IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 11, 2022, accepted March 5, 2022, date of publication April 4, 2022, date of current version April 11, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3164426

Deep Learning L2 Norm Fusion for Infrared & Visible Images

H. SHIHABUDEEN[1], (Member, IEEE), AND J. RAJEESH[2], (Member, IEEE)
[1]College of Engineering Thalassery, APJ Abdul Kalam Technological University, Thalassery, Kerala 670107, India
[2]College of Engineering Kidangoor, Kottayam, Kerala 686583, India

Corresponding author: H. Shihabudeen (shihabgec.h@gmail.com)

This work was supported in part by the Centre for Engineering Research and Development (CERD), APJ Abdul Kalam Technological University (KTU).

ABSTRACT Fusion is a strategy for collecting data from multiple images in order to improve information quality. Infrared images can recognise objects from their surroundings depending mostly on radiation disparity, which works better in all weather conditions as well as irrespective of whether it is day or night. Visible images can integrate texture information with great visual precision and in detail that matches with human visual system. Integrating the benefits of thermal radiation information with precise visual information from infrared and visible modalities is a good idea. The presented algorithm utilises the ℓ_2 norm and a combination of residual networks for combining the complementary information from both image modalities. The encoder consist of convolutional layers with selected residual connections in which the output of each layer is associated with each other layer. The ℓ_2 norm approach is then used to fuse the two featuremaps. At last, decoder recreates the fused image. The large mutual information value of 14.85084 indicates more complementary information retained in the fused image than in the infrared and visible images. The large entropy value of 6.92286 indicates more information content in the fused image and the fused image is equipped with more edge information. The proposed architecture collect more pixel values from both infrared and visible image and the fused image looks more natural as it contain more textual content. The proposed system accomplishes a noteworthy performance with the existing models.

INDEX TERMS Artificial neural networks, fusion, infrared, neural networks, visible.

I. INTRODUCTION

Multi-sensor data fusion advancement have supported a number of areas, including distant identifying, clinical imaging and contemporary military. Infrared (IR) pictures are taken utilizing IR cameras that are sensitive to warm radiation and marks. As a result, they unquestionably show heat signature assignment over the area specified, but they also have a poor dynamic range and lack of nuances. On the other hand, self-evident visible (VI) images often have simple structures and nuances because of the reflected light catch instrument of VI sensors. The targets in the images are vague when the scene is under low-light conditions or the actual locations are not clear. The IR and VI image combines and fuse comparable details from source images to create an informative image that uplifts ongoing uses of visible and infrared image fusion technology [1].

The associate editor coordinating the review of this manuscript and approving it for publication was Jiju Poovvancheri.

Several hybrid models have been implemented already in the domain of infrared and visible fusion. We partition these as 3 groups: Deep Learning-based methods [11]–[19], multi-scale transform (MST)-based systems [2]–[10], and different strategies [20]–[24]. The mixed after effects of MST-based systems for infrared and VI images are unpalatable as the images indicate different information [20]. While MST-based fusion procedures typically yield better outcomes because that multi-scale handling technique is normally appropriate to human visual systems [25]. The warm radiation within IR images is predominantly reflected in pixel power, while the appearance of visible images is fundamentally reflected by the gradient. In order to address the issue, Ma *et al.* presented a novel fusion method called the gradient transform fusion (GTF) method [20]. Their methodology assess the visible and IR image fusion as minimization process, expecting to save the thermal radiation data in IR and the gradient data in VI. But complexities with little degree are ignored in the fusion results, which can be subject to two factors. The first is that the ℓ_1-norm is utilized

to make up for the deficiency of the drops, and the second is that gradient transform fusion overlooks the pixel powers in VI.

Notably, description learning-based approaches also gained a lot of recognition. In the limited space, numerous fusion strategies have been proposed including Histogram of Oriented Gradients (HOG) and Sparse Representation (SR)-based merging strategy [26], Co-sparse Analysis [27], and Joint Sparse Representation (JSR) [28]. Li *et al.* [33] pioneered one low rank representation (LRR) combination technique within the low-position region. It uses LRR instead of using SR to remove the highlights from image and then use ℓ_1-norm as well as the choose-max fuse strategy to recreate the fused image. As Deep Learning (DL) became more prevalent, a slew of DL-based fusion techniques were proposed. Convolutional neural network (CNN) was utilised to extract image highlights and reconstruct the combined image [22], [30]. Only the effects from the last layer are used as the features in these CNN-based hybrid approaches and this results in loosing a large number of valuable data collected by the central layers. The lost features are crucial for fusion technique. Modern fusion approaches mainly collect deep and relevant features from large images and are achieved by utilising the computational capacity of deep learning architectures.

Technological advancements in imaging devices produce images with more fine details, which will be useful for further developments in industrial applications. The fusion of image modalities will collect more features and hence, it will be useful for the generation of enhanced images. Deep learning (DL) has produced cutting-edge outcomes in many computer vision and image processing applications due to its high capabilities in feature extraction and data representation. Deep learning help to collect more deep features from the image modalities. The amount of texture details in the fused image is less in most of the available literature on the subject. Texture details are fine features mostly contributed by the visible image.

To address the aforementioned problem, this paper proposes a model that involves an auto-encoder network with the encoder extracting the critical features in an image and the decoder will reconstruct the fused result. The CNN layer and residual layers are used to create the encoding network, which results in the creation of feature maps in each layer. Proper fusion strategy is adopted to get the fused feature map. Finally, we obtain a fused image as a result of the fusion strategy and by using five convolutional layers in the decoder network.

A. KEY CONTRIBUTIONS

This paper brings out an efficient deep learning model for fusion of Visible and infrared images. The paper has the following highlights;

a. Fusing IR and VI image in an efficient and accurate way
b. L2-norm is used as a fusion strategy
c. An auto encoder network creates the deep learning model.
d. Fused output of IR and VI is obtained in decoder network where it contains 5 convolutional layers

Organization of Paper: The remaining part of the paper is organised as follows: Section II depict related works that are used for fusing IR and VI images. Section III provides the proposed approach and its brief explanation. Section IV brings the performance analysis of presented approach with selected images and with different methods and finally conclusion in Section V

II. RELATED WORKS

A number of fusion procedures have been presented in recent years, many of which are heavily reliant on DL. In contrast to multi-scale decomposition strategies and representation learning approaches, fundamental learning based approaches use a collection of images and the learning have been used to find useful features.

Liu *et al.* [12] put forward one CNN based fusion strategy of IR and VI images, in which weight map obtained from the network could theoretically incorporate activity of pixel data from the source images. This model performs two key tasks: measuring activity levels and assigning weights. When comparing this model to other methods, it achieves a better visual and objective state. Ma *et al.* put forward another fusion idea named FusionGAN [14], a fusion procedure dependent on the Generative Adversarial Network. the generator searches for images that blend infrared warm radiation information to the visible gradient information. The discriminator produce the image created by the generator have more visible subtleties. Because of the discriminator, FusionGAN's combination effects have more nuances than GTF's. Since ill-disposed preparing is unreliable and prickly, there is detail mismatch with FusionGAN's combination performance. After effects of FusionGAN's fusion would be smooth and fuzzy in general, particularly the limits of targets, which is brought about by enhancing the ℓ_2-norm.

Zhang *et al.* [16] proposed CNN for image fusion named IFCNN where notable features are extracted from image and are fused by fusion rule and thereby these fused features will be given to 2 layers of convolutional to gain the fused image data. They also build multi-focus datasets based on RGB-D that have the ability to own ground values. This model can be used to generalize fusing various types of images. Ma *et al.* [19] suggested an approach that uses adverse learning to retain image information. The complete model overcomes the earlier drawbacks of conventional fusion approaches, such as the manual and complex nature of activity-level calculation and merging rules. This also allows the merged image to retain both thermal radiation and abundant textural information in the visible image while sharpening infrared target boundaries in the infrared image. When compared to other methods of evaluation, this approach provides significantly better results.

Li *et al.* [15] suggested a ResNet and Zero-phase portion analysis(ZCA) based image fusion technique. To solve the output degradation of fused images, these integrated models are used. In which ResNet extracts features from an image and then ZCA is utilised to normalise and obtain the weight maps. The final merged output is created using the weighted-average concept and the method performs better when evaluating this with the Github dataset. Li *et al.* [13] have built a deep learning method to produce an image with all of the requisite IR and VI features. They do this by decomposing the input set and then fusing the bases with a weighted-average strategy. DL is used to gather information and data and the fused image is recreated by performing ℓ_1 norm and weighted average on the data. Using dual discriminators, Xu *et al.* [18] suggested a conditional GAN for generating fused images. For merging VI and IR images of various resolutions, they used a dual-discriminator conditional generative adversarial network (DDcGAN). The fusion task is carried out between two discriminators and a generator, with the generator producing a real merged image to deceive the discriminators. The discriminators are prepared to figure out the structural dissimilarity between the likelihood distribution of down-examined fused images and infrared images, just as the structural disparity between the likelihood dispersion of fused image gradients as well as that of the gradients of infrared images. When compared to other models of assessment, this model performs better.

For the exposure fusion issue, Prabhakar *et al.* [32] suggested an approach that use CNN. The researchers utilised a basic CNN unit consisting of two convolution layers with in the encoding net as well as three Convolution layers in the decoding net. The encoder network encodes two images and by using an addition process, two-feature map patterns are produced and fused. The decoding network, which consists of three CNN layers, reconstructs the final fused image. Although this method achieves better performance, it still suffers from two main drawbacks: 1) the proposed system architecture is very simple, and key features could not be retrieved correctly. 2) These methods only use outcomes identified by final layers with in encoding net, resulting in a lack of critical details retrieved by that of the middle layers and this phenomenon would become more severe as the network becomes deeper.

To improve information transfer among the different layers, the Huang *et al.* [29] introduced a new residual block network architecture in which it uses direct connections from any layer to every successive layer. There are three advantages of dense block architecture: 1) it increases the flow of data and gradients through the network, making training of the network become smoother 2) it saves as much information as possible and 3) dense links have a normalizing impact which eliminates overfitting of the model.

III. MATERIALS AND METHODOLOGIES

Fig. 1 briefly describes our proposed model in which the input images are collected from MSCOCO 2014 repository that will be given for entire process. Initially these images will have certain anomalies due to movement of camera and object. To avoid those anomalies, we pre-registered all the images and they are passed to the encoding network which is the most important step. The principle that works behind this method is a block of convolutional neural network with selected residual connections. The pre-registered images are used to train the network which generate the feature map and the decoder network recreates the image from this featuremap. After the training process, Featuremap generated by the encoding network are passed to next important process, which is L_2-norm. This is used for measuring activity level of the two featuremaps and an averaging operation is performed on the featuremaps. Finally fused feature maps are passed over to 5-layer CNN, which is the decoder network and this important stage produce the fused image based on calculation from previous stages.

A. DATASET

For the proposed system, the images that is used for training is taken from MS-COCO 2014 (http://cocodataset.org/) repository with the following highlights [34];

a. It contains more than 80,000 instances of IR and visible images
b. Images are resized to 256×256
c. Learning rate is set to $1x10^{-4}$
d. Images are split into train (79,000), and validation (1000) sets
e. The batch size is set to 4

B. SOFTWARE AND PLATFORM

Tensor Flow is one of the open-source programming libraries created by Google Brain Team. They have created tensor flow to lead research in machine learning and deep neural organizations. Fundamentally, this is a delicate product library used for mathematical calculations utilizing the information stream charts. The hubs in the chart address the numerical activities and the multidimensional information exhibits are addressed by the edges in the diagram (called tensors).In this work, we have used Tensor Flow and it is executed over Google Collaboratory with NVIDIA GTX 1050Ti GPU.

C. PRE-REGISTRATION

Pre-registration is the initial stage in which we feed the dataset. Before this we convert these pairs of images (IR, VI) to grey scale image using conversion concept. Then we will analyse the outliers that are presented in this converted image. These errors that are happened due to movement of camera and object in fusion, to rectify it preregistration algorithm is used. Pre-registration uses area-based and feature-based methods to find the best alignment between images. Area based methods use comparison of intensity values for the pre-registration, while feature based methods look for features like corners, neighbourhoods, coordinates etc. Traditional pre-registration flowchart is as follows in Fig. 2.

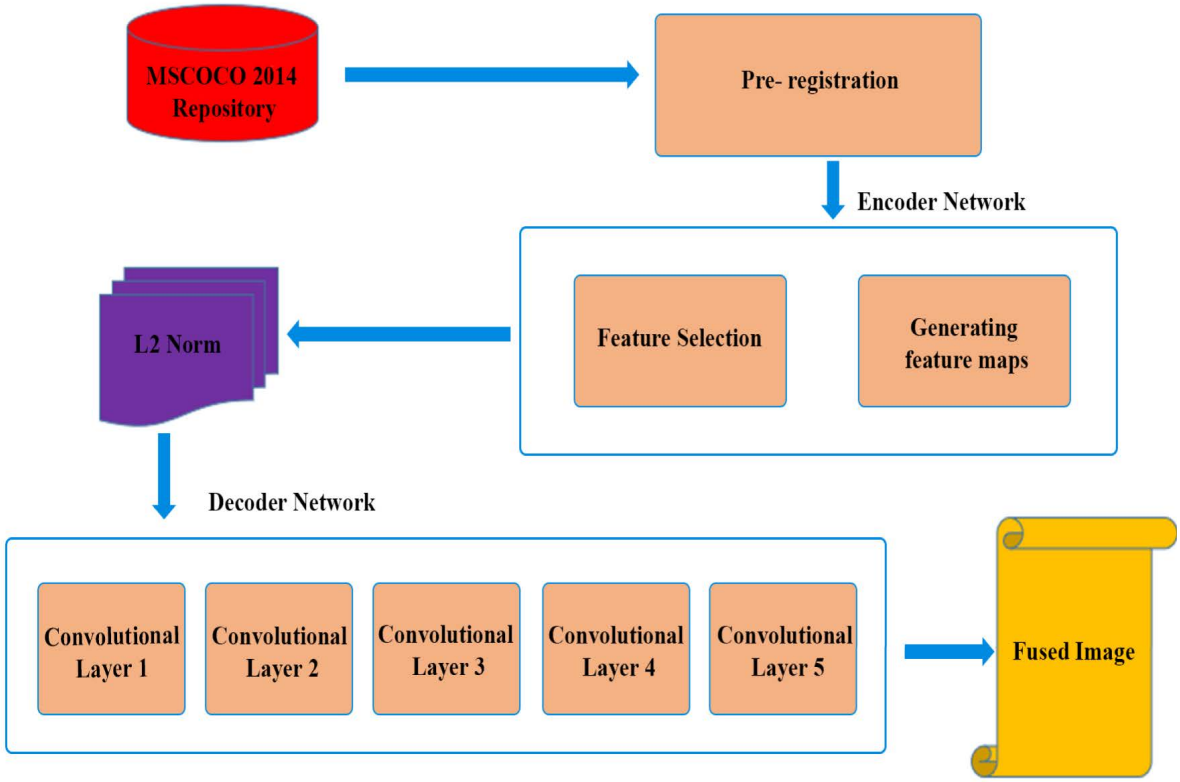

FIGURE 1. Design of proposed model.

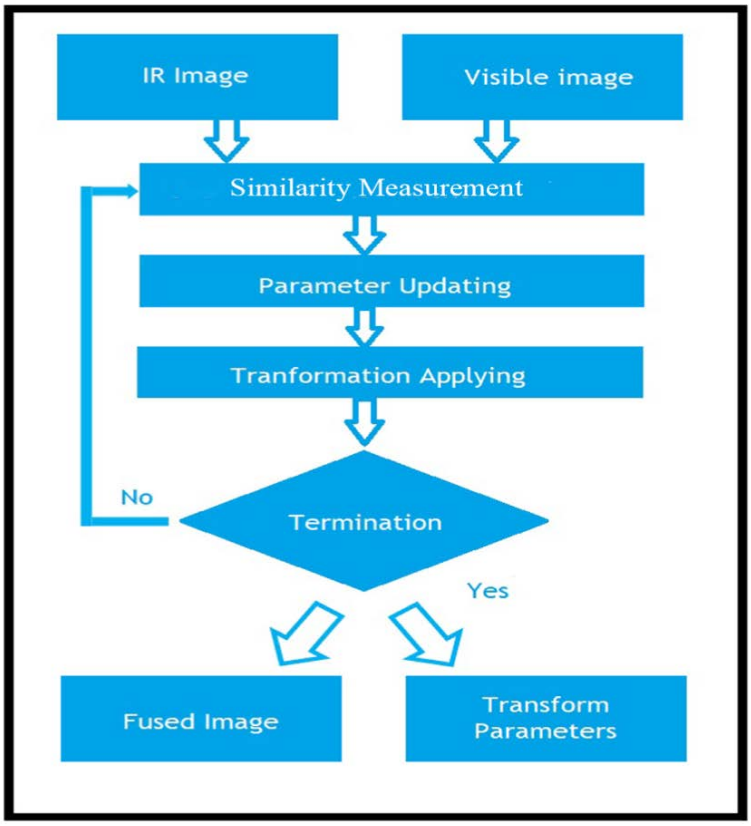

FIGURE 2. Traditional pre-registration flowchart.

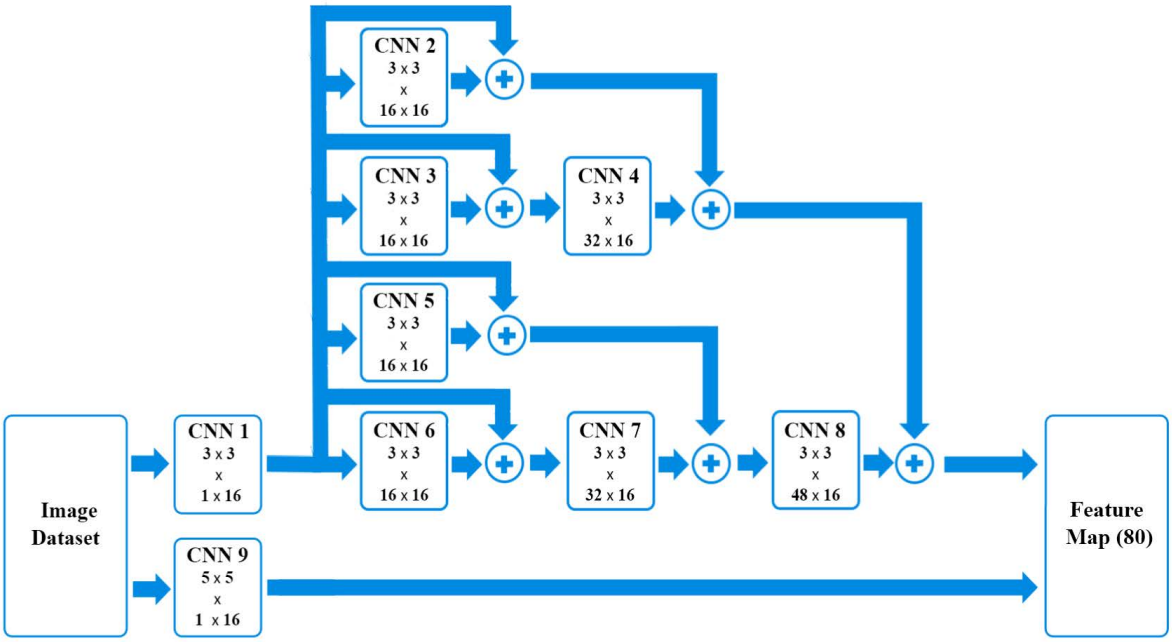

FIGURE 3. Encoding network.

D. ENCODING NETWORK

The encoding network performs the major step in the whole approach. The encoding network has two main functions, namely, feature selection and feature map generation (Fig. 3). Here the pre-registered image is given to this network as it moves to feature selection which is convolutional layer (Conv 1) with 3 × 3 filter and also to layer (Conv 2) with 5 × 5 filter. This is used to extract several rough features. The required number of filters is set to 16 and the output of 1 × 16 are passed towards next process which will generate feature maps. This approach uses convolutional layers with selected residual connections to reduce the training procedure and 3 × 3 filters to extract the relevant features. This will generate 1 × 16 feature maps for each layer. As compared to other CNNs, the output of all layers are send to the consecutive layer which retains the deep features are used for fusion. Residual networks are complex in nature, but they reduce the chance of overfit, which is common in deep convolution network models. Selection of stride as 1 and filter size of 3 can collect more deep features from the selected image. Then by usage of residual concept, it can preserve features that are deep and also it can make sure all notable features are used or not. Also, it reduces the overfitting of data, training and testing time gets increased and finally visual perception also increased [17]. CNN layer with 5 × 5 filter can extract some fine features and can help in the reconstruction of some important features missed due to the fast convergence of residual network explained earlier. The features from this convolutional layer are combined with those from the residual network and driven to the decoding network for reconstruction.

Selection of the depth parameter for the convolutional layer is very crucial as it relates to the feature selection concept. As the layers gets increased, we can easily reduce the overfitting and represent features deeply. Convolutional layer is represented by 4 parameters such as filter size; here we use 3 × 3 filter size in form of W_1 X H_1 X D_1, depth; the output volume is a hyper-parameter. It corresponds to the number of filters (K) that we would like to represent the feature information. stride (s); relates to sliding of the weight values over the image. When the stride is set to 1, we slide the filter masks by one pixel at a time across the entire image, zero padding (P); for controlling special size (F) of output volume. Thus a size volume W_2 X H_2 X D_2, is generated where;

$$W_2 = \frac{W_1 - F + 2P}{s + 1} \tag{1}$$

$$H_2 = \frac{H_1 - F + 2P}{s + 1} \tag{2}$$

$$D_2 = K \tag{3}$$

E. DECODER NETWORK

This is the last and final stage in the proposed model, where the network is dense network and comprises of 5 convolutional layers with 3 × 3 filters. The 64 mapped features will be passed to decoder along with the feature collected by the 5 × 5 filter and finally reconstruct the fused image (Fig. 4). The convolutional layer comprise of ReLu layers and the activation function would be element based, like as max(0,x) thresholding at 0. As a result of this, the volume remains constant. The pool layer will conduct a

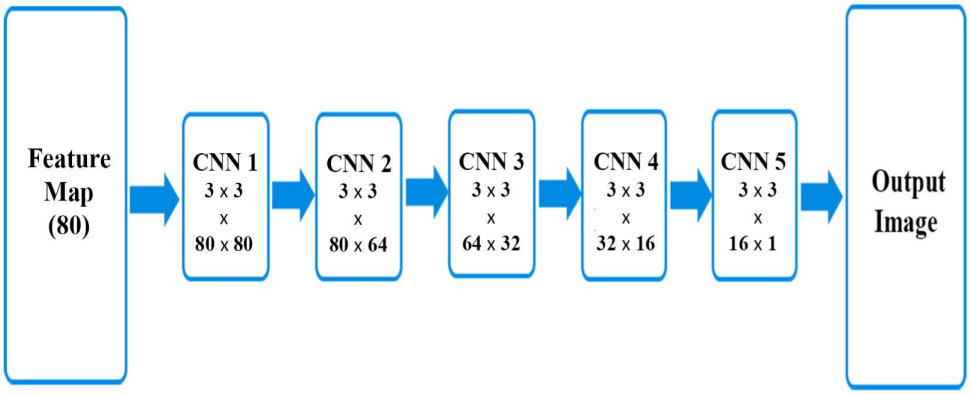

FIGURE 4. Decoding network.

down sampling process, resulting in a volume with spatial dimensions of 16 x 16 x 1.

F. ℓ_2 NORM STRATEGY

Here we use trained network with ℓ_2 norm fusion in which it is used for measuring the activity level. Activity map of each feature point are derived from the formula;

$$C_i(x, y) = \left\| \psi_i^{1:M}(x, y) \right\|_2 \tag{4}$$

Averaging operation is applied to the individual feature map to obtain the activity level measurement of relevant features for the fusion purpose and is given by equation no.6

$$AM_i(x, y) = \frac{\sum_{a=-c}^{c} \sum_{b=-c}^{c} A_i(x + a, y + b)}{(2c + 1)^2} \tag{5}$$

$$\omega_i(x, y) = \frac{AM_i(x, y)}{\sum_{i=1}^{k} AM_i x, y)} \tag{6}$$

The activity map and function map are used to create the final fused image. When searching for a fused coefficient map, Softmax is widely used. The final merged output is generated by

$$FF^m(x, y) = \sum_{i=1}^{k} \omega_i(x, y) \psi_i^m(x, y) \tag{7}$$

G. TRAINING PHASE

During the training stage, we only take account of the encoder-decoder nets and the fusion layer is ignored. The training process try to build encoder and decoder nets capable of recreating the saliency map or the image. Once the encoder and the decoder weights are set, a new fusion technique is selected to combine the feature map from the encoder. Fig. 5 depicts the overall training procedure of the auto encoder network. The modification of fusion layer is possible based on the different applications.

The layer 1 represents the convolution layer resides in the encoder network, which comprises of a 3 x 3 channels,

TABLE 1. Architecture details of auto-encoder network.

Network	CNN Layer	Filter length	Stride	Input Channel	Output Channel	Activation Function
Encoder	1	3	1	1	16	Re Lu
	2	3	1	16	16	Re Lu
	3	3	1	16	16	Re Lu
	4	3	1	32	16	Re Lu
	5	3	1	16	16	Re Lu
	6	3	1	16	16	Re Lu
	7	3	1	32	16	Re Lu
	8	3	1	48	16	Re Lu
	9	5	1	1	16	Re Lu
Decoder	1	3	1	80	80	Re Lu
	2	3	1	80	64	Re Lu
	3	3	1	64	32	Re Lu
	4	3	1	32	16	Re Lu
	5	3	1	16	1	

as seen in Fig. 5.The residual convnets and the output of each convnet is cascaded into the subsequent layer. Five consecutive convolutional layers make up the decoder. It will be used to recover the given data image. Table 1 will give architectural details convolutions layers used the auto encoder network.

The loss function, which decreases the deviation from the actual target to the predicted values, is an error-minimizing function. The total number of absolute differences between n samples is expressed as,

$$L_1 = \sum_{i,j=0}^{n-1} [\, y(i, j) - x(i, j)] \tag{8}$$

Mean square error (MSE) is utilised as a cost function for Conv-net training in several articles and mainly deals with the perceived errors in the image. The ℓ_2 loss function or MSE minimizes the squared difference between the expected and

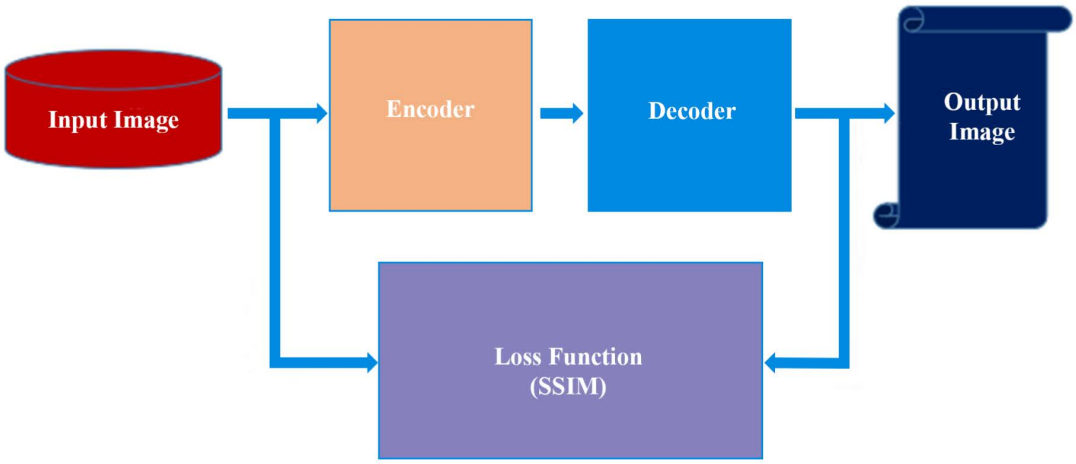

FIGURE 5. Auto encoder network.

existing target values and is defined as;

$$L_2 = \sqrt{\sum_{i,j=0}^{n-1} [\, y(i,j) - x(i,j)]^2} \tag{9}$$

The primary cost feature for training CNN is usually selected as an MSE or ℓ_2 loss function. For its easy optimization behaviour the ℓ_2 cost function is preferred. When compared to the L1 norm, L2 error will be considerably larger than in the presence of noise.

The Structural Similarity Index Measure (SSIM) is a new quality index that will give information about the loss and distortion in an image. It contains 3 components, such as Luminance, Contrast Distortion, and Loss of correlation [42]. The expression of SSIM is,

$$SSIM = \sum_{xf} \left(\frac{2\mu_x \mu_f + C_1}{\mu_x^2 \mu_f^2 + C_1} \right) \left(\frac{2\sigma_x \sigma_f + C_2}{\sigma_x^2 \sigma_f^2 + C_2} \right) \left(\frac{\sigma_{xf} + C_3}{\sigma_x \sigma_f + C_3} \right) \tag{10}$$

$SSIM_{x,f}$ structural similarities of input (x) and fused (f) image,
σ_{xf} the covariance of input and fused images,
σ_x,σ_f Standard deviation of input and fused images
μ_x, μ_f mean value of input and fused images
C_1, C_2, C_3 constants used to stabilise the algorithm

SSIM can correlate well with human's perception of image quality. SSIM mainly deals with the structural similarity between two images and it mainly attempts to model the perceived changes in the structural information of the image. The training process's aim is to develop an auto-encoder network (encoder and decoder) that can extract and reconstruct features more accurately. Since infrared and visible image training data are not enough, we utilize gray level images within the MSCOCO dataset for model training process.

IV. EXPERIMENT RESULTS AND ANALYSIS

During the training and testing processes, we assess the model using subjective and objective parameters as well as comparing it to other models to examine how well it works. SSIM is also used for the training process, which is addressed in Section 3. Under the aforementioned conditions, the proposed model produced a better reconstruction image. Here we use MS COCO dataset [34] images for training our model. Around 79000 images are used as the input from the dataset, with 1000 images used to analyze the auto encoder network. To check the effectiveness of the algorithm, SSIM is used. In general, as we move through the training process, our network will converge and it will take less time for the training process.

The optimal weights of the trained auto encoder model are used for the fusion process. 20 sets of Visible (VI) and Infrared (IR) images were utilised for experimental analysis of the fusion algorithm. Fig. 6 depicts raw IR and VI images as well as their fused images. The fused image contains more complementary information and it is evident in Fig. 6. To evaluate the proposed algorithm's effectiveness, it is compared to similar approaches such as the cross-bilateral filter (CBF) [5], gradient transfer and total variation minimization (GTF) [20], the joint-sparse representation (JSR) [28], DeepFuse [32], Dense fuse [17], Weighted Least Square optimization (WLS) [7], JSR with saliency detection (JSRSD) [31], ResNet-ZCA [15] and FusionGAN [14].

The fused outputs generated by the CBF algorithm have more noise than the information from both images. Due to the artefacts, the relevant features are not clear and CBF is not recommended for merging visible and infrared images. The images generated by GTF and CBF hold more details

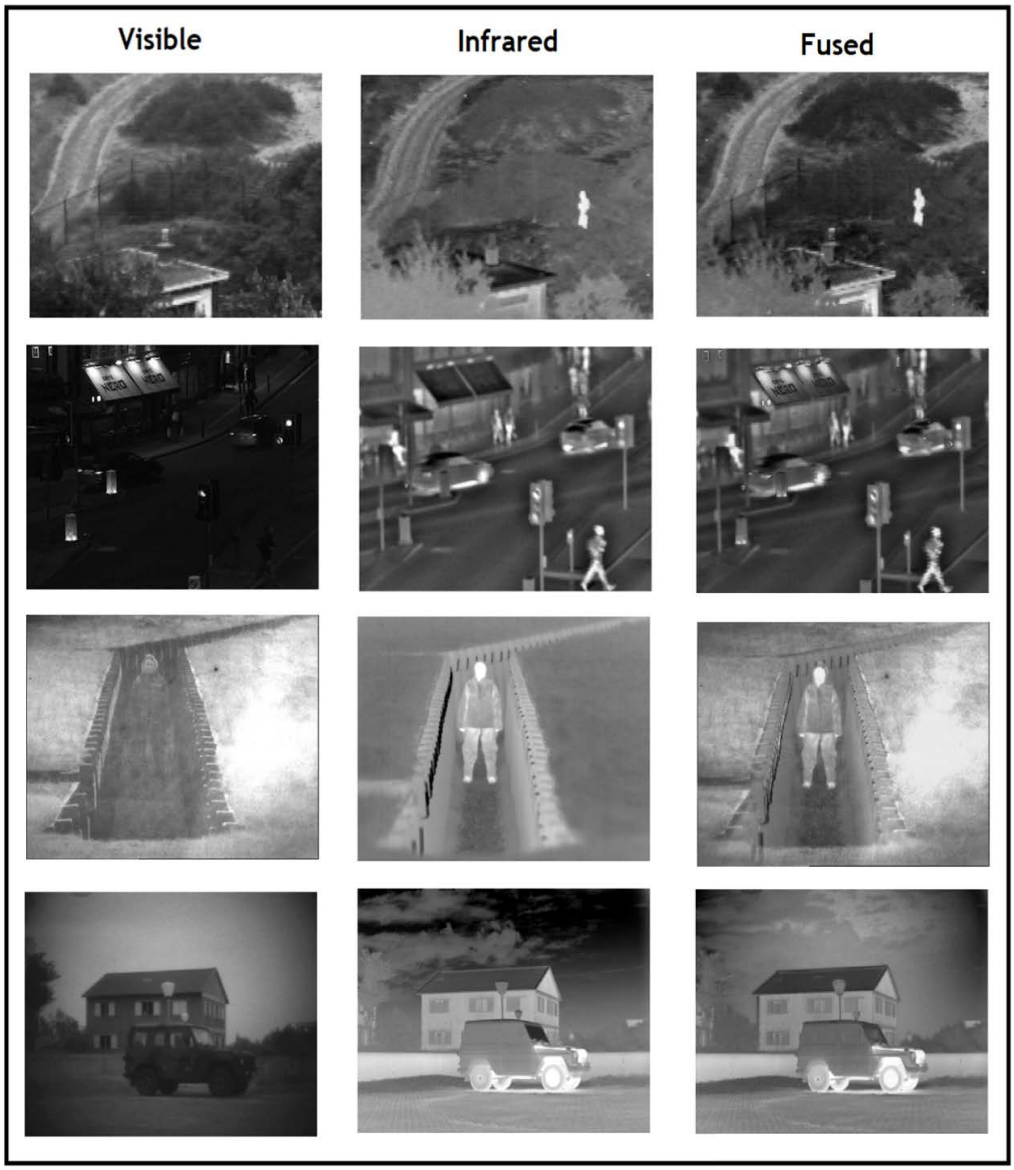

FIGURE 6. Selected visible and infrared images and their fusion results.

from the IR image, but texture information is less. So, the fused output is not suitable for daylight conditions. The fused images generated by JSR and JSRSD are also not good as they contain more artefacts and the complementary information provided is not useful. Fused images by other methods contain less artefacts and noise, and contains more complementary information.

The fused images generated by CNN based methods like Deep fuse, Dense fuse, WLS, Res-Net, Fusion GAN, and the presented method holds more relevant features. The fused output images are consistent with human visual perception. Fused images by WLS provide more textual information when compared with GTF and JSR. When compared to other fused outputs, the Fusion GAN-based merged image retains more information from the Infrared image and appears darker. The proposed fusion method generates a merged image with more textual content that appears more realistic, as well as relevant information from the IR image. Based on the subjective evaluation of the selected fused images, we can conclude that proposed method retains more salient features from infrared and relevant textual information from visible images.

Visual perception and objective assessment are needed to study the effectiveness of the approach. Eleven performance measures were utilised for the evaluation of the implemented fusion method and selected related approaches. They are

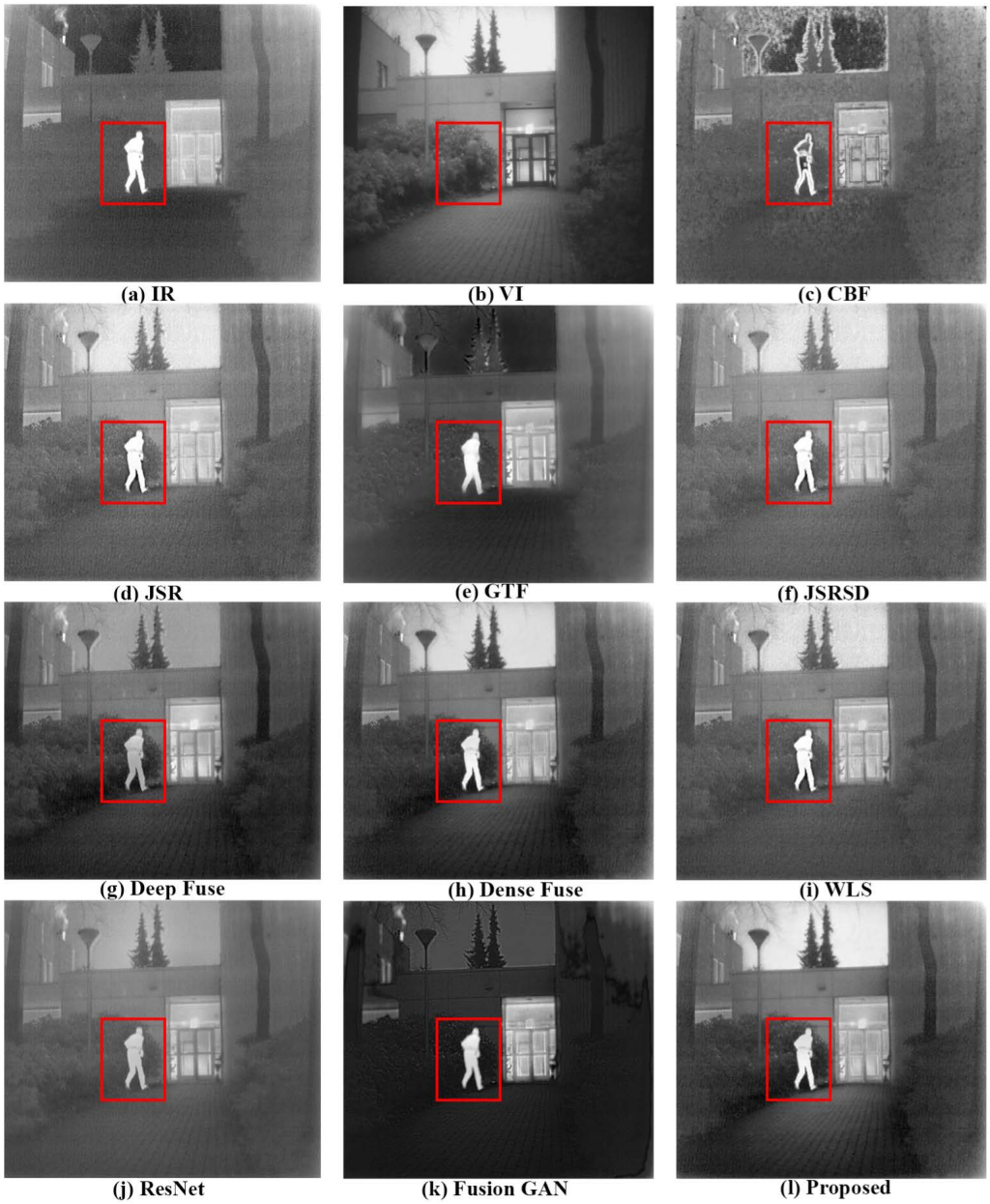

FIGURE 7. Analysis of "man" images with selected fusion methods (a) IR; (b) VI; (c) CBF; (d)JSR; (e) GTF; (f) JSRSD; (g) DeepFuse; (h) DenseFuse; (i) WLS; (j) ResNet-ZCA; (k) Fusion GAN; (l) Proposed.

as follows: Entropy (EN) will determines how much information is retained in the fused image [35]. The high EN value reflects the large number of features in the combined image. Mutual Information (MI) estimates the features that are conveyed to the fused image. A larger MI metric indicates more details are retained from the individual to the fused image [37]. Q_{abf} is an another metric that gives an idea about the quality of the visual information in the merged image [36]. Sum of Correlation Differences (SCD) will give information about the correlation differences of the fused image with individual images [38].

Other metrics used commonly for objective evaluation are FMI_p, FMI_w, and FMI_{dct}, which give mutual information in a fused image with pixel, cosine, and wavelet as features [39]. The quantity of edge pixels preserved in the merged image is indicated by edge preservation index (EPI) [40]. Visual Information Fidelity (VIF) is a metric for assessing picture quality that provides details on information fidelity [41]. $SSIM_a$ and MS-SSIM are modified SSIMs that will check the amount of structural similarity among fused and individual images [43]. $SSIM_a$ will give the average value for similarities between fused and individual images [42].

TABLE 2. The average values of metric for all the images selected for conducting the analysis.

Method	En	Q_{abf}	SCD	FMI_w	FMI_{dct}	$SSIM_a$	MS_{ssim}	MI	FMI_p	VIF	EPI
CBF	6.84494	0.44119	1.38963	0.32013	0.26997	0.60304	0.70879	13.71498	0.87203	0.71489	0.57240
JSR	6.78576	0.32572	1.59136	0.18629	0.14584	0.53906	0.75523	12.72654	0.88463	0.75533	0.59644
GTF	6.63597	0.40993	1.06159	0.41004	0.39244	0.70369	0.80844	13.26865	0.86429	0.41687	0.67011
JSRSD	6.78441	0.32542	1.59123	0.18279	0.14868	0.53963	0.75517	13.38575	0.90207	0.75517	0.47473
Deepfuse	6.68170	0.44615	1.49016	0.41418	0.40222	0.72949	0.72659	13.39869	0.91015	0.80198	0.70197
Densefuse	6.81348	0.46791	1.71264	0.43009	0.38823	0.72052	0.68778	13.34317	0.89921	0.78540	0.68740
WLS	6.64071	0.52134	1.71705	0.37663	0.31781	0.72360	0.71867	13.28143	0.90470	0.80014	0.66775
ResNet	6.19527	0.36341	1.58169	0.39497	0.38155	0.80195	0.74961	12.39054	0.89577	0.79197	0.70180
Fusion GAN	6.36285	0.35585	1.42368	0.40385	0.39357	0.65384	0.73182	12.72570	0.90108	0.77845	0.59323
Proposed	6.92286	0.52147	1.76286	0.43035	0.34063	0.74536	0.79458	14.85084	0.91650	0.80687	0.68471

Table 2 shows the average values of metric values of all the images selected for conducting the analysis. Better metrics are shown in red, while the second best metrics are shown in blue. The proposed method achieves better values for seven metrics (EN, MI, Q_{abf}, SCD, FMI_p, FMI_w, VIF) and second best values for two metrics ($SSIM_a$, MS-SSIM) and comparable values for other metrics. High value of En indicate the presence of more information content in the merged image. When compared to other models, higher values of *Qabf* and SCD imply that the merged image contains fewer artificial noise and the images are more realistic. The fused image contains a large quantity of data from the IR and VI images, as evidenced by the higher MI value.

The objective evaluation demonstrates that the suggested algorithm outperforms other models in terms of fusion performance. So the auto encoder network with the ℓ_2 norm as the fusion strategy can be used as a tool for fusing infrared and visible images.

V. CONCLUSION

We formulate the task of fusing visible (VI) and infrared (IR) images as a ℓ_2-norm minimization concept, along with the intention of producing fused image that resembles the IR image but contains more VI presentation details. We proposed the residual architecture model that includes both convolutional and residual layers in order to obtain an efficient fused image. To develop this model, we feed MS-COCO 2014 dataset containing both IR and VI into pre-registration stage and then pass through encoder network. Fusion is performed by ℓ_2 norm strategy and then to decoding network to reconstruct the fused image. While evaluating our model, we obtained much greater fusion performance compared to other existing models. Future work will look at the semantic relationship and its correction to improve the algorithm's efficiency. This architecture can be used to address a variety of multi-sensor fusion challenges in medical imaging and also in remote sensing applications. The work is also useful for other researchers out to explore, analyse and finally bring new add-ons to build this model even more efficient. They can take this model as an inspiration to build other integrated model using deep learning for obtaining fused image.

ACKNOWLEDGMENT

H. Shihabudeen would like to thank the College of Engineering Thalassery, College of Engineering Kidangoor, and APJ Abdul Kalam Technological University, Kerala, for giving support for carrying out the research work. He also thank his supervisor, who has provided many directions in conducting this research.

REFERENCES

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

[2] W. Gan, X. Wu, W. Wu, X. Yang, C. Ren, X. He, and K. Liu, "Infrared and visible image fusion with the use of multi-scale edge-preserving decomposition and guided image filter," *Infr. Phys. Technol.*, vol. 72, pp. 37–51, Sep. 2015, doi: 10.1016/j.infrared.2015.07.003.

[3] S. Zhenfeng, L. Jun, and C. Qimin, "Fusion of infrared and visible images based on focus measure operators in the curvelet domain," *Appl. Opt.*, vol. 51, no. 12, pp. 1910–1921, Apr. 2012, doi: 10.1364/AO.51.001910.

[4] X. Yan, H. Qin, J. Li, H. Zhou, and J.-G. Zong, "Infrared and visible image fusion with spectral graph wavelet transform," *J. Opt. Soc. Amer. A*, vol. 32, no. 9, pp. 1643–1652, Sep. 2015, doi: 10.1364/JOSAA.32.001643.

[5] B. K. S. Kumar, "Image fusion based on pixel significance using cross bilateral filter," *Signal, Image Video Process.*, vol. 9, no. 5, pp. 1193–1204, Jul. 2015, doi: 10.1007/s11760-013-0556-9.

[6] Z. Fu, X. Wang, J. Xu, N. Zhou, and Y. Zhao, "Infrared and visible images fusion based on RPCA and NSCT," *Infr. Phys. Technol.*, vol. 77, pp. 114–123, Jul. 2016, doi: 10.1016/j.infrared.2016.05.012.

[7] J. Ma, Z. Zhou, B. Wang, and H. Zong, "Infrared and visible image fusion based on visual saliency map and weighted least square optimization," *Infr. Phys. Technol.*, vol. 82, pp. 8–17, May 2017, doi: 10.1016/j.infrared.2017.02.005.

[8] D. P. Bavirisetti and R. Dhuli, "Two-scale image fusion of visible and infrared images using saliency detection," *Infr. Phys. Technol.*, vol. 76, pp. 52–64, May 2016, doi: 10.1016/j.infrared.2016.01.009.

[9] J. Chen, X. Li, L. Luo, X. Mei, and J. Ma, "Infrared and visible image fusion based on target-enhanced multiscale transform decomposition," *Inf. Sci.*, vol. 508, pp. 64–78, Jan. 2020, doi: 10.1016/j.ins.2019.08.066.

[10] J. Zhu, W. Jin, L. Li, Z. Han, and X. Wang, "Fusion of the low-light-level visible and infrared images for night-vision context enhancement," *Chin. Opt. Lett.*, vol. 16, no. 1, Jan. 2018, Art. no. 013501.

[11] X. Ren, F. Meng, T. Hu, Z. Liu, and C. Wang, "Infrared-visible image fusion based on convolutional neural networks (CNN)," in *Intelligence Science and Big Data Engineering*. Cham, Switzerland, 2018, pp. 301–307, doi: 10.1007/978-3-030-02698-1_26.

[12] Y. Liu, X. Chen, J. Cheng, H. Peng, and Z. Wang, "Infrared and visible image fusion with convolutional neural networks," *Int. J. Wavelets, Multiresolution Inf. Process.*, vol. 16, no. 3, p. 1850018, May 2018, doi: 10.1142/S0219691318500182.
[13] H. Li, X.-J. Wu, and J. Kittler, "Infrared and visible image fusion using a deep learning framework," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 2705–2710, doi: 10.1109/ICPR.2018.8546006.
[14] J. Ma, W. Yu, P. Liang, C. Li, and J. Jiang, "FusionGAN: A generative adversarial network for infrared and visible image fusion," *Inf. Fusion*, vol. 48, pp. 11–26, Aug. 2019, doi: 10.1016/j.inffus.2018.09.004.
[15] H. Li, X. J. Wu, and T. S. Durrani, "Infrared and visible image fusion with ResNet and zero-phase component analysis," *Infr. Phys. Technol.*, vol. 102, Nov. 2019, Art. no. 103039, doi: 10.1016/j.infrared.2019.103039.
[16] Y. Zhang, Y. Liu, P. Sun, H. Yan, X. Zhao, and L. Zhang, "IFCNN: A general image fusion framework based on convolutional neural network," *Inf. Fusion*, vol. 54, pp. 99–118, Feb. 2020, doi: 10.1016/j.inffus.2019.07.011.
[17] H. Li and X.-J. Wu, "DenseFuse: A fusion approach to infrared and visible images," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2614–2623, May 2019, doi: 10.1109/TIP.2018.2887342.
[18] H. Xu, P. Liang, W. Yu, J. Jiang, and J. Ma, "Learning a generative model for fusing infrared and visible images via conditional generative adversarial network with dual discriminators," in *Proc. IJCAI*, 2019, pp. 3954–3960.
[19] J. Ma, P. Liang, W. Yu, C. Chen, X. Guo, and J. Wu, "Infrared and visible image fusion via detail preserving adversarial learning," *Inf. Fusion*, vol. 54, pp. 85–98, Feb. 2020, doi: 10.1016/j.inffus.2019.07.005.
[20] J. Ma, C. Chen, C. Li, and J. Huang, "Infrared and visible image fusion via gradient transfer and total variation minimization," *Inf. Fusion*, vol. 31, pp. 100–109, Sep. 2016, doi: 10.1016/j.inffus.2016.02.001.
[21] H. Guo, Y. Ma, X. Mei, and J. Ma, "Infrared and visible image fusion based on total variation and augmented Lagrangian," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 34, no. 11, pp. 1961–1968, Nov. 2017, doi: 10.1364/JOSAA.34.001961.
[22] Y. Liu, X. Chen, H. Peng, and Z. F. Wang, "Multi-focus image fusion with a deep convolutional neural network," *Inf. Fusion*, vol. 36 pp. 191–207, Jul. 2017, doi: 10.1016/j.inffus.2016.12.001.
[23] H. Li, X.-J. Wu, and J. Kittler, "MDLatLRR: A novel decomposition method for infrared and visible image fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 4733–4746, 2020, doi: 10.1109/TIP.2020.2975984.
[24] Y. Zhang, L. Zhang, X. Bai, and L. Zhang, "Infrared and visual image fusion through infrared feature extraction and visual information preservation," *Infr. Phys. Technol.*, vol. 83, pp. 227–237, Jun. 2017, doi: 10.1016/j.infrared.2017.05.007.
[25] G. Piella, "A general framework for multiresolution image fusion: From pixels to regions," *Inf. Fus.*, vol. 4, no. 4, pp. 259–280, Dec. 2003, doi: 10.1016/s1566-2535(03)00046-0.
[26] J.-J. Zong and T.-S. Qiu, "Medical image fusion based on sparse representation of classified image patches," *Biomed. Signal Process. Control*, vol. 34, pp. 195–205, Apr. 2017, doi: 10.1016/j.bspc.2017.02.005.
[27] R. Gao, S. A. Vorobyov, and H. Zhao, "Image fusion with cosparse analysis operator," *IEEE Signal Process. Lett.*, vol. 24, no. 7, pp. 943–947, Jul. 2017, doi: 10.1109/LSP.2017.2696055.
[28] Q. Zhang, Y. Fu, H. Li, and J. Zou, "Dictionary learning method for joint sparse representation-based image fusion," *Opt. Eng.*, vol. 52, no. 5, May 2013, Art. no. 057006, doi: 10.1117/1.OE.52.5.057006.
[29] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," 2016, *arXiv:1608.06993*.
[30] Y. Liu, X. Chen, R. K. Ward, and Z. J. Wang, "Image fusion with convolutional sparse representation," *IEEE Signal Process. Lett.*, vol. 23, no. 12, pp. 1882–1886, Dec. 2016, doi: 10.1109/LSP.2016.2618776.
[31] C. H. Liu, Y. Qi, and W. R. Ding, "Infrared and visible image fusion method based on saliency detection in sparse domain," *Infr. Phys. Technol.*, vol. 83, pp. 94–102, Jun. 2017, doi: 10.1016/j.infrared.2017.04.018.
[32] K. R. Prabhakar, V. S. Srikar, and R. V. Babu, "DeepFuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4724–4732, doi: 10.1109/ICCV.2017.505.
[33] H. Li and X.-J. Wu, "Multi-focus image fusion using dictionary learning and low-rank representation," vol. 10666, pp. 675–686, 2017, doi: 10.1007/978-3-319-71607-7_59.
[34] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Computer Vision—ECCV 2014*, Sep. 2014, pp. 740–755, doi: 10.1007/978-3-319-10602-1_48.
[35] J. W. Roberts, F. B. Ahmed, and J. A. Van Aardt, "Assessment of image fusion procedures using entropy, image quality, and multispectral classification," *J. Appl. Remote Sens.*, vol. 2, no. 1, May 2008, Art. no. 023522, doi: 10.1117/1.2945910.
[36] C. S. Xydeas and V. Petrović, "Objective image fusion performance measure," *Electron. Lett.*, vol. 36, no. 4, pp. 308–309, Feb. 2000, doi: 10.1049/el:20000267.
[37] L. Paninski, "Estimation of entropy and mutual information," *Neural Comput.*, vol. 15, no. 6, pp. 1191–1253, Jun. 2003, doi: 10.1162/089976603321780272.
[38] V. Aslantas and E. Bendes, "A new image quality metric for image fusion: The sum of the correlations of differences," *AEU Int. J. Electron. Commun.*, vol. 69, no. 12, pp. 1890–1896, Dec. 2015, doi: 10.1016/j.aeue.2015.09.004.
[39] M. B. A. Haghighat, A. Aghagolzadeh, and H. Seyedarabi, "A non-reference image fusion metric based on mutual information of image features," *Comput. Elect. Eng.*, vol. 37, no. 5, pp. 744–756, Sep. 2011, doi: 10.1016/j.compeleceng.2011.07.012.
[40] J. Joseph, S. Jayaraman, R. Periyasamy, and S. Renuka, "An edge preservation index for evaluating nonlinear spatial restoration in MR images," *Current Med. Imag. Rev.*, vol. 13, no. 1, pp. 58–65, Jan. 2017.
[41] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006, doi: 10.1109/TIP.2005.859378.
[42] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004, doi: 10.1109/TIP.2003.819861.
[43] K. Ma, K. Zeng, and Z. Wang, "Perceptual quality assessment for multi-exposure image fusion," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3345–3356, Nov. 2015, doi: 10.1109/TIP.2015.2442920.

H. SHIHABUDEEN (Member, IEEE) received the B.Tech. degree from the University of Calicut, Kerala, India, in 2009, and the M.Tech. degree from the Cochin University of Science and Technology, Kerala, in 2011.

He started his career as a Teacher in the year 2012. In 2016, he joined as a Part Research Scholar with APJ Abdul Kalam Technological University, Kerala. He has received Research Seed Money funding from the Centre of Engineering Research & Development in the year 2019 and currently working as an Assistant Professor with the College of Engineering Kidangoor. He has published eight research papers in international journals, including national and international conferences. He has a total of 11 years of experience in teaching and research related activities. His research interests include image processing for fusion applications like infrared and visible fusion, multi-focus fusion, and deep learning enabled target detection.

J. RAJEESH (Member, IEEE) received the B.E. degree in electronic and communication and the M.E. degree in communication system from Madurai Kamaraj University, Tamil Nadu, India, in 1990 and 2002, respectively, and the Ph.D. degree in communication systems from Anna University, Chennai, India, in 2012.

From 1992 to 2002, he was a Biomedical Engineer with the Dr. Jeyasekaran Medical Trust. He started his teaching career as a Lecturer with the Noorul Islam College of Engineering and later worked as an Associate Professor at Noorul Islam University. He is currently working as a Professor with the College of Engineering Kidangoor. He has a total of 27 years of experience in teaching and research related activities. He has published more than 30 research papers in international journals, including national and international conferences. He received few funded projects and consultancy works, including funding from AICTE and DST. His research interests include computer-aided diagnosis for cancer detection, detection of alzheimer's disease and its progression, and image processing for computer-aided diagnosis systems.