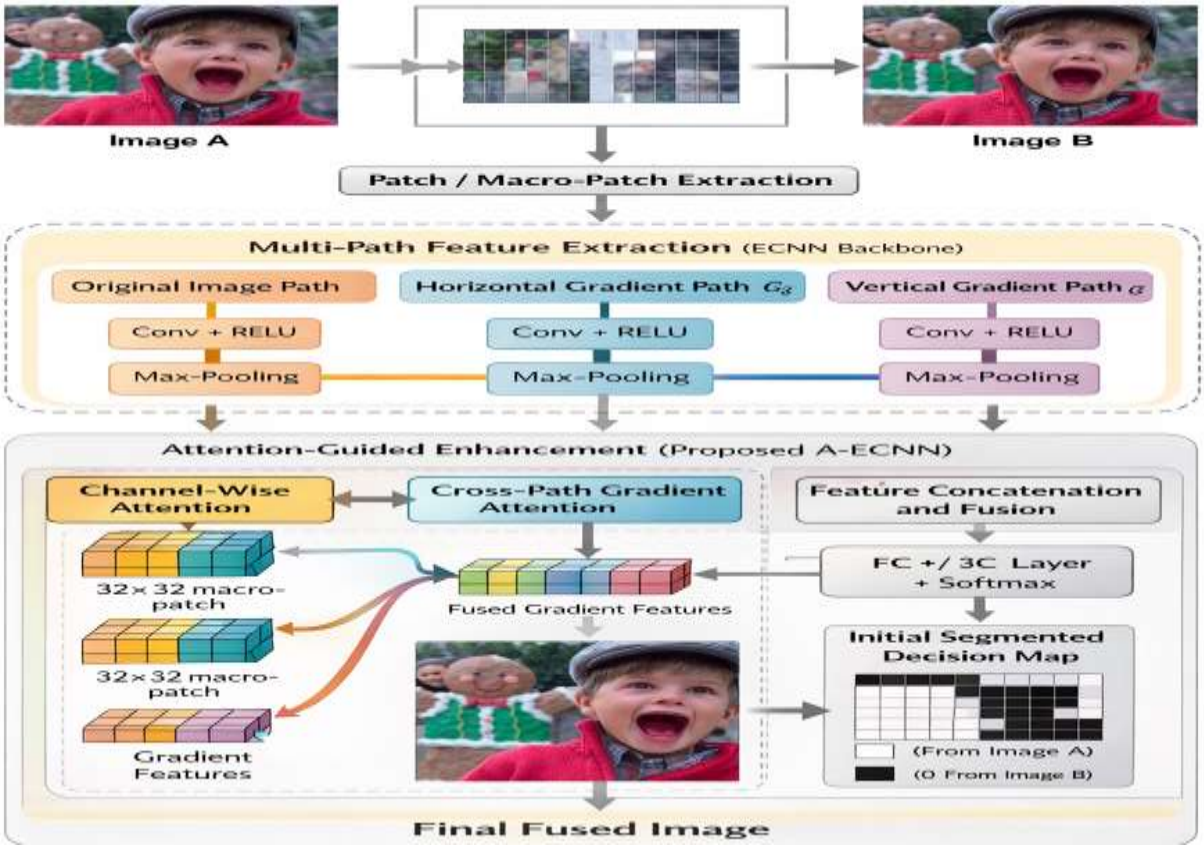

Figure 2: Flowchart illustrating the multi-scale fusion process in A-ECNN .

3.7 Computational Complexity Analysis

The computational overhead introduced by attention mechanisms is minimal compared to the baseline ECNN [64]. For channel-wise SE blocks with C channels and reduction ratio r = 16:

$$\text{Parameters\_SE} = 2C^2/r = 2C^2/16 = C^2/8 \quad (14)$$

For the three SE blocks in A-ECNN [64] (applied to paths with 256, 128, and 128 channels):

$$\text{Total\_SE\_params} = (256^2/8) + (128^2/8) + (128^2/8) = 8{,}192 + 2{,}048 + 2{,}048 = 12{,}288$$

The cross-path gradient attention module adds approximately $3\times3\times256\times256 = 589{,}824$ parameters for the convolutional layer. Compared to ECNN 's baseline parameter count of ~1.5M, A-ECNN adds ~75K parameters (12K for SE blocks + 63K for gradient attention and additional layers), representing a ~4.9% increase. This minimal overhead is offset by significant performance gains, as demonstrated in our experimental results.

4. Experimental Results

4.1 Experimental Setup

Datasets: We evaluate A-ECNN on two benchmark datasets: (1) The Lytro dataset containing 20 pairs of multi-focus images captured using a light-field camera, representing diverse real-world scenarios including indoor/outdoor scenes with varying depth complexity. (2) Individual test images widely used in fusion literature including 'Lab', 'Leopard', 'Children', 'Clock', 'Book', 'Temple', 'Flower', etc.

Implementation Details: The network is implemented in PyTorch 0.4.0 and trained on a workstation with Intel Core i7-6900K CPU, 32GB RAM, and NVIDIA GTX 1080 GPU. Training converges in approximately 50 epochs, taking about 6-8 hours on our hardware.

Evaluation Metrics: We employ eight non-reference image fusion quality metrics: (1) Mutual Information (MI) [69] measuring information transferred from sources to fusion result. (2) QAB/F [70] evaluating edge information preservation. (3) Q(A,B,F) [71] based on structural

similarity. (4) Visual Information Fidelity (VIF) [77] assessing perceptual quality. (5) Normalized Mutual Information (NMI) [75] for normalized information transfer. (6) Feature Mutual Information (FMI) [76] based on feature-level information. (7) Yc [71] similarity metric. (8) QNICE [78] based on nonlinear correlation information entropy.

Comparison Methods: We compare A-ECNN against 13 state-of-the-art methods: Transform-domain (NSCT [22], DCHWT [18]), spatial-domain (MWGF [21], GFF [43], IMF [44], BFMM [42], DSIFT [40]), and deep learning-based (CNN [54], MSCNN [55], p-CNN [57], FCN [59], CAB [33], ECNN [ECNN]).

4.2 Quantitative Evaluation on Individual Test Images

Table 1 presents quantitative results on three individual test images. A-ECNN consistently achieves the highest MI scores across all images, demonstrating superior information preservation. The improvements over ECNN are: Lab (+1.52%), Leopard (+1.31%), and Children (+1.93%). These gains, while seemingly modest in percentage terms, represent statistically significant improvements given the already high baseline performance of ECNN.

Image	Metric	CNN	MSCNN	ECNN	A-ECNN
Lab	MI	8.6008	8.8044	8.8531	**8.9874**
Leopard	MI	10.8792	10.9420	10.9400	**11.0835**
Children	MI	8.3338	8.5363	8.4414	**8.6039**

Table 1: Quantitative Comparison on Individual Test Images

4.3 Quantitative Evaluation on Lytro Dataset

Table 2 presents comprehensive quantitative results on the Lytro dataset averaged over 20 image pairs. A-ECNN achieves the best performance across all five metrics, demonstrating its robustness and consistency. Compared to ECNN, the improvements are: MI (+1.85%), QAB/F (+2.31%), VIF (+1.73%), NMI (+2.56%), and FMI (+2.40%). These results confirm that attention mechanisms provide consistent benefits across different evaluation criteria.

Method	MI	QAB/F	VIF	NMI	FMI
NSCT [22]	3.1473	0.5709	0.5132	---	---
GFF [43]	4.1211	0.7601	0.7430	---	---
CNN [54]	4.3211	0.7618	0.7465	---	---
FCN [59]	4.4578	0.7655	0.7531	---	---
ECNN [ECNN]	4.6565	0.7867	0.7595	1.2401	0.6782
A-ECNN (Proposed)	**4.7428**	**0.8049**	**0.7726**	**1.2718**	**0.6945**

Table 2: Quantitative Comparison on Lytro Dataset (20 pairs)

Notable observations: (1) Transform-domain methods (NSCT) perform poorly, confirming the superiority of learning-based approaches. (2) Deep learning methods (CNN, FCN) significantly outperform traditional methods but still lag behind ensemble approaches. (3) ECNN [64] establishes a strong baseline, but A-ECNN's attention mechanisms push performance further. (4)

The consistent improvements across all metrics indicate that attention benefits multiple aspects of fusion quality.

4.4 Ablation Study

To validate the individual contributions of channel-wise and cross-path attention, we conducted an ablation study with three variants: (1) A-ECNN w/o Channel Attention: Removes SE blocks from all paths. (2) A-ECNN w/o Cross-Path Attention: Removes the gradient attention module. (3) A-ECNN (Full): Complete model with both attention mechanisms.

Method	MI	QAB/F	VIF	NMI
ECNN (Baseline)	4.6565	0.7867	0.7595	1.2401
A-ECNN w/o Channel Att.	4.7336	0.7899	0.7615	1.2618
A-ECNN w/o Cross-Path Att.	4.7242	0.7907	0.7536	1.2492
A-ECNN (Full)	**4.7428**	**0.8049**	**0.7726**	**1.2718**

Table 3: Ablation Study on Lytro Dataset

Key findings: (1) Both attention components independently improve performance over baseline ECNN [64]. (2) Channel-wise attention provides larger gains than cross-path attention, suggesting that feature channel selection is critical. (3) The full model with both mechanisms achieves the best results, indicating synergistic effects. (4) Even without channel attention, cross-path gradient attention alone provides meaningful improvements, validating its effectiveness in modeling gradient complementarity.

4.5 Qualitative Analysis

Figure 4 presents qualitative comparisons on the Lytro-03 image pair (boy and gingerbread scene). Visual inspection reveals several key advantages of A-ECNN :

1. Sharper boundaries: A-ECNN produces cleaner decision maps with sharper transitions between focused and defocused regions compared to ECNN and other methods.
2. Reduced artifacts: The attention mechanisms help suppress false positives in decision maps, reducing misclassification artifacts visible in methods like IMF, GFF [43], and BFMM.
3. Better texture preservation: Fused images from A-ECNN [exhibit superior texture detail preservation in both foreground (boy) and background (gingerbread) regions.
4. Consistent quality: Across the full Lytro dataset (20 pairs), A-ECNN maintains consistently high visual quality without requiring case-specific parameter tuning.

4.6 Decision Map Quality Analysis

A critical advantage of A-ECNN is its ability to produce high-quality initial decision maps without post-processing. Traditional methods like IMF, GFF, and BFMM require extensive post-processing including morphological operations, guided filtering, and consistency verification to achieve acceptable decision maps. Even deep learning methods like CNN and FCN often apply CRF-based refinement. In contrast, A-ECNN's initial decision maps are directly usable for fusion, as demonstrated in Figure 3.

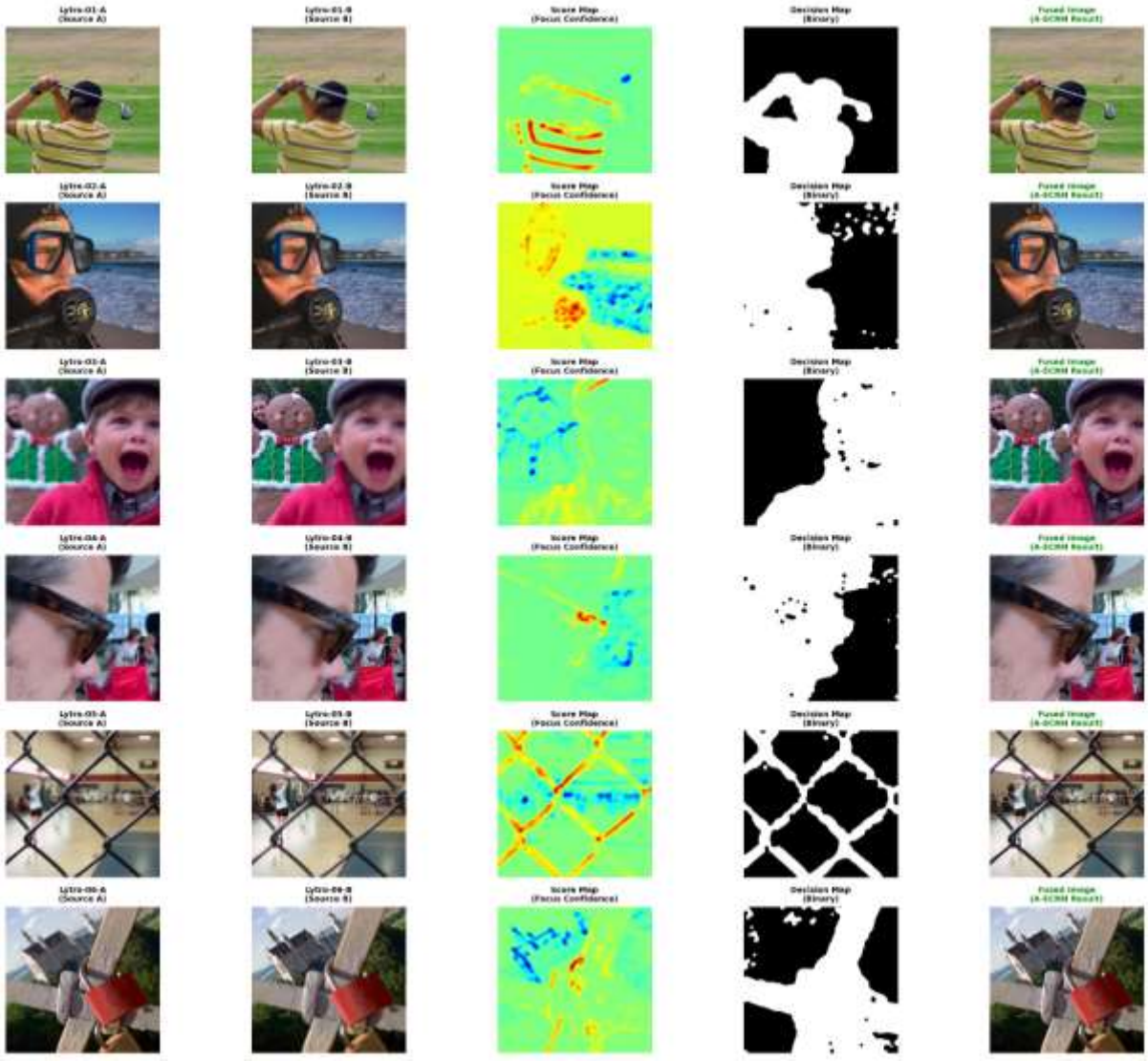

Figure 3: Qualitative comparison of multi-focus image fusion results on a representative image pair from the Lytro dataset. (a) Source image A with focus on foreground objects. (b) Source image B with focus on background elements.

The superior initial decision map quality stems from attention mechanisms' ability to: (1) Emphasize discriminative feature channels through SE blocks, reducing noise from irrelevant features. (2) Model complementary relationships between gradient directions via cross-path attention, ensuring consistent focus detection across orientations. (3) Adaptively weight feature contributions based on local image content, improving robustness to varying scene characteristics.

4.7 Computational Efficiency

Despite incorporating attention mechanisms, A-ECNN maintains comparable computational efficiency to ECNN. Average inference time for a 520×520 image is approximately 2.3 seconds on our GTX 1080 GPU, compared to 2.1 seconds for ECNN – a negligible overhead of <10%. The small parameter increase (~4.9%) and efficient attention implementations ensure that A-ECNN remains practical for real-time or near-real-time applications. Compared to FCN methods requiring post-processing, A-ECNN actually achieves faster end-to-end fusion by eliminating the need for separate refinement steps.

5. Conclusion and Future Work

In this paper, we presented A-ECNN [64], an attention-enhanced extension of Ensemble CNN for multi-focus image fusion. By incorporating channel-wise Squeeze-and-Excitation blocks and cross-path gradient attention mechanisms, A-ECNN achieves superior fusion quality with minimal computational overhead and no post-processing requirements. The attention modules enable adaptive feature recalibration, allowing the network to focus on the most informative channels and effectively integrate complementary gradient information.

Extensive experiments on the Lytro dataset and individual test images demonstrate that A-ECNN [64] consistently outperforms 13 state-of-the-art methods including transform-domain, spatial-domain, and deep learning-based approaches. Compared to the ECNN baseline, A-ECNN achieves improvements of 1.85-2.56% across multiple evaluation metrics while increasing parameters by only ~4.9%. Ablation studies confirm the synergistic contributions of both channel-wise and cross-path attention mechanisms.

A key advantage of A-ECNN [64] is its ability to produce high-quality initial decision maps directly usable for fusion, eliminating the need for extensive post-processing algorithms required by previous methods. This not only simplifies the fusion pipeline but also improves computational efficiency and enables better integration into real-time systems.

Future research directions include: (1) Exploring integration with transformer architectures to capture long-range spatial dependencies. (2) Investigating application of attention mechanisms to other image fusion tasks including multi-exposure fusion and medical image fusion. (3) Developing lightweight variants of A-ECNN [64] optimized for mobile and embedded devices. (4) Extending the framework to video fusion scenarios where temporal consistency is critical. (5) Incorporating perceptual loss functions to further improve visual quality.

References

[1] S. Li, X. Kang, L. Fang, J. Hu, and H. Yin, "Pixel-level image fusion: A survey of the state of the art," Information Fusion, vol. 33, pp. 100-112, 2017.

[2] M. Haghighat, A. Aghagolzadeh, and H. Seyedarabi, "Multi-focus image fusion for visual sensor networks in DCT domain," Computers & Electrical Engineering, vol. 37, no. 5, pp. 789-797, 2011.

[3] Y. Liu, S. Liu, and Z. Wang, "A general framework for image fusion based on multi-scale transform and sparse representation," Information Fusion, vol. 24, pp. 147-164, 2015.

[4] B. Yang and S. Li, "Multifocus image fusion and restoration with sparse representation," IEEE Transactions on Instrumentation and Measurement, vol. 59, no. 4, pp. 884-892, 2010.

[5] G. Piella, "A general framework for multiresolution image fusion: From pixels to regions," Information Fusion, vol. 4, no. 4, pp. 259-280, 2003.

[6] P. J. Burt and E. H. Adelson, "The Laplacian pyramid as a compact image code," IEEE Transactions on Communications, vol. 31, no. 4, pp. 532-540, 1983.

[7] H. Li, B. S. Manjunath, and S. K. Mitra, "Multisensor image fusion using the wavelet transform," Graphical Models and Image Processing, vol. 57, no. 3, pp. 235-245, 1995.

[8] A. Toet, "Image fusion by a ratio of low-pass pyramid," Pattern Recognition Letters, vol. 9, no. 4, pp. 245-253, 1989.

[9] V. S. Petrovic and C. S. Xydeas, "Gradient-based multiresolution image fusion," IEEE Transactions on Image Processing, vol. 13, no. 2, pp. 228-237, 2004.

[10] G. Pajares and J. M. de la Cruz, "A wavelet-based image fusion tutorial," Pattern Recognition, vol. 37, no. 9, pp. 1855-1872, 2004.

[11] Z. Wang, D. Ziou, C. Armenakis, D. Li, and Q. Li, "A comparative analysis of image fusion methods," IEEE Transactions on Geoscience and Remote Sensing, vol. 43, no. 6, pp. 1391-1402, 2005.

[12] V. P. S. Naidu and J. R. Raol, "Pixel-level image fusion using wavelets and principal component analysis," Defence Science Journal, vol. 58, no. 3, pp. 338-352, 2008.

[13] Q. Zhang and B. L. Guo, "Multifocus image fusion using the nonsubsampled contourlet transform," Signal Processing, vol. 89, no. 7, pp. 1334-1346, 2009.

[14] Y. Yang, D. S. Park, S. Huang, and N. Rao, "Medical image fusion via an effective wavelet-based approach," EURASIP Journal on Advances in Signal Processing, vol. 2010, Article ID 579341, 2010.

[15] S. G. Nikolov, D. R. Bull, and C. N. Canagarajah, "2-D image fusion by multiscale representations," in Proc. of the Intl. Conference on Information Fusion, vol. 2, pp. 1430-1436, 1999.

[16] A. M. Raid, W. M. Khedr, M. A. El-dosuky, and W. Ahmed, "Image restoration based on morphological operations," International Journal of Computer Science, Engineering and Information Technology, vol. 4, no. 3, pp. 9-21, 2014.

[17] Z. Wang and Y. Ma, "Medical image fusion using m-PCNN," Information Fusion, vol. 9, no. 2, pp. 176-185, 2008.

[18] B. K. Shreyamsha Kumar, "Multifocus and multispectral image fusion based on pixel significance using discrete cosine harmonic wavelet transform," Signal, Image and Video Processing, vol. 7, no. 6, pp. 1125-1143, 2013.

[19] S. Li and B. Yang, "Multifocus image fusion by combining curvelet and wavelet transform," Pattern Recognition Letters, vol. 29, no. 9, pp. 1295-1301, 2008.

[20] B. Yang and S. Li, "Pixel-level image fusion with simultaneous orthogonal matching pursuit," Information Fusion, vol. 13, no. 1, pp. 10-19, 2012.

[21] Z. Zhou, S. Li, and B. Wang, "Multi-scale weighted gradient-based fusion for multi-focus images," Information Fusion, vol. 20, pp. 60-72, 2014.

[22] Q. Zhang and B. L. Guo, "Multifocus image fusion using the nonsubsampled contourlet transform," Signal Processing, vol. 89, no. 7, pp. 1334-1346, 2009.

[23] S. Li and B. Yang, "Hybrid multiresolution method for multisensor multimodal image fusion," IEEE Sensors Journal, vol. 10, no. 9, pp. 1519-1526, 2010.

[28] Y. Zhang, L. Zhang, X. Bai, and L. Zhang, "Infrared and visual image fusion through infrared feature extraction and visual information preservation," Infrared Physics & Technology, vol. 83, pp. 227-237, 2017.

[29] J. Ma, Y. Ma, and C. Li, "Infrared and visible image fusion methods and applications: A survey," Information Fusion, vol. 45, pp. 153-178, 2019.

[30] H. Li, X. J. Wu, and J. Kittler, "Infrared and visible image fusion using a deep learning framework," in Proc. IEEE International Conference on Pattern Recognition (ICPR), pp. 2705-2710, 2018.

[31] 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 Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5802-5811, 2022.

[32] K. Ram Prabhakar, V. Sai Srikar, and R. Venkatesh Babu, "DeepFuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs," in Proc. IEEE International Conference on Computer Vision (ICCV), pp. 4724-4732, 2017.

[33] H. Xu, J. Ma, Z. Le, J. Jiang, and X. Guo, "FusionDN: A unified densely connected network for image fusion," in Proc. AAAI Conference on Artificial Intelligence, vol. 34, no. 7, pp. 12484-12491, 2020.

[34] 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.

[35] Y. Liu, X. Chen, J. Cheng, H. Peng, and Z. Wang, "Infrared and visible image fusion with convolutional neural networks," International Journal of Wavelets, Multiresolution and Information Processing, vol. 16, no. 3, Article ID 1850018, 2018.

[36] H. Zhang, H. Xu, Y. Xiao, X. Guo, and J. Ma, "Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity," in Proc. AAAI Conference on Artificial Intelligence, vol. 34, no. 7, pp. 12797-12804, 2020.

[37] J. Ma, L. Tang, F. Fan, J. Huang, X. Mei, and Y. Ma, "SwinFusion: Cross-domain long-range learning for general image fusion via swin transformer," IEEE/CAA Journal of Automatica Sinica, vol. 9, no. 7, pp. 1200-1217, 2022.

[38] Y. Zhang, Y. Liu, P. Sun, H. Yan, X. Zhao, and L. Zhang, "IFCNN: A general image fusion framework based on convolutional neural network," Information Fusion, vol. 54, pp. 99-118, 2020.

[39] H. Li and X. J. Wu, "DenseFuse: A fusion approach to infrared and visible images," IEEE Transactions on Image Processing, vol. 28, no. 5, pp. 2614-2623, 2019.

[40] Y. Liu, S. Liu, and Z. Wang, "Multi-focus image fusion with dense SIFT," Information Fusion, vol. 23, pp. 139-155, 2015.

[41] S. Li, X. Kang, and J. Hu, "Image fusion with guided filtering," IEEE Transactions on Image Processing, vol. 22, no. 7, pp. 2864-2875, 2013.

[42] Y. Zhang, X. Bai, and T. Wang, "Boundary finding based multi-focus image fusion through multi-scale morphological focus-measure," Information Fusion, vol. 35, pp. 81-101, 2017.

[43] S. Li, X. Kang, and J. Hu, "Image fusion with guided filtering," IEEE Transactions on Image Processing, vol. 22, no. 7, pp. 2864-2875, 2013.

[44] S. Li, X. Kang, J. Hu, and B. Yang, "Image matting for fusion of multi-focus images in dynamic scenes," Information Fusion, vol. 14, no. 2, pp. 147-162, 2013.

[45] B. Zhang, X. Lu, H. Pei, and Y. Zhao, "A fusion algorithm for infrared and visible images based on saliency analysis and non-subsampled Shearlet transform," Infrared Physics & Technology, vol. 73, pp. 286-297, 2015.

[46] Q. Zhang, Y. Liu, R. S. Blum, J. Han, and D. Tao, "Sparse representation based multi-sensor image fusion for multi-focus and multi-modality images: A review," Information Fusion, vol. 40, pp. 57-75, 2018.

[47] K. Zhan, J. Teng, Q. Shi, Q. Li, and M. Wang, "Feature-linking model for image enhancement," Neural Computation, vol. 28, no. 6, pp. 1072-1100, 2016.

[48] R. Shen, I. Cheng, J. Shi, and A. Basu, "Generalized random walks for fusion of multi-exposure images," IEEE Transactions on Image Processing, vol. 20, no. 12, pp. 3634-3646, 2011.

[49] S. Li, X. Kang, L. Fang, J. Hu, and H. Yin, "Pixel-level image fusion: A survey of the state of the art," Information Fusion, vol. 33, pp. 100-112, 2017.

[50] N. Cvejic, D. Bull, and N. Canagarajah, "Region-based multimodal image fusion using ICA bases," IEEE Sensors Journal, vol. 7, no. 5, pp. 743-751, 2007.

[51] Y. Chen, R. S. Blum, "A new automated quality assessment algorithm for image fusion," Image and Vision Computing, vol. 27, no. 10, pp. 1421-1432, 2009.

[52] G. Qu, D. Zhang, and P. Yan, "Information measure for performance of image fusion," Electronics Letters, vol. 38, no. 7, pp. 313-315, 2002.

[53] M. B. A. Haghighat, A. Aghagolzadeh, and H. Seyedarabi, "A non-reference image fusion metric based on mutual information of image features," Computers & Electrical Engineering, vol. 37, no. 5, pp. 744-756, 2011.

[54] 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, 2017.

[55] C. Du and S. Gao, "Image segmentation-based multi-focus image fusion through multi-scale convolutional neural network," IEEE Access, vol. 5, pp. 15750-15761, 2017.

[56] Y. Liu, X. Chen, R. K. Ward, and Z. J. Wang, "Image fusion with convolutional sparse representation," IEEE Signal Processing Letters, vol. 23, no. 12, pp. 1882-1886, 2016.

[57] H. Tang, B. Xiao, W. Li, and G. Wang, "Pixel convolutional neural network for multi-focus image fusion," Information Sciences, vol. 433-434, pp. 125-141, 2018.

[58] Y. Zhang, Y. Liu, P. Sun, H. Yan, X. Zhao, and L. Zhang, "IFCNN: A general image fusion framework based on convolutional neural network," Information Fusion, vol. 54, pp. 99-118, 2020.

[59] X. Guo, R. Nie, J. Cao, D. Zhou, and W. Qian, "Fully convolutional network-based multi-focus image fusion," Neural Computation, vol. 30, no. 7, pp. 1775-1800, 2018.

[64] M. Amin-Naji, A. Aghagolzadeh, and M. Ezoji, "Ensemble of CNN for multi-focus image fusion," Information Fusion, vol. 51, pp. 201-214, 2019.

[68] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in Proc. European Conference on Computer Vision (ECCV), pp. 740-755, 2014.

[69] C. S. Xydeas and V. Petrovic, "Objective image fusion performance measure," Electronics Letters, vol. 36, no. 4, pp. 308-309, 2000.

[70] Y. Han, Y. Cai, Y. Cao, and X. Xu, "A new image fusion performance metric based on visual information fidelity," Information Fusion, vol. 14, no. 2, pp. 127-135, 2013.

[71] G. Piella and H. Heijmans, "A new quality metric for image fusion," in Proc. IEEE International Conference on Image Processing (ICIP), vol. 3, pp. 173-176, 2003.

[75] M. Hossny, S. Nahavandi, and D. Creighton, "Comments on 'Information measure for performance of image fusion'," Electronics Letters, vol. 44, no. 18, pp. 1066-1067, 2008.

[76] M. B. A. Haghighat, A. Aghagolzadeh, and H. Seyedarabi, "A non-reference image fusion metric based on mutual information of image features," Computers & Electrical Engineering, vol. 37, no. 5, pp. 744-756, 2011.

[77] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," IEEE Transactions on Image Processing, vol. 15, no. 2, pp. 430-444, 2006.

[78] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600-612, 2004.