Multimodal Sensor Fusion Using Symmetric Skip Autoencoder Via an Adversarial Regulariser

Snigdha Bhagat, Shiv Dutt Joshi, *Member, IEEE*, Brejesh Lall, *Member, IEEE*, and Smriti Gupta

***Abstract*—The fusion of the spatial characteristics, of visual image, and spectral aspects, of infrared image, is of immense practical importance. In this work, we propose a novel spatially constrained adversarial autoencoder that extracts deep features from the infrared and visible images to obtain a more exhaustive and global representation. A residual autoencoder architecture, regularised by a residual adversarial network has been employed, to generate a more realistic fused image. The residual module serves as a primary building block for the encoder, decoder, and adversarial network, as an add on the symmetric skip connections, perform the functionality of embedding the spatial characteristics directly from the initial layers of encoder structure to the decoder part of the network. The spectral information in the infrared image is incorporated by adding the feature maps over several layers in the encoder part of the fusion structure. The encoder section is made up of two separate branches to carry out independent inference on both the visual as well as infrared images. The loss function has been designed to incorporate the characteristics of both the modalities by optimizing over the textural content of the visible image and the spectral content of its infrared counterpart. In order to efficiently optimize the network's parameters, an adversarial regulariser network has been proposed that would perform supervised learning on the fused image and the original visual image since the visual image contains most of the structural content in comparison to the infrared image. The adversarial game has been incorporated in the structure by the addition of classification loss in the generator and discriminator loss functions in addition to the content loss.**

***Index Terms*—Autoencoder, generative modeling, image fusion, TNO dataset.**

I. Introduction

INFORMATION fusion is a technique to integrate relevant information from disparate sensors to merge, collate or juxtapose data in order to obtain a robust image output which can facilitate several subsequent processing tasks [1]. When we are dealing with the problem of visible and infrared image fusion, the underlying issue is the limitation of bandwidth in capturing image data. The effectiveness of fusion is a very subjective process, and thus it is often assessed by the level of artifacts and abnormalities in the fused outcome. Visible images provide texture content and the details of the underlying structure that is in accordance with the human visual system while infrared images capture the image content in a different frequency band. The infrared imaging, captures signals having wavelength higher than that of visible light, which is not visible to the human eye. It creates images based on differences in surface temperature by detecting infrared radiation (heat) that emerges from objects and their surrounding environment and is thus often used to improve the night time vision. Visible and infrared images thus form a pair of complementary data. Thus the central motive of image fusion is to extract salient information from both the modalities and remove unnecessary details without creating artifacts in the fused image.

Image fusion plays a crucial role in video surveillance, modern military, vegetation monitoring, and satellite cloud imaging applications. It has been used profoundly in unmanned aerial vehicles for applications such as target detection, localization, military scrutiny, and forest fire control [2]. Further, it has immense applications in future autonomous automotive systems for the design of advanced driver assistance systems since the fused image would provide a perceptually meaningful image in night time and low visibility conditions [3]. Light detection and ranging (LiDAR) remote sensing system that is used to monitor span and growth of vegetation typically in terms of coverage area and height can also be facilitated by image fusion since usually the images are limited such that they capture only the top view of the canopies. Thus portable ground-based LiDAR device along with thermal imaging could help construct a 3-D thermal reconstruction of the crop that would provide a more comprehensive estimate of the growth of plants [4]. Traditionally any image fusion problem can be visualized as the formulation of a map between the input images, i.e., the visual and infrared images and the fused image so as to effectively incorporate the spatial characteristics of the visual image and the spectral characteristics of the infrared image. In the literature the image fusion problem has been addressed utilizing different schemes including multiscale transform [1], [5] [6], sparse representation [7], [8], neural network [9], [10], subspace [11], [12], saliency-based [13], [14] methods, hybrid model-based approaches [15], [16], and other methods [17], [18].

Any fusion framework involves three basic components: First, an image transformation model, second activity level measurement, and third the formulation of fusion rule. Several deep

Manuscript received September 3, 2020; revised October 17, 2020; accepted October 29, 2020. Date of publication November 3, 2020; date of current version January 6, 2021. *(Corresponding author: Snigdha Bhagat.)*

Snigdha Bhagat and Shiv Dutt Joshi are with the Department of Electrical Engineering, IIT Delhi, New Delhi 110016, India (e-mail: snigdha.bhagat@ee.iitd.ac.in; sdjoshi@ee.iitd.ac.in).

Brejesh Lall is with the Bharti School of Telecommunications Technology & Management, IIT Delhi, New Delhi 110016, India (e-mail: brejesh@ee.iitd.ac.in).

Smriti Gupta is with Mastercard, India (e-mail: Smriti.Gupta@mastercard.com).

This article has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors.

Digital Object Identifier 10.1109/JSTARS.2020.3035633

learning framework based algorithms have also been successfully applied to find a solution to the image fusion problem, because of their ability to extract image features. Li *et al.* [20] in his paper for multifocus image fusion performed decomposition of the input images into low and high-frequency components for training two separate neural networks for end-to-end training. To generate a fused low-frequency sub-band image, a siamese network is deployed to find high-level feature maps which are then used to find a fusion map for pixel-level fusion. For high-frequency sub-band fusion, a residual neural network is trained using a texture preserving loss function. On similar lines, Tang *et al.* [21] improved the algorithm by proposing a pixel level CNN that would classify focused and defocused pixels. These methods have successfully achieved robust and perceptually relevant state-of-the-art performance.

Unlike multifocus image fusion, visible and infrared images do not have a clear-cut evaluation metric since the output can only be evaluated based on perceptual quality and ground truths cannot be constructed artificially. Besides, there is a dire shortage of publicly available databases for training deep networks. Despite all these issues, there is no doubt that visible and infrared image fusion, based on convolution neural networks, is worth exploring since deep neural networks can model complex linear and nonlinear characteristics of the input images considerably. Piao *et al.* [22] in his paper has proposed a deep network for the fusion of infrared and visible images on different scales by using multiscale wavelet decomposition. In order to determine the fusion rule, the Siamese network is used that determines the saliency of each pixel from the two images. Li *et al.* [23] in his paper has fusion by first decomposing the source images into base and detail part out of which the base parts are fused using weighted averaging. The detail part is fused using the VGG network by extracting features at multiple layers in order to generate a weight map for fusion. Fused detail content is obtained by employing the max selection rule. The fused image is then constructed by a weighted combination of the base and detail content. On similar lines, Huang *et al.* [24] presented a nonsub sampled contourlet transform technique that would decompose the image into several high and low frequency sub-bands. The task of image fusion is carried out by utilizing the activity maps generated by utilizing PCNN architecture for the fusion of high frequency sub-band components and utilizing image characteristics such as phase congruency, local sharpness, and signal strength in order to fuse several low frequency sub-bands. The task of feature extraction is carried out using a pretrained network, and the formulation of activity level measurement and fusion rule requires manual intervention, thus making the process highly unreliable.

Image fusion essentially involves generating new data from the distributions of multiple input images of different modalities. Thus in the past few years, generative modeling techniques have been employed for the task of multimodal image fusion [25], [26]. Further, the above-mentioned strategies fail to provide an end-to-end solution without any manual intrusion. FusionGAN proposed by Ma *et al.* [25] tried to solve this problem by proposing an end-to-end architecture based on deep generative networks. FusionGAN posed this problem as an adversarial game where the generator performs the task of retaining the infrared image information while incorporating the gradient information from the visual image and on the other hand discriminator tries to drive the fused image closer to a visible image. This has been proposed as a mini-max problem between the generator and the discriminator. Hou *et al.* [26] also tried to overcome the same limitation in his paper and developed a dynamically adaptive end-to-end deep fusion framework called the visible and infrared image fusion network (VIF-Net), the deep network has been trained on a composite loss function that consists of M-SSIM and total variation (TV). Xu *et al.* [27] in his paper has also modeled the problem of image fusion as an adversarial process in which the generator module is fed with visible image and IR image, concatenated as two separate two channels in order to generate an initial estimate of the fused image. The discriminator module improves the information content in the fused image by formulating it as a classification module. In this article, we have employed an encoder–decoder architecture with residual connections regularised by an adversarial network; the encoder part has a downsample path that maps the input signal to a lower-dimensional space called the latent space from which the data is reconstructed by the decoder part which in turn provides an upsample path to transform the latent space to the original signal space. The network is optimized to obtain an efficient latent space representation at the output of the encoder such that the original signal can be reconstructed effectively. The residual connection, on the other hand, helps avoid the problem of model degradation caused due to increased depth of the network. The adversarial regulariser helps the autoencoder fusion network to generate a more realistic fused image that can effectively incorporate the characteristics of the visual image. Based on the study of techniques in the literature, below stated conclusions can be drawn.

1) Several multiscale transform and saliency-based techniques require activity level measurement maps to formulate the fusion rule that in turn requires manual intervention and increases the computational complexity of the algorithm.
2) The fused image needs to incorporate the texture details of the visible image and huge contrast variation of the infrared image generated. The visible image contains an enormous amount of details in comparison to the infrared image; thus, the prime focus is to retain the visible image content entirely and add the extra information from the infrared image.
3) The fused image can be expressed as a convex combination of the visible and infrared images. Thus in order to generate the distribution of the fused image, generative modeling techniques can be employed by learning from the distributions of infrared and visible images.

In view of the above-stated limitations in the literature, we propose an algorithm for image fusion that takes care of the shortcomings mentioned above. Following are some of the considerations which have gone into the design of the proposed algorithm.

1) In order to address the limitation of human intervention in the formulation of activity maps, the autoencoder based

network architecture proposed provides an end-to-end structure to obtain the fused image.

2) A novel loss function is proposed that takes into account incorporating all the characteristic of visible image and the relevant contrast information from the infrared part contrary to the preexisting techniques that give equal weightage to the input images. The texture content of the visible image has been preserved by minimizing the texture difference between the fused image and visible image and the spectral content of an infrared image has been incorporated by processing the visible and infrared image separately and merging the features at the end of each layer.
3) Since the objective is to obtain the distribution of the fused image, a generative adversarial network architecture has been employed. The generator module is an autoencoder network that performs the functionality of fusing characteristics of the visible and infrared image at several intermediate layers in the feature space. It attempts to generate the fused image as close as the visible image to fool the discriminator module that is trained to differentiate between the visible and fused images. The discriminator module is trained only on visible images since most of the detailed high-frequency information lies only in the visible image.
4) The structure of the generator, i.e., the autoencoder module is that of a residual network to facilitate the propagation of the error to the initial few layers efficiently. It also ensures the simultaneous generation of a more robust and realistic fused image. Further, it plays a significant role in avoiding the model degradation problem caused due to the increasing depth of the network.
5) The symmetric skip connections in the autoencoder structure help bypass the spectral and spatial features directly to the decoder section of the autoencoder. This is essential since the downsample operation in the encoder helps extract all the semantic information from both the input images and the upsample part aids reconstruction of a high resolution fused image.

The rest of the article is organized as follows. Section II discusses the preliminaries required for the proposed network design, and a brief introduction of the building blocks of the network has been provided. Section III introduces the proposed algorithm that includes the network structure, the loss function and all the relevant details. In Section IV, the results of the proposed technique have been compared to the other state-of-the-art techniques, and an introduction to subjective and objective techniques for fusion quality assessment has been provided.

II. Preliminaries and Methods

A. Autoencoder Constrained by Adversarial Regulariser

Generative modeling is a class of machine learning algorithms in which the network tries to learn the underlying distribution of the data from the given set of data points in order to generate new data points of the same class. Considering the data samples of a class as the training set the network tries to generate the best fit continuous distribution which, when sampled, can create new data points of the same class with some variations. But since it is not always possible to learn the exact distribution from the given data points so we try to model a distribution that can best approximate the true data. This is where neural networks come handy since they learn a function that can model the data distribution. Variational Autoencoders (VAE) and Generative adversarial networks (GAN) are the two most commonly used approaches when it comes to generative modeling. Autoencoders serve as a useful tool when we need to obtain a compressed representation such that the data point can be perfectly reconstructed from such compressed latent space representation. In this article, we propose the use of autoencoders for interpolation by a convex combination of the latent codes, which would, in turn, semantically mix the characteristics of both the input images. Since intuitively the distribution of fused image would be a weighted combination of the distribution of visible and infrared images. In this article, we propose to use adversarial training such that the autoencoder serves as the generator model, and the discriminator can be used as a classifier that would differentiate between the fused image and input images. Depending on the output, the error can be used to train the generator and the discriminator network. Adversarial training would help the generator to produce a more realistic fused image until and unless it can fool the discriminator model.

B. Residual Networks

Image classification has advanced in the past few years due to the availability of large datasets for training and powerful GPUs that has enabled the training of very deep architectures. Simonyan *et al.* [28], authors of VGG, successfully proved that accuracy can be increased by adding more and more layers to a network. Before this, in 2009, Yoshua Bengio [29] gave convincing theoretical and mathematical evidence for the effectiveness of a deeper neural network over their shallow counterparts. The residual network was first proposed by He *et al.* in [30], it was observed that as the depth of the network increases the accuracy of the network first saturates and then starts decreasing. This had nothing to do with overfitting, and thus the dropout layer could not solve this problem. It could be argued that this could be posed as an optimization problem since as the depth increases it becomes harder to train and propagate the error throughout the entire deep network due to vanishing gradients. Since the gradient is multiplied by the weight matrix at each step during back-propagation, thus if the value of gradients is small due to successive multiplications its value would diminish. Since neural networks are universal function approximators, any deep or shallow neural network should be able to learn any simple or complex functions, but due to the curse of dimensionality and vanishing gradients, deep networks often are not able to learn simple identity mapping.

Traditionally in a neural network, the output of one layer feeds the next layer, and in a residual network, the output would feed the next layer and another layer after 2–3 hops. Let us assume

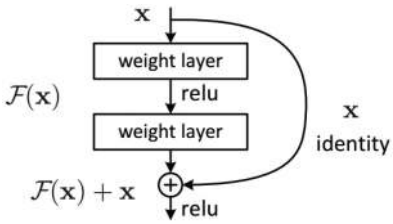


Fig. 1. Single residual block.

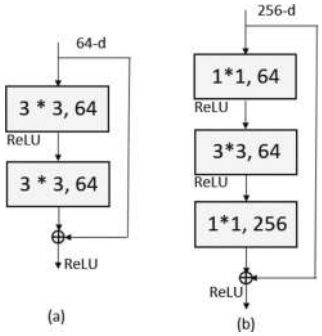


Fig. 2. (a) Basic residual building block. (b) Bottleneck unit.

the network is trying to learn the function $G(x)$ from given input x. The difference is also called as the residue between the input and output can thus be represented as $F(x) = G(x) - x$. In the case of residual networks as shown in Fig. 1 the set of stacked nonlinear layers try to learn the residual mapping $F(x)$ instead of the true desired mapping $G(x)$. To solve this issue, residual networks come in handy since they can learn simple mapping functions. As pointed out by Veit *et al.* [31] this strategy also solves the problem of vanishing gradients since the error can be propagated efficiently to the initial few layers and they can also learn as fast as the last few layers. He successfully visualized a deep residual network as an ensemble of several shallow networks with variable lengths. Thus the residual network becomes easy to optimize and can enjoy accuracy gains from significantly increased depth. ResNet can have a very deep network of up to 152 layers and still learn functions with a good accuracy since it has to learn the residual representation function instead of the signal representation.

III. Proposed Method

A. Problem Formulation

In this article, we propose a residual encoder–decoder architecture for image fusion along with a discriminator network that can perform the task of adversarial training, as shown in Fig. 3. The architecture of the proposed modules have been shown in Figs. 4 and 5. It has three primary units, namely, encoder, decoder, and the discriminator block. The generator network is the residual encoder–decoder network, which tries to fuse the features of both the modalities by combining the latent space representations at several intermediate layers, as shown in Fig. 3. Intuitively the feature maps obtained at several layers are fused in order to combine the content of both the images to generate the fused image. Thus the fused image ($\mathbf{F}$) can be represented as RAE ($\mathbf{V}$, $\mathbf{I}$) where RAE is the residual autoencoder function.

The generator network is conditioned by the reconstruction loss and a total variation loss which tries to maximize the texture content in the fused image as given in (3). The discriminator network is a classifier network that generates a scalar value to estimate the probability of a given image to be a real image and not the synthetically produced output of the generator network. Thus the classification loss due to the discriminator network is also propagated to the generator network so that it can generate a more realistic image such that the discriminator would classify it as the input visual image rather than its fused counterpart. The discriminator is fed with only the visual image and not the infrared image since most of the texture content is in the visual image the infrared image provides details which are mostly highlighted by huge contrast variations. The infrared details are incorporated in the fused image by virtue of content loss in the generator cost function. Since the objective is to gain the ability to distinguish between the real and synthetically generated samples, cross entropy loss function has been employed in order to train the generator module. The overall training target of the generator module is to minimize the following objective function, given in (1), and the discriminator module also denoted as D would, in turn, try to maximize the same. $D(\mathbf{V})$ is an estimate of the probability by the discriminator module that the input image is real and the $\mathbb{E}$ operator performs average over all such instances. $D(\mathbf{F}) \equiv D(RAE(\mathbf{V}, \mathbf{I}))$ on the other hand, gives an estimate of the probability that a synthetically generated/fake instance is real

$$\min_{RAE}\max_{D}\{\mathbb{E}[\log(1 - D(\mathbf{F})] + \mathbb{E}[\log(D(\mathbf{V})]\}. \quad (1)$$

The generator cost function is thus a combination of the reconstruction loss function also denoted as L_{content} and the classification loss of the discriminator module as denoted as $L_{\text{gen|disc}}$ which serves as a regulariser as given in (2). The functionality of the discriminator is to discriminate between real and synthetically generated data, thus it is trained to minimize the cross-entropy loss. Since cross-entropy can quantify the difference between two probability distributions for a given random variable or set of random variables the loss function can be expressed as given in (4). α and β are hyper-parameters whose values can be varied in accordance with the requirement. α in this context helps control the weightage of the regularization function, that in this case is the total variation regularization. β on the other hand controls the weightage of individual input in the fused image. The value of β can be optimized in accordance with the requirement. A high β value increases the weightage of the infrared image and vice versa. The results proposed in the article have been evaluated by taking the value of $\alpha = 0.2$ and $\beta = 0.3$. These values have been found out using the grid search method

$$L_{\text{generator}} = L_{\text{content}} + L_{\text{gen|disc}} \quad (2)$$

$$L_{\text{content}} = \beta\,\mathbb{E}[||\mathbf{F} - \mathbf{I}||_2^2] + (1 - \beta)\,\mathbb{E}[||\mathbf{F} - \mathbf{V}||_2^2] + \mathbb{E}[\alpha\,\|(\mathbf{F} - \mathbf{V})\|_{TV}] \quad (3)$$

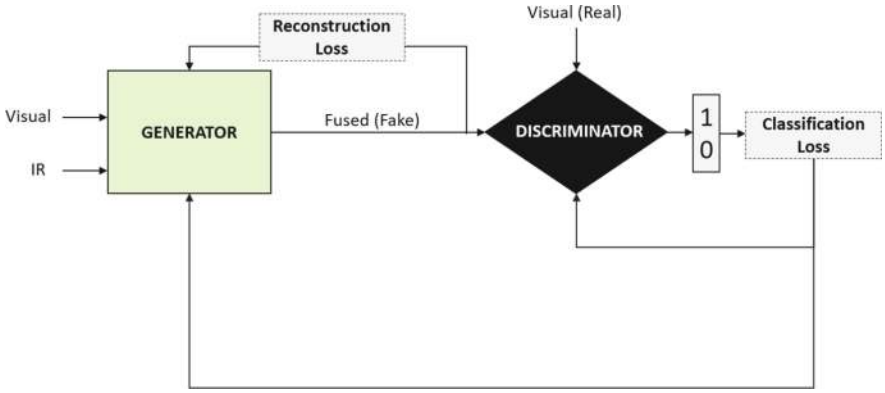


Fig. 3. Overall structure.

$$L_{\text{disc}} = -\mathbb{E}[\log(1 - D(\mathbf{F}))] - \mathbb{E}[\log(D(\mathbf{V}))] \quad (4)$$

$$L_{\text{gen|disc}} = \mathbb{E}[\log(1 - D(\mathbf{F}))]. \quad (5)$$

B. Network Architecture

The architecture of the generator module is shown in Fig. 3. The network architecture of the encoder network has been adopted from the ResNet architecture proposed by Kaiming He *et al.* [30]. The idea of combining the latent space representation in order to fuse the content of two images was inspired by works of Berthelot *et al.* [32] on interpolation of data by fusing the latent space representation of images and Jiang *et al.* [33] on semantic segmentation in which RGB image and the depth image are combined in order to segment different objects in a room.

The generator architecture is a residual autoencoder network with symmetric skip connections. Layer_1_v - Layer_4_v and Layer_1_i - Layer_4_i constitute the encoder part of the network for visible and infrared images, respectively, and Layer 5–9 is the decoder part of the architecture. Two separate encoder branches are used to encode the visible and infrared images, respectively. Different colors have been used in the architecture to denote different kinds of layers which have been elaborately explained in Fig. 4. In the encoder section, the notation $Layer_m i[n]$ or $Layer_m v[n]$ has been used to better demonstrate the layer structure where 'm' denotes the layer number and n denotes the nth Bottleneck unit in that particular layer. Similarly, n stand for nth Transbasic block in the decoder section for the visual and infrared image, respectively. The structure of the Bottleneck unit and the Transbasic block is explained in Fig. 4. The bottleneck unit has been shown alongside the basic residual unit in 2. The only difference between the two units is a $1 * 1$ convolution since as the network grows deeper performing $3 * 3$ convolution turns out to be very expensive. Thus the $1 * 1$ convolution reduces the number of input feature channels before the convolution operation and scales it back to the original feature space. This strategy helps reduce the utilization of RAM of the GPU.

The encoder section extracts several feature maps by virtue of the bottleneck layers. The bottleneck layer on the encoder side has downsample units that aids reduction in the dimensionality of input images in order to obtain a compressed latent space representation. Similarly, we have upsampled units in the decoder section (Transbasic Block with upsample operations) to transform reduced feature vector space back to the original dimensions. The upsample and downsample operations are executed by performing convolution and transpose convolution with a stride of two. The adders in the encoder section at the end of each block are for the fusion of the latent space representations at several levels. The lower half of the network architecture from Layer 5–8 followed by a final convolution and deconvolution layer compositely comprise the decoder section of the autoencoder network. The elaborate structure of the layers has been explained in Fig. 5(b)–(g). Initial four layers in the decoder block have upsampled units which increase the dimension of feature maps by a factor of two. As evident from Fig. 5 the last unit of the 5th–8th layer consists of a Transbasic block with an upsampling operation. The residual layers with upsample operation in the decoder section have inverse order in comparison to the residual layers in the encoder section with downsampling operation. The CONV and CONV TRANSPOSE blocks shown in Fig. 5(e) are the standard PyTorch convolution and convolution transpose operations applied on input image with several input channels.

Since the network is deep and we need to overcome the problem of vanishing gradient and provide an effective way of learning simple mapping functions, we have residual connections as denoted by blue lines. The agant layer in the residual connections is composed of a 1*1 convolution operator followed by the batch norm layer, which helps to reduce the dimension of feature space which in turn reduces the computation complexity. The initial structure of the discriminator module is similar to that of a branch of an encoder network. It is followed by a series of fully connected layers that complies the output of previous layers in a weighted combination to obtain a compressed representation. In the end, the output of the sequence of fully connected layers is passed through a sigmoid layer to scale it between zero to one and generate the probability of a certain image belonging to a particular class.

IV. Experimentation and Results

A. Experimental Conditions

In order to evaluate the performance of the proposed algorithm, the network architecture has been trained on image

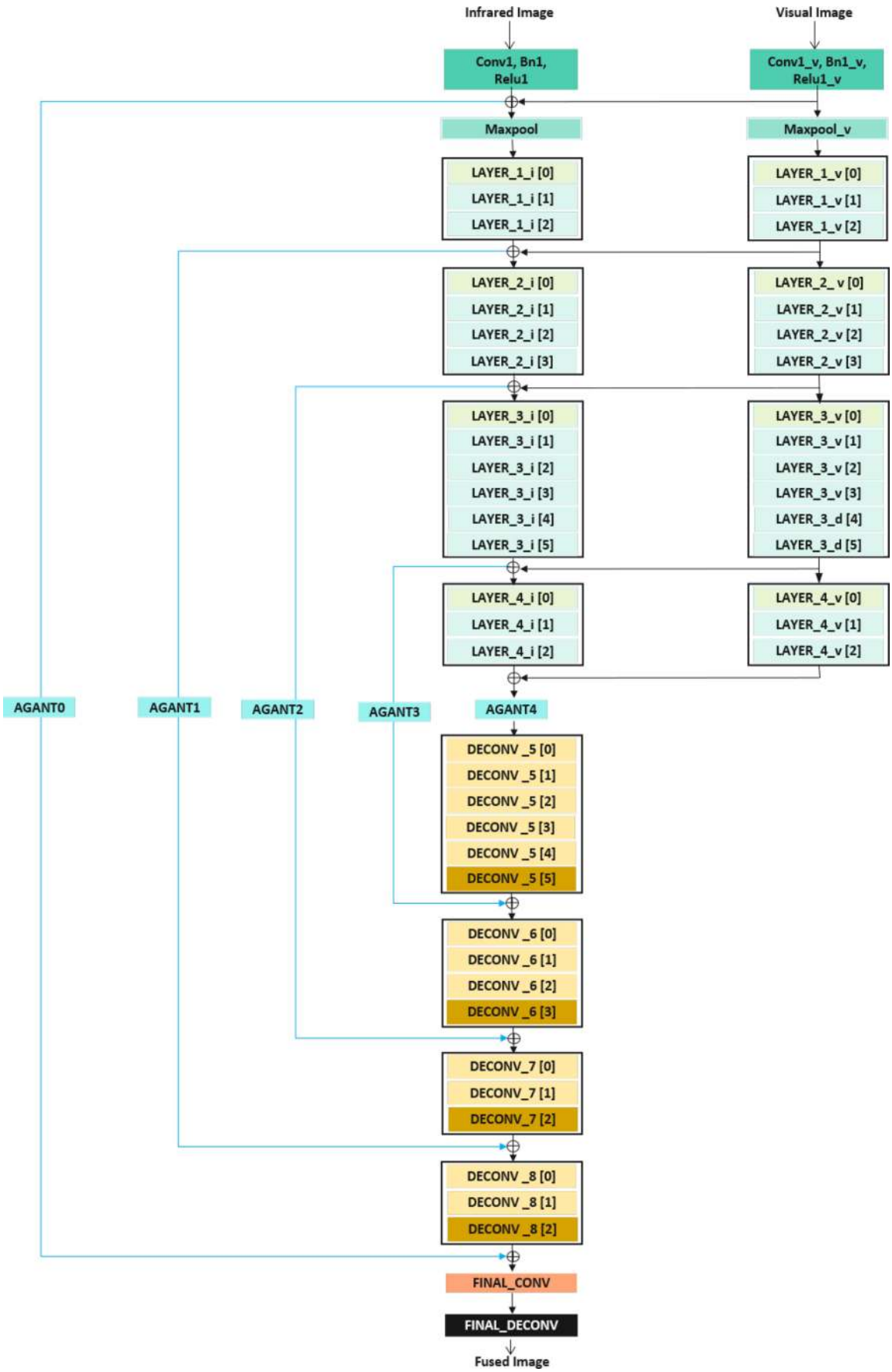


Fig. 4. Fusion autoencoder module (Generator).

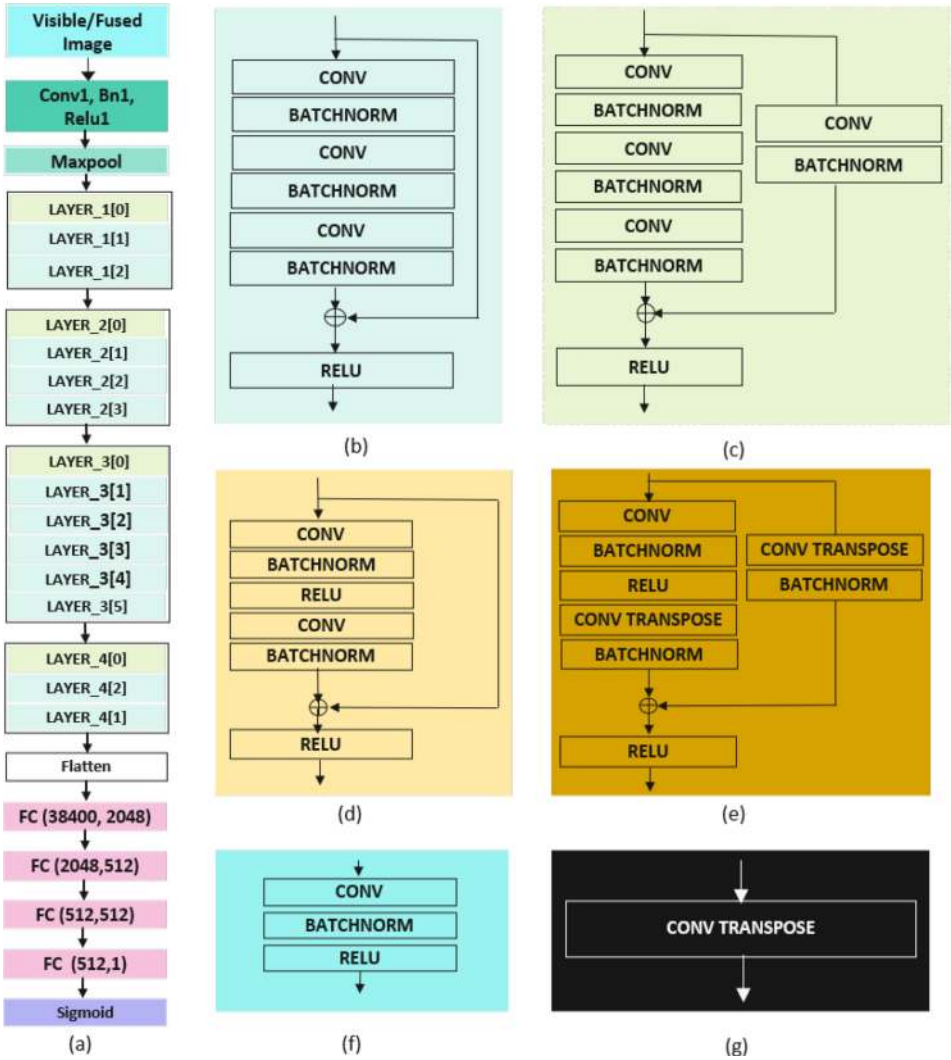


Fig. 5. (a) Network architecture of discriminator module. (b) Bottleneck block without downsampling operation. (c) Bottleneck block with downsampling operation. (d) Transbasic block without upsampling operation. (e) Transbasic block with upsampling operation. (f) Agant layer. (g) Final deconv layer.

from two separate publicly available databases. First is the TNO multiband image database [34] that consists of images in visual range (390–700 nm), near IR (700–1000 nm) & long-wave IR (8–12 μm). These images are typically geometrically registered night time images in military and surveillance scenarios comprising of huge variety in terms of type of objects, targets, and background. The TNO dataset is composed of three smaller datasets, i.e., TNO Image fusion dataset, Kayak Image fusion sequence, the TRICLOBS dynamic multiband image dataset. TNO image fusion dataset consists of images in military surveillance scenarios, the Kayak Image fusion dataset consists of images of three approaching kayaks in a cluttered maritime background and the TRICLOBS dynamic multiband dataset consists of colored motion sequences of civilian surveillance scenarios in an urban environment. The second database is the OSU Thermal database (OSU-T), it consists of 10 sequences with a total of 284 images. This database inherently captures pedestrians walking on the roads with cameras mounted on rooftops of an eight storey building with a 75 mm lens and Raytheon 300D thermal sensor core where the gain and focus are manually controlled.

These publicly available datasets have been combined and the network has been trained on a total of 1200 images. The entire dataset was divided into three parts, i.e., the training set (60%), validation set (20%) and test set (20%). All the images were scaled to a dimension of 480*640. The training process was carried out on an N-series virtual machine which was equipped with NVIDIA Tesla K80 GPU with NVIDIA GRID 2.0 technology. The fusion process was performed on a Linux based system with 56 GB RAM and 340 GB temporary local memory provided by Microsoft Azure platform in python.

B. *Performance Metrics*

In this article, we have proposed the use of information-based metrics like Entropy and Mutual Information and other parameters which can provide quantitative insight to the image quality

TABLE I
PERFORMANCE METRICS

Metric	Description	Formulation
Entropy	Entropy is a function of information content of an image. It can also be defined as average uncertainty in the image source.	$H(\mathbf{X}) = \sum_{i=1}^{255} p_k \log_2 p_k$ Where K - number of gray levels in fused image, p_k is probability associated with gray level K.
Mutual Information	It is a metric that quantifies the amount of information that one variable contains about the other.	$MI = \frac{MI(\mathbf{F},\mathbf{V}) + MI(\mathbf{F},\mathbf{I})}{2}$ $MI(\mathbf{X},\mathbf{Y}) = H(\mathbf{X}) + H(\mathbf{Y}) - H(\mathbf{X},\mathbf{Y})$
$Q^{AB/F}$	$Q^{AB/F}$ models the quality of fusion in terms of retained edge information with respect to the input images.	$Q^{VI/F} = \frac{\sum_{n=1}^{N}\sum_{m=1}^{M} Q^{VF}(n,m)w^V(n,m) + Q^{IF}(n,m)w^I(n,m)}{w^V(n,m) + w^I(n,m)}$ Where N, M are the image dimensions. $w^I(n,m)$ and $w^V(n,m)$ are the weights assigned for optimal combination of gradient information. The weights are directly proportional to the gradient magnitude information $g_V(n,m)$ in order to give more weightage to the high edge strength. $Q^{VF}(n,m) = Q_g^{AF}(n,m)Q_\alpha^{AF}(n,m)$ Q_g^{VF} quantifies the relative magnitude of edges preserved and Q_α^{VF} quantifies the relative orientation of edges preserved.
Structural Similarity index	It is a full reference perceptual metric that measures the structural similarity of two images on the basis of three impact factors, namely luminance, contrast and structure.	$SSIM(x,y) = [l(x,y)]^\alpha [c(x,y)]^\beta [s(x,y)]^\gamma$ $SSIM = \frac{SSIM(\mathbf{V},\mathbf{F}) + SSIM(\mathbf{I},\mathbf{F})}{2}$
Visual Information Fidelity	It is an information fidelity criterion that models the Shannon information shared between the reference image and the fused image by utilising the natural scene statistics, human visual system (HVS) and image degradation based priors [54].	$VIF = \sum_{j \in \text{sub-bands}} \frac{\mathbf{I}\left(\vec{\mathbf{C}}^{N,j}, \vec{\mathbf{F}}^{N,j} \middle| \mathbf{S}^{N,j}\right)}{\mathbf{I}\left(\vec{\mathbf{C}}^{N,j}, \vec{\mathbf{E}}^{N,j} \middle| \mathbf{S}^{N,j}\right)}$ The numerator quantifies the amount of distorted image information that is transmitted to brain via HVS and the denominator indicates the effectiveness of transmission of the reference signal. $\vec{C}$ denotes the random field from sub-band of reference signal and E, F are respectively the visual signals extracted by the brain after the reference image and the distorted images pass through the HVS channel.

like structural similarity index (SSIM), visual information fidelity (VIF), and $Q^{AB/F}$. SSIM is a perceptual quality based model, that models image degradation as perceived changes in the structural content of the image. visual information fidelity (VIF) is used to compare the quality of the fused image in reference to the input visual and infrared images based on natural scene statistics [52], [53], [54]. It quantifies the amount of information present in the original image. It models the amount of information that can be extracted from the fused image, which is relevant to its original image. $Q^{AB/F}$ is a gradient-based quality index that measures the amount of edge information in the fused image. A detailed description of the performance evaluation metrics has been tabulated in I.

TABLE II
OBJECTIVE SCORE OF PROPOSED FUSION METHOD ON 7 BENCHMARK IMAGE PAIRS

	VIF	$Q^{AB/F}$	SSIM	MI	Entropy
Athena	1.3922	0.3066	0.7571	3.3386	7.0596
Bench	1.1924	0.5524	0.5749	3.732	7.281
Bunker	1.3130	0.2585	0.6256	3.5013	7.0987
Tank	1.2945	0.2496	0.7470	3.9900	7.3848
Sandpath	1.1215	0.3594	0.6478	3.1729	6.8665
Nato_camp	1.1920	0.4077	0.70965	3.1738	6.8165
Kaptein	1.4673	0.2423	0.6763	3.3626	7.0055
Average	**1.2818**	**0.3395**	**0.6769**	**3.4673**	**7.0732**

VIF = Visual Information Fidelity; SSIM = Structural Similarity; MI = Mutual Information; EN = Entropy

C. Results

The performance of an image fusion algorithm cannot be judged by objectively evaluating the value of specific performance metrics since the quality of the output depends on the application in which the fused image has to be used. Thus the fusion performance is evaluated on both qualitative and quantitative metrics. The quality of the fused image can be evaluated using subjective and objective scores. The subjective scores are a function of the visual quality of the fused image and how does the image look perceptually. The perceptual quality depends on how natural the fused image looks visually, the amount of image distortion, and the visibility of texture and edge details in the fused image.

There is no absolute objective metric that can quantify the quality of the fused image; thus, a combination of several metrics are used to judge the perceptual quality of the fused image. The value of all these metrics for different categories of images in the TNO dataset was evaluated and recorded in Table II. The proposed algorithm was compared to 19 existing state-of-the-art algorithms to compare the fusion performance. It was observed that the values of SSIM and VIF were better than that of the existing techniques and the value of entropy (EN) and mutual

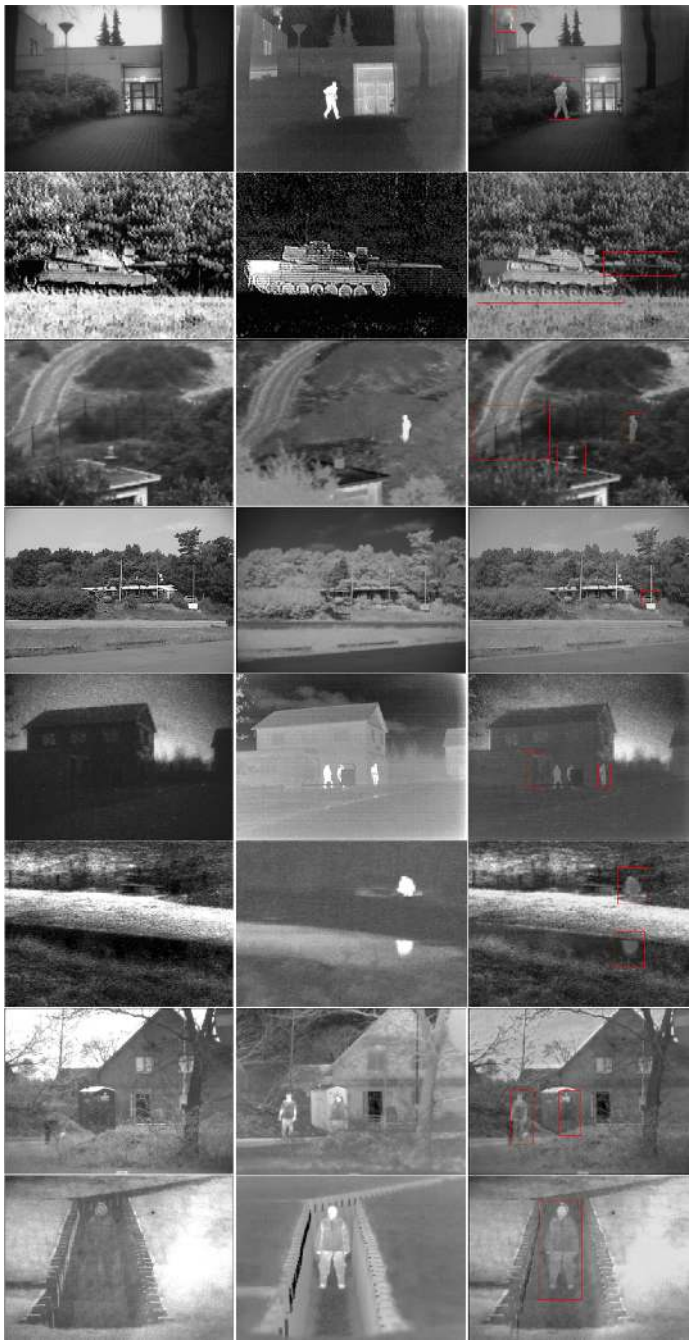


Fig. 6. Subjective results of the proposed algorithm on the TNO dataset, left to right Visual Image, Infrared image, and Fused Image.

information (MI) which are metrics to evaluate the information contained in the fused image also exceeds most of the existent algorithms. A high SSIM score along with a high mutual information indicates that the fused image resembles the input source images to a great extent and has essentially incorporated all the characteristics of the input images. An exceptionally high VIF score also indicates that the generated fused image has high perceptual similarity with the input source images as also evident from the subjective results recorded in Fig. 6 for various sets of input images. The algorithm has been tested on a variety of environmental and imaging conditions. Table II encapsulates the result obtained under a variety of conditions.

TABLE III
PERFORMANCE EVALUATION FOR STATE-OF-THE-ART ALGORITHMS

	VIF	$Q^{AB/F}$	SSIM	MI	EN
ASR [35]	0.3767	0.5125	0.4898	2.0770	6.4384
LP [36]	0.4363	0.6011	0.4938	1.9353	6.7053
NSCT [37]	0.4213	0.5753	0.4945	1.883	6.585
SCNN [22]	0.4780	**0.6181**	0.6582	2.9402	7.1697
GTF [38]	0.3440	0.3804	0.4236	2.1623	6.5819
FPDE [39]	0.3338	0.4167	0.4617	1.9024	6.3974
DDCTPCA [40]	0.3927	0.5068	0.4851	1.8382	6.5567
CBF [41]	0.3696	0.4752	0.4843	1.722	6.5989
HMSD [42]	0.3943	0.5284	0.4891	2.6005	6.9609
ADF [43]	0.3270	0.3823	0.4786	2.2094	6.3511
TSIFVS [44]	0.3632	0.5059	0.4898	1.8646	6.6270
Wavelet [45]	0.3028	0.2939	0.4869	2.4895	6.3003
IFEVIP [46]	0.4061	0.4805	0.4865	**3.8723**	6.8685
GFF [47]	0.4681	0.6180	0.4344	3.5612	6.989
FusionGAN [25]	0.2306	0.4672	0.6039	2.346	6.901
DDcGAN [48]	0.3192	0.5974	0.509	3.256	**7.340**
DenseFuse [49]	0.3501	0.4419	0.6684	2.892	6.8418
DPAL [50]	0.4034	0.3356	0.6425	3.001	7.052
TEMTD [51]	0.3627	0.3910	0.4872	1.6907	6.478
Proposed	**1.2818**	0.3395	**0.6769**	3.4673	7.0732

The images in Athena and Bench category had data points for application in military surveillance scenarios. The images were captured diurnally and in all weather conditions, they exhibited very low visibility due to smoke and shadowing effect. As evident from the results presented, the proposed algorithm works best for these conditions and exhibits a high VIF, $Q^{AB/F}$, SSIM score. The images in the tank category that exhibited low spectral sensitivity of around 8–12 μm also yielded high MI, entropy scores, that quantify the information present in the fused image. The images in the Nato camp category were captured just before dawn, when the overall environment is hazy. The images exhibited low visual and thermal contrast since all the objects in the scene would have about the same temperature after losing excess heat in the night. It can be observed that the proposed algorithm performs better in terms of subjective quality, SSIM and $Q^{AB/F}$ score in such conditions. Although the average $Q^{AB/F}$ value is not at par with other techniques for certain types of input images. It is because the proposed network focuses on encapsulating most of the characteristics of visible images and complementary behavior from its IR counterpart, the gradient information transferred from the IR to fused image is low that in turn weights down the value of the $Q^{AB/F}$ score as validated by evaluating the values of Q^{AF} and Q^{BF} separately. Thus, we can summarize that the proposed algorithm performs well in low visibility, low spectral sensitivity, and bad weather conditions producing perceptually meaningful results. Further, the basic objective of incorporating the characteristics of the visible image and the complementary information of the infrared image has also been satisfied as evident from the subjective results recorded in Fig. 6 for various sets of input images. The results of the comparison have been tabulated in Table III. The perceptual performance of other state-of-the-art algorithms has also been recorded in Fig. 7.

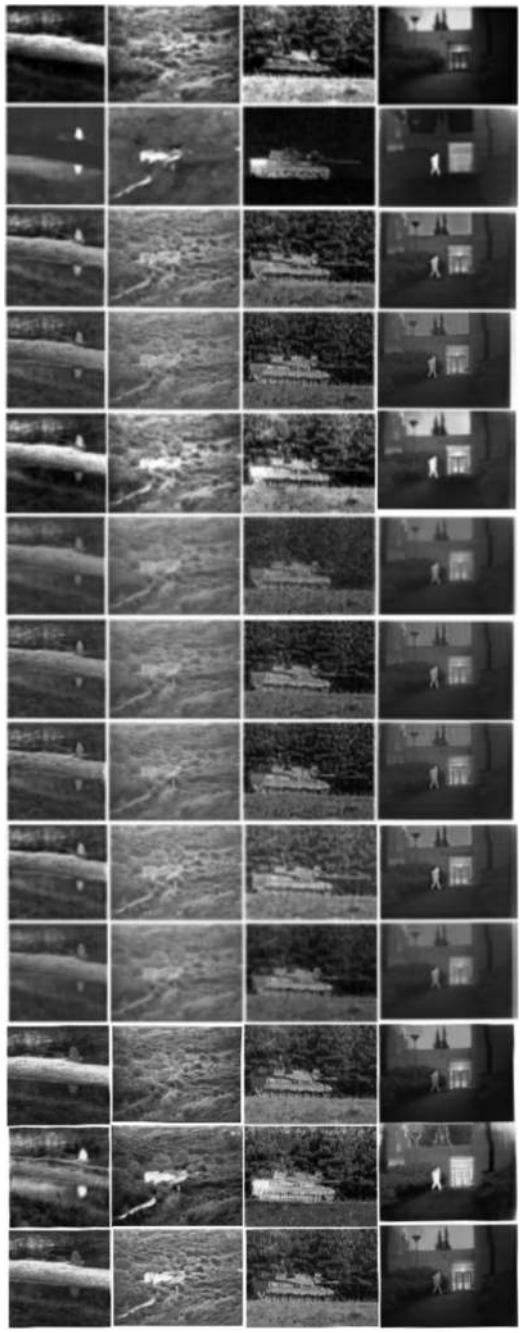


Fig. 7. Qualitative metric evaluation results on infrared and visible image pairs from the TNO database. From left to right: Bench, Bunker, Tank, Kaptein. From top to bottom: Visible, Infrared, LP, NSCT, SCNN, FPDE, DDCTPCA, CBF, HMSD, Wavelet, FusionGAN, DDcGAN, Proposed algorithm.

D. Discussion and Conclusion

In this article, we have proposed a visible, infrared image fusion network based on a generative modeling scheme that has a residual autoencoder as the generator and convolution neural network-based discriminator. The basic building block in the proposed autoencoder architecture is a residual module since it helps avoid the model degradation problem in the case of deep networks. Further, the autoencoder structure with symmetric skip connections from encoder to decoder sections proved effective to propagate the feature maps directly to the decoder section. It can incorporate the details of features obtained in the initial few layers directly to the decoder section that helps to generate a high resolution fused image. The discriminator section behaves as an adversarial regulariser that performs the task of classification of the generated fused image and input visual image. The discriminator is fed only with the visible image since it carries most of the information content, and the infrared image merely provides high contrast details of objects or individuals that are not visible in normal lighting conditions.

The structure has been designed so as to inherently performs the task of combining the spectral content from the infrared image and spatial content from the visible image. This network architecture can also be extended for various other applications like multispectral and panchromatic images and later on generalized for hyper-spectral imaging with a large number of spectral bands. The structure of the discriminator module is similar to a branch of the encoder part in the generator network so as to ensure that two networks are equally powerful. This can be intuitively explained since we are trying to perform the task of improving the feature modeling by pitting the generator and discriminator against each other. There are two basic issues that one faces while training a GAN. The first one is called the mode collapse in which the generator network becomes more powerful and tries to model all the input instances to a single true output. This is undesirable as we require that the generator output to be diverse. The other issue is faced when the discriminator network is more powerful and it learns to fool the generator network by classifying every input to the network as false. This in-turn results in less error gradient propagation and the generator fails to learn the target distribution. Thus it can be intuitively stated that we need to have the generator and discriminator module to have an approximate similar structure.

The algorithm has been evaluated on TNO and the OSU-T database and the performance metrics have been compared to 19 other state-of-the-art techniques. It has been observed that the proposed technique provides the best SSIM and VIF values. The values of entropy and mutual information are also at par with the other methods.

REFERENCES

[1] S. Li, Y. Bin, and H. Jianwen, "Performance comparison of different multi-resolution transforms for image fusion," *Inf. Fusion*, vol. 12, no. 2, pp. 74–84, 2011.

[2] Y. Zuo *et al.*, "Airborne infrared and visible image fusion combined with region segmentation," *Sensors*, vol. 17, no. 5, pp. 1127, 2017.

[3] I. Shopovska, J. Ljubomir, and P. Wilfried, "Deep visible and thermal image fusion for enhanced pedestrian visibility," *Sensors*, vol. 19, no. 17, pp. 3727, 2019.

[4] F. Narvàez, *et al.*, "LiDAR and thermal images fusion for ground-based 3D characterisation of fruit trees," *Biosyst. Eng.*, vol. 151, pp. 479–494, 2016.

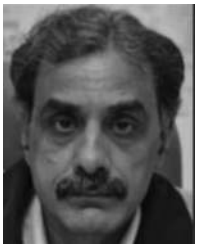
[5] G. Pajares and Jesus Manuel De La Cruz, "A wavelet-based image fusion tutorial," *Pattern Recognit.*, vol. 37, no. 9, pp. 1855–1872, 2004.

[6] Z. Zhang and S. B. Rick, "A categorization of multiscale-decomposition-based image fusion schemes with a performance study for a digital camera application," in *Proc. IEEE*, vol. 87, no. 8, pp. 1315–1326, 1999.

[7] J. Wang, J. Peng, X. Feng, G. He, and J. Fan, "Fusion method for infrared and visible images by using non-negative sparse representation," *Infrared Phys. Technol.* vol. 67, pp. 477–489, 2014.
[8] S. Li, Y. Haitao, and F. Leyuan, "Group-sparse representation with dictionary learning for medical image denoising and fusion," *IEEE Trans. Biomed. Eng.* vol. 59, no. 12, pp. 3450–3459, Dec. 2012.
[9] T. Xiang, Y. Li, and G. Rongrong, "A fusion algorithm for infrared and visible images based on adaptive dual-channel unit-linking PCNN in NSCT domain," *Infrared Phys. Technol.* vol. 69, 2015.
[10] W. Kong, Z. Longjun, and L. Yang, "Novel fusion method for visible light and infrared images based on NSST–SF–PCNN," *Infrared Phys. Technol.* vol. 65, pp. 103–112, 2014.
[11] D. P. Bavirisetti, X. Gang, and L. Gang, "Multi-sensor image fusion based on fourth order partial differential equations," in *Proc. 2017 20th Int. Conf. Inf. Fusion*, 2017, pp. 1–9.
[12] W. Kong, L. Yang, and Z. Huaixun, "Adaptive fusion method of visible light and infrared images based on non-subsampled shearlet transform and fast non-negative matrix factorization," *Infrared Phys. Technol.*, vol. 67, pp. 161–172, 2014.
[13] X. Zhang *et al.*, "Infrared and visible image fusion via saliency analysis and local edge-preserving multi-scale decomposition," *JOSA A* vol. 34, no. 8, pp. 1400–1410, 2017.
[14] J. Zhao, Y. Chen, H. Feng, Z. Xu, and Q. Li, "Infrared image enhancement through saliency feature analysis based on multi-scale decomposition," *Infrared Phys. Technol.* vol. 62, pp. 86–93, 2014.
[15] Y. Liu, L. Shuping, and W. Zengfu, "A general framework for image fusion based on multi-scale transform and sparse representation," *Inf. Fusion*, vol. 24, pp. 147–164, 2015.
[16] J. Ma *et al.*, "Infrared and visible image fusion based on visual saliency map and weighted least square optimization," *Infrared Phys. Technol.*, vol. 82, pp. 8–17, 2017.
[17] J. Ma *et al.*, "Infrared and visible image fusion via gradient transfer and total variation minimization," *Inf. Fusion*, vol. 31, 2016.
[18] J. Zhao *et al.*, "Fusion of visible and infrared images using global entropy and gradient constrained regularization," *Infrared Phys. Technol.*, vol. 81, pp. 201–209, 2017.
[19] J. Yang *et al.*, "PanNet: A deep network architecture for pan-sharpening," *Proc. IEEE Int. Conf. Comput. Vis.*, 2017.
[20] J. Li, Y. Genji, and F. Hui, "Multifocus image fusion using wavelet-domain-based deep CNN," *Comput. Intell. Neurosci.*, 2019.
[21] H. Tang *et al.*, "Pixel convolutional neural network for multi-focus image fusion," *Inf. Sci.*, vol. 433, pp. 125–141, 2018.
[22] J. Piao, C. Yunfan, and S. Hyunchul, "A new deep learning based multi-spectral image fusion method," *Entropy*, vol. 21, no. 6, p. 570, 2019.
[23] H. Li, Xiao-Jun Wu, and K. Josef, "Infrared and visible image fusion using a deep learning framework," in *Proc. 2018 24th Int. Conf. Pattern Recognit.*, pp. 2705–2710, 2018.
[24] X. Huang, G. Qi, H. Wei, Y. Chai, and J. Sim, "A novel infrared and visible image information fusion method based on phase congruency and image entropy," *Entropy*, vol. 21, no. 12, 2019, Art. no. 1135.
[25] J. Ma *et al.*, "FusionGAN: A generative adversarial network for infrared and visible image fusion," *Inf. Fusion*, vol. 48, pp. 11–26, 2019.
[26] R. Hou *et al.*, "VIF-Net: An unsupervised framework for infrared and visible image fusion," *IEEE Trans. Comput. Imag.*, vol. 6, pp. 640–651, Art. no. 2020.
[27] D. Xu *et al.*, "Infrared and visible image fusion with a generative adversarial network and a residual network," *Appl. Sci.*, vol. 10, no. 2, 2020, Art. no. 554.
[28] K. Simonyan and Z. Andrew, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[29] Y. Bengio, "Learning deep architectures for AI," *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, 2009.
[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Patt. Recognit.*, 2016.
[31] A. Veit, J. W. Michael, and B. Serge, "Residual networks behave like ensembles of relatively shallow networks," *Adv. Neural Inf. Process. Syst.*, 2016.
[32] D. Berthelot *et al.*, "Understanding and improving interpolation in autoencoders via an adversarial regularizer," 2018, *arXiv:1807.07543*.
[33] J. Jiang *et al.*, "RedNet: Residual encoder-decoder network for indoor RGB-D semantic segmentation," 2015, *arXiv:1806.01054*.
[34] A. Toet, "The TNO multiband image data collection," *Data in Brief*, vol. 15, p. 249, 2017.
[35] Y. Liu and W. Zengfu, "Simultaneous image fusion and denoising with adaptive sparse representation," *IET Image Process.* vol. 9, no. 5, pp. 347–357, 2014.
[36] Z. Liu *et al.*, "Image fusion by using steerable pyramid," *Pattern Recognit. Lett.*, vol. 22, no. 9, pp. 929–939, 2001.
[37] K. He *et al.*, "Infrared and visible image fusion based on target extraction in the nonsubsampled contourlet transform domain," *J. Appl. Remote Sens.*, vol. 11, no. 1, 2017, Art. no. 015011.
[38] J. Ma *et al.*, "Infrared and visible image fusion via gradient transfer and total variation minimization," *Inf. Fusion*, vol. 31, 2016.
[39] K. Vanitha, D. Satyanarayana, and M. N. G. Prasad, "A new hybrid medical image fusion method based on fourth-order partial differential equations decomposition and DCT in SWT domain," in *Proc. 10th Int Conf Comput. Comm. Netw. Technol* ., 2019.
[40] V. P. S. Naidu, "Hybrid DDCT-PCA based multi sensor image fusion," *J. Opt.*, vol. 43, no. 1, pp. 48–61, 2014.
[41] X. Yan *et al.*, "Infrared and visible image fusion with spectral graph wavelet transform," *JOSA A*, vol. 32, no. 9, pp. 1643–1652, 2015.
[42] Z. Zhou *et al.*, "Perceptual fusion of infrared and visible images through a hybrid multi-scale decomposition with Gaussian and bilateral filters," *Inf. Fusion*, vol. 30, pp. 15–26, 2016.
[43] D. P. Bavirisetti and D. Ravindra, "Fusion of infrared and visible sensor images based on anisotropic diffusion and Karhunen-Loeve transform," *IEEE Sensors J.*, vol. 16, no. 1, pp. 203–209, Jan. 2016.
[44] D. P. Bavirisetti and D. Ravindra, "Two-scale image fusion of visible and infrared images using saliency detection," *Infrared Phys. Technol.*, vol. 76, pp. 52–64, 2016.
[45] X. Yan *et al.*, "Infrared and visible image fusion with spectral graph wavelet transform," *JOSA A* vol. 32, no. 9, pp. 1643–1652, 2015.
[46] Y. Zhang *et al.*, "Infrared and visual image fusion through infrared feature extraction and visual information preservation," *Infrared Phys. Technol.*, vol. 83, pp. 227–237, 2017.
[47] S. Li, K. Xudong, and H. Jianwen, "Image fusion with guided filtering," *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2864–2875, Jul. 2013.
[48] J. Ma *et al.*, "DDcGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 4980–4995, 2020.
[49] H. Li and X. Wu, "DenseFuse: A fusion approach to infrared and visible images," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2614–2623, May 2019.
[50] J. Ma *et al.*, "Infrared and visible image fusion via detail preserving adversarial learning," *Inf. Fusion*, vol. 54, pp. 85–98, 2020.
[51] J. Chen *et al.*, "Infrared and visible image fusion based on target-enhanced multiscale transform decomposition," *Inf. Sci.*, vol. 508, pp. 64–78, 2020.
[52] CS. Xydeas and P. Vladimir, "Objective image fusion performance measure," *Electr. Lett.*, vol. 36, no. 4, pp. 308–309, 2020.
[53] Z. Wang *et al.*, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.
[54] H. R. Sheikh and Alan C. Bovik, "A visual information fidelity approach to video quality assessment," in *Proc. Ist Int. Workshop Video Process. Qual. Metrics Consum. Electron.*, vol. 7, Jan. 2005.

Snigdha Bhagat received the B.Tech. degree in electronics and communication engineering from the National Institute of Technology, Aizawl, India, in 2014, and the M.Tech. degree in communication engineering from Indian Institute of Technology (IIT) Delhi, New Delhi, India, in 2014. She is currently working toward the Ph.D. degree with IIT Delhi, with the dissertation based on quantification and fusion of differential information.

In due course of Ph.D. degree, she has been a part of industrial projects on face recognition and vehicle number plate recognition. She is currently a Teacher Trainee with VNIT Nagpur, Nagpur, India. Her research interests include signal processing, image processing, and deep learning.



Shiv Dutt Joshi (Member, IEEE) received the B.E. degree (Hons.) in electrical and electronics engineering from the Birla Institute of Technology, Pilani, India, in 1981, and the M.Tech. degree in communications and radar engineering, and the Ph.D. degree in signal processing from IIT Delhi, New Delhi, India, in 1983 and 1988, respectively.

He is currently a Professor with the Electrical Engineering Department, IIT Delhi. His research interests include the development of fast algorithms for stochastic signal processing, speech processing, modeling of stochastic processes, and group theoretical approach to signal processing.

Brejesh Lall (Member, IEEE) received the B.E. and the M.E. degrees in electronics and communication from Delhi College of Engineering, DU Delhi, India, in 1991, and 1992, respectively. He completed Ph.D. degree, in 1997 from IIT Delhi in the area of multirate signal processing. During the Ph.D. he worked on "some studies on characterization and modeling of stochastic processes in the multiscale framework."

He joined Hughes Software Systems, in 1997 and worked there for nearly eight years with the Signal Processing Group. He returned to his alma mater and joined IIT Delhi as a faculty member, in 2005. Since July 2005, he has been in the Electrical Engineering Department and has contributed to research and teaching in the general area of Signal Processing. He has successfully completed numerous sponsored projects and consultancies and is working on several others. He is the current head of Bharti School of Telcom Technology and Management, and the co-ordinator of two centers of excellence, viz. Airtel IIT Delhi Centre of Excellence in Telecommunications and Ericsson IIT Delhi 5G Center of Excellence.

Smriti Gupta completed the B.Tech. degree, in 2017 in civil engineering from Indian Institute of Technology, Delhi, India.

She has more than three years of experience in Data Analysis and Data Science domain. She joined Mastercard's Data and Services Team as Data Analyst in 2017. She switched to Mastercard's A.I. Garage as a Specialist, in 2019. Her projects include time series forecasting for mastercard products, delinquency score, and retrial prediction.