FusionDN: A Unified Densely Connected Network for Image Fusion

Han Xu,[1] **Jiayi Ma,**[1,*] **Zhuliang Le,**[1] **Junjun Jiang,**[2] **Xiaojie Guo**[3]
[1]Electronic Information School, Wuhan University, Wuhan 430072, China
[2]School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China
[3]College of Intelligence and Computing, Tianjin University, Tianjin 300350, China
{xu_han, lezhuliang}@whu.edu.cn, {jyma2010, xj.max.guo}@gmail.com, junjun0595@163.com

Abstract

In this paper, we present a new unsupervised and unified densely connected network for different types of image fusion tasks, termed as *FusionDN*. In our method, the densely connected network is trained to generate the fused image conditioned on source images. Meanwhile, a weight block is applied to obtain two data-driven weights as the retention degrees of features in different source images, which are the measurement of the quality and the amount of information in them. Losses of similarities based on these weights are applied for unsupervised learning. In addition, we obtain a single model applicable to multiple fusion tasks by applying elastic weight consolidation to avoid forgetting what has been learned from previous tasks when training multiple tasks sequentially, rather than train individual models for every fusion task or jointly train tasks roughly. Qualitative and quantitative results demonstrate the advantages of FusionDN compared with state-of-the-art methods in different fusion tasks.

Introduction

Image fusion is a significant branch of image enhancement and has been applied in military and civilian fields, such as computer vision and surveillance. Due to limitations of devices and techniques, the image captured by one type of sensor or under a single shooting setting cannot characterize all the information in the scenario. For example, visible images are affected by the light condition and infrared images cannot present details (Ma, Ma, and Li 2019; Ma et al. 2016). Images taken by common devices can only capture details within a very limited range, resulting in under/over-exposure regions. Also, under a certain focal setting of optical lens, only the objects within the depth-of-field (DOF) have sharp appearances while others are blurred (Liu et al. 2017b). Under these circumstances, more images taken by different sensors or under different settings are needed to describe the scenario in an all-round way. However, the large amount of redundant information in these images is a waste of storage space. Thus, the target of fusion is to preserve the vital information in them and merge it into a single image. Then the fused result with more high-quality information can provide better visual perception.

According to imaging principles, image fusion tasks can be broadly divided into two categories, such as fusing images obtained by multi-modality imaging and by digital photography. Many traditional methods have been proposed to solve these kinds of fusion problems respectively. These methods can be divided into spatial-domain methods and transform-domain methods (Zhang et al. 2020). In spatial-domain methods, fusion is completed based on small blocks or regions (Zhang, Bai, and Wang 2017). Transform-domain methods transform source images to other domains, and the fusion process is completed in these transformed domains, including multi-scale transform (e.g., pyramid, wavelet, shearlet, discrete coscine transform), sparse representation (Zhang et al. 2018), hybrid (Paramanandham and Rajendiran 2018), subspace and other methods.

In recent years, due to the strong ability of extracting image features, deep learning has been successfully applied to image fusion. i) For multi-modality images, (Liu et al. 2017a) proposed a siamese convolutional network to generate a weight map for fusing medical images. In (Li and Wu 2018), the encoding network and the decoder are designed to extract and fuse features for the infrared and visible image fusion and the method is also applied to multi-focus image fusion. Also in infrared and visible image fusion, FusionGAN (Ma et al. 2019) and its variants (Ma et al. 2020; Xu et al. 2019) established an adversarial game between a generator and a discriminator. The discriminator forces fused images to have more details in visible images. In remote sensing image fusion, (Masi et al. 2016) proposed a CNN for projection, mapping, and reconstruction to solve the pansharpening problem. ii) For digital photography images, for the first time, (Prabhakar, Srikar, and Babu 2017) introduced deep learning in multi-exposure image fusion by building a novel CNN and applying MEF-SSIM to realize unsupervised learning. In (Liu et al. 2017b), for fusing multi-focus images, a deep CNN trained by high-quality image patches and corresponding artificially made blurred versions is adopted to encode a direct mapping between source images and the focus map.

However, for different types of source images, the vital in-

[*]Corresponding author.

formation varies greatly. For example, as representations of multi-modality images, infrared images represent the thermal radiation information with high contrast pixel intensities, while visible images mainly capture the reflected light information with abundant gradient variations. In images obtained by digital photography, features to be extracted are the objects with clearer representation. Nevertheless, the extraction process is difficult to implement in accordance with a uniform rule. Some deep learning-based methods solve it by training the model on the dataset of one fusion task and applying the trained model to other tasks. But due to lack of training on other datasets, the fused results are unsatisfactory. Moreover, the major stumbling block in utilizing deep learning for image fusion is the lack of ground-truth fused images for supervised learning. Some methods solve this problem by artificially making ground-truth images. However, for the image fusion problem, sometimes there is no uniform standard to measure whether the artificially made ground-truth images are appropriate or not. And it is not only time-consuming and demanding but also difficult to be universal for different fusion tasks.

To overcome these challenges, in this paper, we propose a unified densely connected network for image fusion that overcomes catastrophic forgetting, termed as *FusionDN*. Given two source images, the densely connected network is applied to generate the fused image. Meanwhile, a weight block is applied to obtain two data-driven weights as the retention degrees of features in different source images. Thus, for all fusion tasks, ground-truth fused images are not required. In addition, rather than train different models for different fusion tasks individually, we obtain a single model applicable to multiple fusion tasks by applying elastic weight consolidation (EWC) to avoid forgetting what has been learned from previous tasks when training multiple tasks sequentially. Both qualitative and quantitative results reveal the advantages of FusionDN compared with state-of-the-art methods.

Contributions of our work include the following aspects:

- Considering the lack of ground-truth images as the stumbling block in image fusion, we propose a new unsupervised network for image fusion. Since the loss function is data-driven, the network can be applied to different fusion tasks, i.e., it is a unified network for image fusion.
- We implement a single model to accomplish different fusion tasks. It overcomes the disadvantage of training the model only on single fusion task in existing methods, and overcomes the storage and computation issues, or catastrophic forgetting. Thus it is not only a unified framework, but also a unified model for several fusion tasks. The code is available at: https://github.com/hanna-xu/FusionDN.

Proposed Method

Problem Formulation

First of all, it should be noted that there are differences between source images of different fusion tasks. Some of them are single-channel and some are three-channel (usually RGB) images. If source images are three-channel data, we convert them from RGB to YCbCr color space. We are devoted to fusing the Y channel (luminance channel) values, as the structural details and the brightness variation are in this channel. And values of Cb and Cr channels (chrominance channels) are fused in a traditional way. Then, the fused components of these channels are transferred to RGB color space to obtain the final fused image. In this way, all fusion problems are unified into the single-channel image fusion.

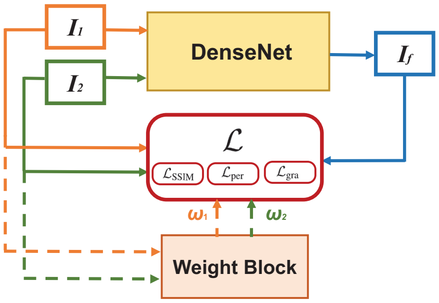
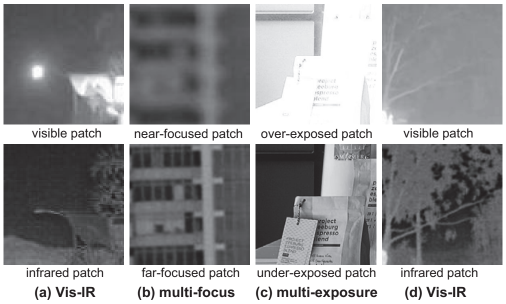

I_1
I_2
DenseNet
I_f
$\mathcal{L}$
$\mathcal{L}_{\text{SSIM}}$
$\mathcal{L}_{\text{per}}$
$\mathcal{L}_{\text{gra}}$
ω_1
ω_2
Weight Block

Figure 1: Overall procedure of our proposed FusionDN.

Given two single-channel source images I_1 and I_2, since the vital information in different types of images varies greatly, as a unified framework, it is difficult to predetermine the features to be extracted and fused. In view of this situation, rather than designing feature extraction and reconstruction methods, from a new perspective we determine the retention degrees of features of different source images in the fused image according to their own properties. Since this retention degree varies with the specific source image, our method is a data-driven method by applying two data-driven weights, i.e., ω_1 and ω_2. They are determined by the specific properties of images instead of pre-setting artificially in advance. As shown in Figure 1, the weight block is employed to generate the weights of different source images and then feed them into the loss function of DenseNet. The DenseNet is trained to extract and reconstruct the features of source images according to the weights and sub loss functions.

As for assessing the weight of each source image, the primary consideration is to preserve the higher-quality information in images with a higher weight. For example, it is embodied in the regions with less noise in visible images compared with corresponding infrared images, the objects within the DOF with sharper appearance in multi-focus images, the objects with more suitable brightness and less distortion in multi-exposure images and so on, as shown in Figure 2(a)-(c). Therefore, to evaluate the quality of the information contained in each source image, deep neural networks for image quality assessment (IQA) (Bosse et al. 2017) is employed here to realize the assessment. For example, it assesses whether the source image quality declines due to problems such as Gaussian blur, noise, compression and local block-wise distortions of different intensity. Moreover, because the original high-quality images are difficult to obtain or these images do not really exist, we employ the no-reference (NR) model instead of the full-reference model. Then, we can obtain two image quality scores, i.e., IQA_1 and IQA_2, of I_1 and I_2, correspondingly.



(a) Vis-IR (b) multi-focus (c) multi-exposure (d) Vis-IR

Figure 2: Some example patches of different fusion tasks.

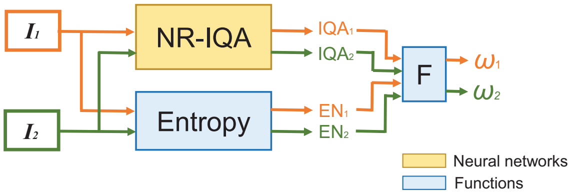


Figure 3: Illustration of the specific process of weight block.

Nevertheless, there is a problem that IQA is merely the evaluation of image quality regardless of other aspects of the image. A typical example is shown in Figure 2(d). The visible patch has a higher image quality compared with the infrared patch but the infrared patch has a more complete scenario representation. Intuitively, we prefer that the information in the infrared patch be preserved more in the fused image. This is the reflection of another fusion criterion. In theory, the more information from source images preserved in the fused image, the better. While the amount of information is not within the measurement of image quality. To tackle this problem, in addition to NR-IQA, we apply the objective metric entropy to measure the amount of information in each source image on the basis of information theory. Mathematically, it is defined as follows:

$$EN = -\sum_{l=0}^{L-1} p_l \log_2 p_l, \tag{1}$$

where L is the number of gray levels and generally set as 256. p_l is the probability of the corresponding level. On the one hand, the larger value of EN means that there is more information contained. On the other hand, EN can be affected by noise easily. Solely relying on EN to assign weights may result in a great deal of noise and distortions in the fused result. Complementarily, IQA can assess the noise and other problems of reducing image quality.

Therefore, the two metrics IQA and EN can complement and make up for each other's weakness. So we take both of the quality and the amount of information into consideration to get a more comprehensive evaluation criterion. With the function F applied for some subsequent operations to determine the final weights ω_1 and ω_2, the specific process in the weight block can be illustrated as Figure 3.

Loss Functions

With a weight λ to control the trade-off between the quality and the amount of information in source images, we can obtain their respective scores, i.e., s_1 and s_2:

$$s_1 = IQA_1 + \lambda EN_1, \quad s_2 = IQA_2 + \lambda EN_2. \tag{2}$$

The final weights are assigned according to the scores. Since the difference between the scores is much smaller compared with the values themselves, the weights obtained by direct normalization cannot reflect the difference between them. Thus, to enhance and embody the difference in weights, s_1 and s_2 are exponentially stretched. With a positive number c to scale values and the subsequent normalization processing, the final weights of source images can be defined as follows:

$$\omega_1 = \frac{\exp(\frac{s_1}{c})}{\exp(\frac{s_1}{c}) + \exp(\frac{s_2}{c})}, \ \omega_2 = \frac{\exp(\frac{s_2}{c})}{\exp(\frac{s_1}{c}) + \exp(\frac{s_2}{c})}. \tag{3}$$

Eqs. (2)-(3) are the operations in F in Figure 3. ω_1 and ω_2 are employed in the loss function of DenseNet to control the retention degrees of features in different source images.

The higher the retention degree, the higher the similarity between the fused image and the source image. As for constraining the similarity between different images, structural similarity index measure (SSIM) is the most widely used metric which models the loss and distortion according to the similarities in light, contrast and structure information (Wang et al. 2004). Mathematically, SSIM between images x and y can be defined as follows:

$$\text{SSIM}_{x,y} = \sum_{x_i, y_i} \frac{2\mu_{x_i}\mu_{y_i} + C_1}{\mu_{x_i}^2 + \mu_{y_i}^2 + C_1} \cdot \frac{2\sigma_{x_i}\sigma_{y_i} + C_2}{\sigma_{x_i}^2 + \sigma_{y_i}^2 + C_2} \cdot \frac{\sigma_{x_i y_i} + C_3}{\sigma_{x_i}\sigma_{y_i} + C_3}, \tag{4}$$

where μ denotes the mean value, σ denotes the standard deviation/covariance, C_1, C_2 and C_3 are the parameters to make the metric stable. Thus, the loss of DenseNet based on SSIM can be formulated as:

$$\mathcal{L}_{\text{SSIM}} = \omega_1(1 - \text{SSIM}_{I_f,I_1}) + \omega_2(1 - \text{SSIM}_{I_f,I_2}). \tag{5}$$

While in some task, the features to be extracted may be the image itself. For example, the features to be preserved in multi-focus images are the region within the DOF. In other words, the fused image is expected to reconstruct the focused region in the source image and not to mix the blurred information in the other one as much as possible. For such high standard reconstruction problem, it is not enough to solely rely on the three aspects of constraints defined in SSIM. So, in additional to the pixel-wise similarity, we also encourage I_f to represent the high perceptual similarity with I_1 and I_2. Therefore, we adopt the pre-trained VGG-16 network as the feature extractor and duplicate the single-channel I_1 and I_2 to make RGB channels before feeding them into the VGG network. The output of the convolutional layers before max-pooling layers is the extracted feature map used in the perceptual loss, which can be defined as:

$$\mathcal{L}_{\text{per}}(x, y) = \sum_j \frac{1}{H_j W_j C_j} \|\phi_j(x) - \phi_j(y)\|_2^2 \tag{6}$$

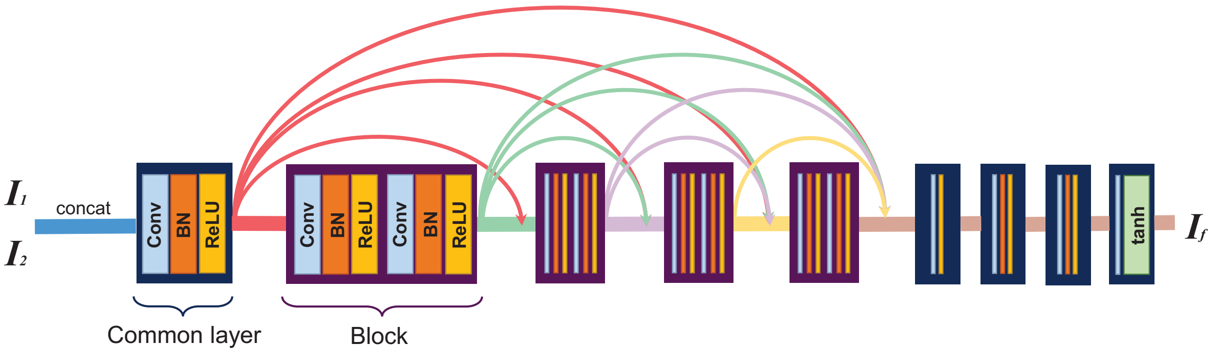


Figure 4: Architecture of the densely connected network (DenseNet). BN: batch normalization. ReLU/tanh: activate function.

where $\phi_j(x)$ are the extracted feature map by the convolutional layer before the jth max-pooling layer of size $H_j \times W_j \times C_j$. Then the perceptual loss of DenseNet can be specifically defined as:

$$\mathcal{L}_{\text{per}} = \omega_1 \mathcal{L}_{\text{per}}(I_f, I_1) + \omega_2 \mathcal{L}_{\text{per}}(I_f, I_2). \tag{7}$$

In the perceptual loss, features extracted from higher layers in VGG-16 cannot reconstruct detail information, such as texture and exact shape, but only content and overall spatial structure. In addition, in order to make the fused image exhibit sharper appearance, we add another term to constrain the gradient difference to slightly improve the results. The gradient loss is defined by the squared Frobenius norm between the gradient variations of images x and y as Eq. (8) and the gradient loss of DenseNet can be obtained as Eq. (9):

$$\mathcal{L}_{\text{gra}}(x, y) = \frac{1}{HW} \|\nabla x - \nabla y\|_F^2, \tag{8}$$

$$\mathcal{L}_{\text{gra}} = \omega_1 \mathcal{L}_{\text{gra}}(I_f, I_1) + \omega_2 \mathcal{L}_{\text{gra}}(I_f, I_2). \tag{9}$$

These three components compose the loss function. With α and β controlling the trade-off, it can be defined as:

$$\mathcal{L} = \mathcal{L}_{\text{SSIM}} + \alpha \mathcal{L}_{\text{per}} + \beta \mathcal{L}_{\text{gra}}. \tag{10}$$

Network Architecture of DenseNet

There are five common convolutional layers and four blocks in our network DenseNet to generate I_f. The input of the network is the concatenated I_1 and I_2. As for the first six layers, since it has been proven that CNNs can be significantly deeper and can be trained efficiently if they contain shorter connections between layers close to the input and those close to the output, we employ the densely connected layers from densely connected convolutional networks (Huang et al. 2017) in our DenseNet. Short direct connections are built between each layer and all layers in a feed-forward fashion, as shown in Figure 4, which are able to address the issue of vanishing gradients and strengthen feature propagation while substantially reducing the number of parameters in the network (Zhang, Sindagi, and Patel 2018). Then, the features extracted by these layers are fed into subsequent four common layers to reduce the channels of feature maps gradually and generate the final fused image.

Table 1: Input/output channels of all convolutional layers.

	Input channels		Output channels	
Common layer 1	2		48	
Block 1	conv1	conv2	conv1	conv2
	48	48	48	48
Block 2	conv1	conv2	conv1	conv2
	96	48	48	48
Block 3	conv1	conv2	conv1	conv2
	144	48	48	48
Block 4	conv1	conv2	conv1	conv2
	192	48	48	48
Common layer 2	240		240	
Common layer 3	240		128	
Common layer 4	128		64	
Common layer 5	64		1	

Moreover, in the densely connected layers, we apply the block composed by two convolutional layers to replace the common convolutional layer. By introducing the additional convolutional layer, the block can be trained to learn higher level features, which are used in short direct connections. In this way, the number of parameters can be reduced compared with building short connections for all convolutional layers when we deepen our DenseNet.

The specific settings of all layers are shown in Table 1. To avoid information loss, the reflect padding is applied before convolution. All kernel sizes are set as 3×3 and all strides are set as 1 with no pooling layers. To accelerate deep network training by reducing internal covariate shift, batch normalization is applied.

Single Model for Multi-Fusion Task with EWC

The diversity of source images in different fusion tasks leads to the difference of the features extracted by the network, which is directly reflected in parameters. Therefore, rather than training different models with the same architecture for different tasks individually, we are committed to training a single model for all fusion tasks. In other words, we prefer to employ a single model to continuously learn different tasks without forgetting what has been learned from previous tasks. In this way, all the trained fusion tasks can be

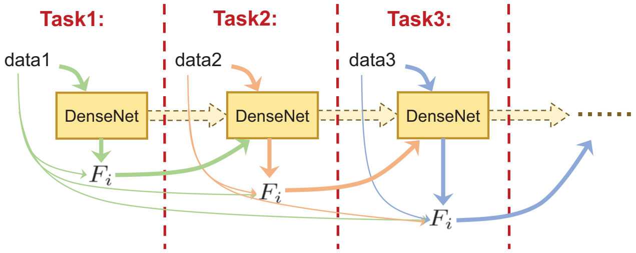


Figure 5: Intuitive description of data flow between different tasks and the process of EWC. Thin lines indicate that only a small subset of data are kept, which are merely used to calculate F_i and not applied to train DenseNet.

accomplished with the same set of parameters.

The simplest solution is to jointly train multiple tasks, and it has been shown to be much more effective than sequentially training. Nevertheless, as the number of tasks increases, there are two urgent problems. One is that always keeping the data of previous tasks results in the storage issue. The other one is that using all the data for training causes the computation issue, including the difficulty and time cost of computation. Therefore, we adopt the strategy of sequentially training different fusion tasks instead of jointly training. To avoid the most serious catastrophic forgetting problem caused by sequentially training, the EWC algorithm is applied to safeguard against this (Kirkpatrick et al. 2017).

In addition to the loss of each task itself, i.e., $\mathcal{L}(\theta)$ in Eq. (11), an extra regularization term is also included in the total loss function of the current task $\mathcal{L}'(\theta)$. To retain what the network has learned from previous tasks, the squared distances between parameter values of the current task θ_i and those of the previous task θ_i^* are added up to form this term. The subscript i represents the ith parameter in the network.

$$\mathcal{L}'(\theta) = \mathcal{L}(\theta) + \gamma \textstyle\sum_i F_i(\theta_i - \theta_i^*)^2, \tag{11}$$

where F_i represents the penalty value of corresponding squared distance. If θ_i^* has an important impact on what has been learned before, the squared distance between θ_i and θ_i^* should be small and we give such squared distance a larger penalty value F_i. F_i can be assigned as the diagonal terms of Fisher information matrix and can be easily approximated by computing square of gradients with the data of previous tasks, as defined in Eq. (12):

$$F_i = \mathbb{E}\left[\left(\frac{\partial}{\partial \theta_i^*} \log p(D^*|\theta^*)^2 | \theta^* \right] , \tag{12}$$

where D^* represents the data of previous tasks. Thereinto, $\log p(D^*|\theta^*)$ can be approximately replaced by $-\mathcal{L}(\theta^*)$. With γ to control the trade-off, the loss function of current task can be formulated as Eq. (11). For multiple tasks, the sequential training process and the data flow can be intuitively shown in Figure 5. In our model, θ are the parameters in DenseNet and $\mathcal{L}(\theta)$ is the loss function defined in Eq. (10). Then, Eq. (11) can be specifically converted as follows:

$$\mathcal{L}' = \mathcal{L}_{\text{SSIM}} + \alpha\mathcal{L}_{\text{per}} + \beta\mathcal{L}_{\text{gra}} + \gamma \textstyle\sum_i F_i(\theta_i - \theta_i^*)^2. \tag{13}$$

DenseNet is designed for fusing images of one channel. If the source images are three-channel images, the fused Y channel value can be obtained by DenseNet. And as for values of chrominance channels (Cb and Cr channels), if C_1 and C_2 are not equal to τ, they can be fused as follows:

$$\text{C}_f = \frac{\text{C}_1(|\text{C}_1 - \tau|) + \text{C}_2(|\text{C}_2 - \tau|)}{|\text{C}_1 - \tau| + |\text{C}_2 - \tau|}, \tag{14}$$

Otherwise, C_f can be directly set as τ. C_1 and C_2 are the Cb/Cr channel values of two source images, and C_f is the corresponding fused channel value of the fused image. τ is set as 128. Then, the final fused image in RGB color space can be obtained by the corresponding color space transform formula.

Experimental Results and Discussions

Training Details

We perform FusionDN with EWC on three fusion tasks: 1) visible and infrared image fusion; 2) multi-exposure image fusion; and 3) multi-focus image fusion. These are the specific tasks in Figure 5. The training and test sets are from three publicly available datasets: *RoadScene* for task1[1], the dataset provided by (Cai, Gu, and Zhang 2018)[2] for task2, and *Lytro Multi-focus*[3] for task3. Thereinto, *RoadScene* dataset is a new infrared and visible image dataset released by ourselves to remedy shortcomings in existing datasets. The new dataset has 221 aligned Vis and IR image pairs containing rich scenes such as roads, vehicles, pedestrians and so on. These image pairs are highly representative scenes from the FLIR video[4]. We preprocess the background thermal noise in the original IR images, accurately align the Vis and IR image pairs, and cut out the exact registration regions to form this dataset. It solves the problems in existing datasets such as few image pairs, low spatial resolution and extreme lack of detailed information in infrared images.

Source images in the training datasets are cropped to patches of size 64×64. As for multi-focus images, due to the lack of aligned dataset, the source images are enlarged and flipped (either horizontally or vertically) to obtain more training data. λ is set as 12, 15 and 11, respectively. c is set as 13, 8 and 1 correspondingly. α is set as 5e-5 and β is set as 3e3. The parameters in DenseNet are updated by RMSPropOptimizer with the learning rate set as 1e-4. In the training phase, 177 Vis/IR pairs, 60 under/over-exposure pairs, and 10 far/near-focused pairs are used for training, respectively. In the testing phase, the numbers of image pairs in these 3 fusion tasks are 44, 30 and 10 correspondingly.

Results

We conduct qualitative and quantitative experiments to validate the effectiveness. For each task, we compare our results with five state-of-the-art methods respectively. These comparative methods contain traditional, deep learning-based

[1] https://github.com/hanna-xu/RoadScene

[2] http://rit-mcsl.org/fairchild//HDRPS/HDRthumbs.html

[3] https://mansournejati.ece.iut.ac.ir/content/lytro-multi-focus-dataset

[4] https://www.flir.com/oem/adas/adas-dataset-form/

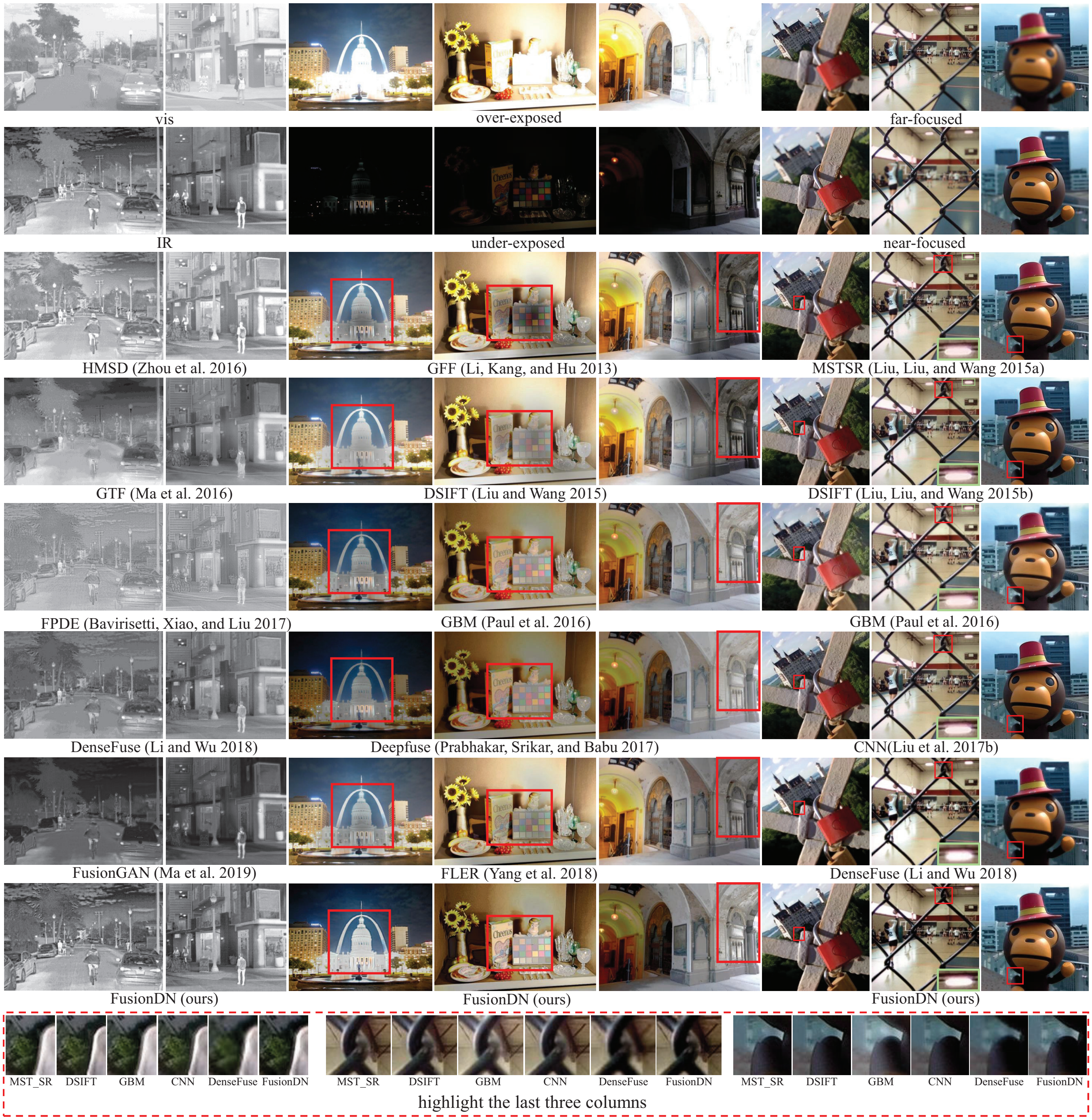


Figure 6: Qualitative comparison of FusionDN with corresponding state-of-the-art methods on infrared and visible (the first and second columns), multi-exposure (from the third to the fifth columns), and multi-focus (the last three columns) image pairs.

and general fusion methods, such as HMSD (Zhou et al. 2016), GTF (Ma et al. 2016), FPDE (Bavirisetti, Xiao, and Liu 2017), DenseFuse (Li and Wu 2018) and FusionGAN (Ma et al. 2019) for infrared and visible fusion, GFF (Li, Kang, and Hu 2013), DSIFT (Liu and Wang 2015), GBM (Paul, Sevcenco, and Agathoklis 2016), Deepfuse (Prabhakar, Srikar, and Babu 2017) and FLER (Yang et al. 2018) for multi-exposure fusion, MSTSR (Liu, Liu, and Wang 2015a), DSIFT (Liu, Liu, and Wang 2015b), GBM, CNN (Liu et al. 2017b) and DenseFuse for multi-focus fusion.

Qualitative Comparisons The qualitative results are intuitively shown in Figure 6. For the infrared and visible image

Table 2: Mean values on infrared and visible image fusion. Bold indicates the best and italics denote the second best

	EN	MG	VIF	N_{ABF}
HMSD	7.3414	*0.0304*	*1.0542*	*0.2404*
GTF	*7.4717*	0.0150	0.4664	0.0599
FPDE	6.9044	0.0249	0.4428	0.1575
DenseFuse	7.2588	0.0199	0.4530	0.2131
FusionGAN	6.9150	0.0138	0.3466	0.0778
FusionDN (ours)	**7.5832**	**0.0305**	**1.0915**	**0.2668**

Table 3: Mean values on multi-exposure image fusion.

	SD	EN	SSIM	VIF
GFF	*0.1974*	*7.3990*	1.9492	2.5763
DSIFT	0.1760	7.2753	1.9525	2.1207
GBM	0.1417	6.9354	1.9628	*3.1403*
Deepfuse	0.1467	6.8950	1.9649	1.9497
FLER	0.1296	6.8322	*1.9582*	2.4506
FusionDN (ours)	**0.1999**	**7.4488**	**1.9654**	**5.3011**

Table 4: Mean values on multi-focus image fusion.

	SD	EN	VIF	SCD
MSTSR	0.1909	7.3801	1.1465	2.5763
DSIFT	0.2044	7.4603	1.2848	2.1207
GBM	0.1875	7.4596	*1.3592*	*3.1403*
CNN	0.2040	7.4586	1.2740	1.9497
DenseFuse	*0.2195*	*7.5783*	1.3454	2.4506
FusionDN (ours)	**0.2312**	**7.6800**	**1.5949**	**5.3011**

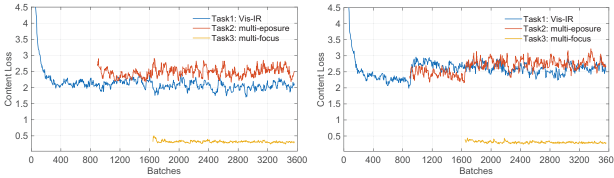


Figure 7: Changes of the content loss with (i.e., the left plot) or without (i.e., the right plot) EWC.

fusion, our results have two advantages. First, our results can preserve the high contrast in the infrared image by highlighting thermal targets or the highly illuminated regions in the visible image. The second one is that on the basis of retaining the thermal radiation information, our method can add more texture details to the background and objects to make them more similar to those in the visible image. As for multi-exposure image fusion, our results show more evidently appropriate exposure compared with others. By avoiding the dark regions in other methods, our results can exhibit more clear details, as shown in the red boxes in the third to fifth columns. In multi-focus image fusion, although we do not artificially produce a large number of clear and blurred images as the training data, nor do we directly extract focused regions and fill them in fused results, we also achieve comparable results. As shown in the last three columns, our results can preserve the sharp appearance in both source images as our method tries to reconstruct the focused regions in source images as much as possible after judging the relative blurring between them. For ease of observation, regions in red boxes are highlighted and shown in the last row. It is worth noting that since we apply a single model to perform multi-fusion tasks and have previously trained it on the multi-exposure fusion task, the model can also modify the over-exposure region in multi-focus images, as shown in green boxes in the seventh column, which are highlighted and shown in the lower right corner.

Quantitative Comparisons We also perform objective metrics to evaluate fused results, and different appropriate metrics are used for different tasks. For each task, the first two metrics are devoted to evaluating the properties of fused images, while the rest two metrics evaluate the relevance between fused images and source images. More concretely, we employ the entropy (EN) (Liu et al. 2014), mean gradient (MG) and standard deviation (SD) to evaluate the amount of information, edges and textures, and contrast in the fused image. And the visual information fidelity (VIF) (Sheikh and Bovik 2006), petrovic metric parameter (N_{ABF}) (Petrovic and Xydeas 2005), SSIM (Wang et al. 2004) and sum of the correlations of differences (SCD) (Aslantas and Bendes 2015) are used to evaluate the distortion, noise or artifacts added due to fusion, similarity, and amount of complementary information transferred between the fused image and source images. The specific results are shown in the Tables 2–4. The optimal mean values of our method on these metrics show that our results contain more information, texture details, stronger contrast, and a higher similarity with the source image with less distortion or artifacts.

Comparative Experiments In our method, we employ EWC to train a single model for all tasks without catastrophic forgetting. To validate its effectiveness, we sequentially train the three tasks without EWC. The difference is shown in the changes of the content loss in Figure 7. With EWC, when we train the next task, the content losses of previous tasks are basically the same as the losses when they were trained, as shown in the left plot of Figure 7. However, without EWC, because of the differences between specific fusion tasks, when we train the model on the next fusion task, the content losses of previous tasks increase evidently, as shown in the right plot of Figure 7, which is the representation of a decline in the performance of the single model on previous tasks. Thus, with EWC, FusionDN can obtain a single model applicable to the above three fusion tasks.

Conclusion

In this paper, a new unsupervised deep learning fusion method, called *FusionDN*, is proposed by using a unified densely connected network to generate fused images. A weight block is applied to obtain two data-driven weights as the retention degrees of features in different source images. These weights are obtained based on measuring the quality and the amount of information in source images. Moreover, we obtain a single model applicable to multiple fusion tasks that overcomes catastrophic forgetting and avoids the storage and computation issues of jointly training. This single model can generate high-quality fused results in infrared and visible, multi-exposure, and multi-focus image fusions compared with state-of-the-art methods. Also, based on the FLIR video, we release a new aligned infrared and visible image dataset, i.e., *RoadScene*, which provides a new choice for image fusion benchmark evaluation.

Acknowledgments

The authors gratefully acknowledge the financial supports from the National Natural Science Foundation of China under Grant nos. 61773295 and 61772512, and the Natural Science Fund of Hubei Province under Grant no. 2019CFA037.

References

Aslantas, V., and Bendes, E. 2015. A new image quality metric for image fusion: the sum of the correlations of differences. *Aeu-international Journal of electronics and communications* 69(12):1890–1896.

Bavirisetti, D. P.; Xiao, G.; and Liu, G. 2017. Multi-sensor image fusion based on fourth order partial differential equations. In *Proc. Int. Conf. Inf. Fusion*, 1–9.

Bosse, S.; Maniry, D.; Müller, K.-R.; Wiegand, T.; and Samek, W. 2017. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Trans. Image Process.* 27(1):206–219.

Cai, J.; Gu, S.; and Zhang, L. 2018. Learning a deep single image contrast enhancer from multi-exposure images. *IEEE Trans. Image Process.* 27(4):2049–2062.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 4700–4708.

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proc. Natl. Acad. Sci.* 114(13):3521–3526.

Li, H., and Wu, X.-J. 2018. Densefuse: A fusion approach to infrared and visible images. *IEEE Trans. Image Process.* 28(5):2614–2623.

Li, S.; Kang, X.; and Hu, J. 2013. Image fusion with guided filtering. *IEEE Trans. Image Process.* 22(7):2864–2875.

Liu, Y., and Wang, Z. 2015. Dense sift for ghost-free multi-exposure fusion. *J. Visual Commun. Image Represent.* 31:208–224.

Liu, L.; Liu, B.; Huang, H.; and Bovik, A. C. 2014. No-reference image quality assessment based on spatial and spectral entropies. *Signal Processing: Image Communication* 29(8):856–863.

Liu, Y.; Chen, X.; Cheng, J.; and Peng, H. 2017a. A medical image fusion method based on convolutional neural networks. In *Proc. Int. Conf. Inf. Fusion*, 1–7.

Liu, Y.; Chen, X.; Peng, H.; and Wang, Z. 2017b. Multi-focus image fusion with a deep convolutional neural network. *Inf. Fusion* 36:191–207.

Liu, Y.; Liu, S.; and Wang, Z. 2015a. A general framework for image fusion based on multi-scale transform and sparse representation. *Inf. Fusion* 24:147–164.

Liu, Y.; Liu, S.; and Wang, Z. 2015b. Multi-focus image fusion with dense sift. *Inf. Fusion* 23:139–155.

Ma, J.; Chen, C.; Li, C.; and Huang, J. 2016. Infrared and visible image fusion via gradient transfer and total variation minimization. *Inf. Fusion* 31:100–109.

Ma, J.; Yu, W.; Liang, P.; Li, C.; and Jiang, J. 2019. Fusiongan: A generative adversarial network for infrared and visible image fusion. *Inf. Fusion* 48:11–26.

Ma, J.; Liang, P.; Yu, W.; Chen, C.; Guo, X.; Wu, J.; and Jiang, J. 2020. Information fusion of passive sensors for detection of moving targets in dynamic environments. *Inf. Fusion* 54:85–98.

Ma, J.; Ma, Y.; and Li, C. 2019. Infrared and visible image fusion methods and applications: A survey. *Inf. Fusion* 45:153–178.

Masi, G.; Cozzolino, D.; Verdoliva, L.; and Scarpa, G. 2016. Pansharpening by convolutional neural networks. *Remote Sens.* 8(7):594.

Paramanandham, N., and Rajendiran, K. 2018. Multi sensor image fusion for surveillance applications using hybrid image fusion algorithm. *Multimed. Tools Appl.* 77(10):12405–12436.

Paul, S.; Sevcenco, I. S.; and Agathoklis, P. 2016. Multi-exposure and multi-focus image fusion in gradient domain. *J. Circuit. Syst. Comp.* 25(10):1650123.

Petrovic, V., and Xydeas, C. 2005. Objective image fusion performance characterisation. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, 1866–1871. IEEE.

Prabhakar, K. R.; Srikar, V. S.; and Babu, R. V. 2017. Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs. In *Proc. IEEE Int. Conf. Comput. Vis*, 4724–4732.

Sheikh, H. R., and Bovik, A. C. 2006. Image information and visual quality. *IEEE Transactions on image processing* 15(2):430–444.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; Simoncelli, E. P.; et al. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* 13(4):600–612.

Xu, H.; Liang, P.; Yu, W.; Jiang, J.; and Ma, J. 2019. Learning a generative model for fusing infrared and visible images via conditional generative adversarial network with dual discriminators. In *Proc. Int. Joint Conf. Artificial Intelligence*, 3954–3960.

Yang, Y.; Cao, W.; Wu, S.; and Li, Z. 2018. Multi-scale fusion of two large-exposure-ratio images. *IEEE Signal Process. Lett.* 25(12):1885–1889.

Zhang, Y.; Bai, X.; and Wang, T. 2017. Boundary finding based multi-focus image fusion through multi-scale morphological focus-measure. *Inf. Fusion* 35:81–101.

Zhang, Q.; Liu, Y.; Blum, R. S.; Han, J.; and Tao, D. 2018. Sparse representation based multi-sensor image fusion for multi-focus and multi-modality images: A review. *Inf. Fusion* 40:57–75.

Zhang, Y.; Liu, Y.; Sun, P.; Yan, H.; Zhao, X.; and Zhang, L. 2020. Ifcnn: A general image fusion framework based on convolutional neural network. *Inf. Fusion* 54:99–118.

Zhang, H.; Sindagi, V.; and Patel, V. M. 2018. Multi-scale single image dehazing using perceptual pyramid deep network. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 902–911.

Zhou, Z.; Wang, B.; Li, S.; and Dong, M. 2016. Perceptual fusion of infrared and visible images through a hybrid multi-scale decomposition with gaussian and bilateral filters. *Inf. Fusion* 30:15–26.