This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2020.3012548, IEEE Transactions on Pattern Analysis and Machine Intelligence

U2Fusion: A Unified Unsupervised Image Fusion Network

Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling

Abstract—This study proposes a novel *unified* and *unsupervised* end-to-end image *fusion* network, termed as *U2Fusion*, which is capable of solving different fusion problems, including multi-modal, multi-exposure, and multi-focus cases. Using feature extraction and information measurement, U2Fusion automatically estimates the importance of corresponding source images and comes up with adaptive information preservation degrees. Hence, different fusion tasks are unified in the same framework. Based on the adaptive degrees, a network is trained to preserve the adaptive similarity between the fusion result and source images. Therefore, the stumbling blocks in applying deep learning for image fusion, *e.g.*, the requirement of ground-truth and specifically designed metrics, are greatly mitigated. By avoiding the loss of previous fusion capabilities when training a single model for different tasks sequentially, we obtain a unified model that is applicable to multiple fusion tasks. Moreover, a new aligned infrared and visible image dataset, *RoadScene* (available at https://github.com/hanna-xu/RoadScene), is released to provide a new option for benchmark evaluation. Qualitative and quantitative experimental results on three typical image fusion tasks validate the effectiveness and universality of U2Fusion. Our code is publicly available at https://github.com/hanna-xu/U2Fusion.

Index Terms—Image fusion, unified model, unsupervised learning, continual learning.

1 Introduction

IMAGE fusion has a wide variety of applications, ranging from security to industrial and civilian fields [1], [2]. With the limitation of hardware devices or optical imaging, an image captured with one type of sensor or one single shooting setting can merely capture a part of the information. For instance, information of reflected lighting, with brightness in a limited range and within a predefined depth-of-field, is a typical representation of incomplete information. The target of image fusion is to generate a synthesized image by integrating complementary information from several source images that are captured with different sensors or optical settings. A schematic illustration of different image fusion tasks is shown in Fig. 1. A single fusion image with superior scene representation and better visual perception is suitable for subsequent visual tasks, such as video surveillance, scene understanding, and target recognition, *etc*. [3], [4].

Typically, image fusion operates on multi-modal, multi-exposure, or multi-focus images. To solve these problems, a large number of algorithms have been developed. They can be roughly divided into two categories: those based on a traditional fusion framework and those based on end-to-end models [9]. Although these algorithms have achieved promising results in their respective fusion tasks, some problems remain to be solved. In methods based on the traditional fusion framework, the finite choices of fusion rules and the complexity of manual design limit the improvement of the performance. In end-to-end models, the fusion problem is solved by relying on ground truth for supervised learning or the specifically designed metrics for unsupervised learning. *However, universal ground truth or no-reference metric for multiple tasks does not exist*. These issues form the major stumbling blocks in the unity of models and the application of supervised or unsupervised learning.

Meanwhile, *different fusion tasks often share the similar goal, that is, to synthesize an image by integrating vital and complementary information from several source images.* Nevertheless, in different tasks, the vital information to be integrated varies largely as source images are of different types (see detailed explanation in Sec. 3.1), thus limiting the effectiveness of most methods to specific tasks. With the strong ability of feature representation in neural networks, the varied information can be represented in a unified way. It potentially leads to a unified fusion framework, which will be explored in this study.

Moreover, by solving different fusion problems in a unified model, *these tasks can promote one another*. For instance, given that the unified model has been trained for multi-exposure image fusion, it is capable of improving the fusion performance of under/overexposed regions in the multi-modal or multi-focus images. *Thus, by gathering the strengths of multiple tasks, the unified model can achieve better results for each single fusion task with stronger generalization than multiple individually trained models.*

To address these issues, we propose a *unified unsupervised* image *fusion* network known as *U2Fusion*. For information preservation, a feature extractor is first adopted to extract abundant and comprehensive features from source images. Then, the richness of information in features is measured to define the relative importance of these features, which indi-

- *H. Xu and J. Ma are with the Electronic Information School, Wuhan University, Wuhan, 430072, China (email: xu_han@whu.edu.cn, jyma2010@gmail.com).*
- *J. Jiang is with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, 150001, China (email: jiangjunjun@hit.edu.cn).*
- *X. Guo is with the College of Intelligence and Computing, Tianjin University, Tianjin, 300350, China (email: xj.max.guo@gmail.com).*
- *H. Ling is with the Department of Computer Science, Stony Brook University, NY, 11794, USA. (email: haibin.ling@gmail.com).*

Manuscript received Apr. 12, 2020; revised Jul. 13, 2020; accepted Jul. 25, 2020. (Corresponding author: Jiayi Ma.)

0162-8828 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Carleton University. Downloaded on August 03,2020 at 08:47:48 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2020.3012548, IEEE Transactions on Pattern Analysis and Machine Intelligence

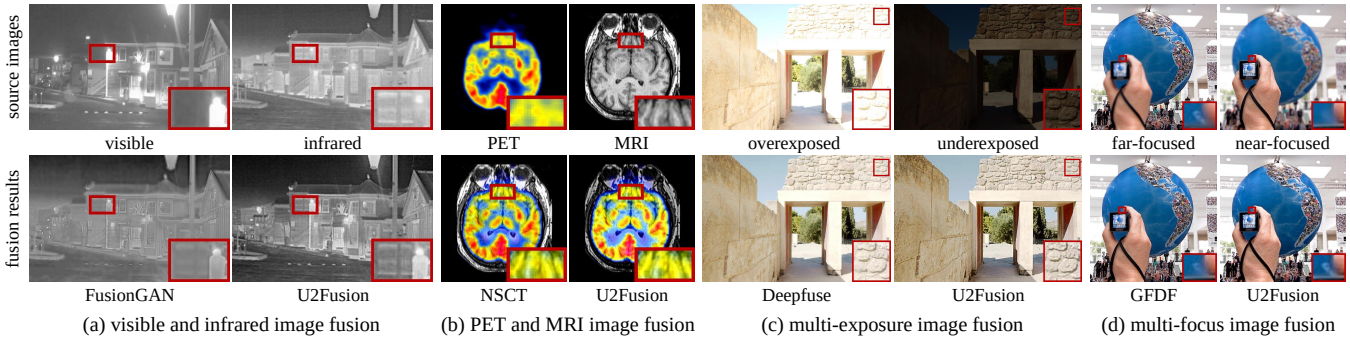


Fig. 1. Schematic illustration of different image fusion tasks (first row: source images, second row (from left to right): fusion results of FusionGAN [5], U2Fusion, NSCT [6], U2Fusion, Deepfuse [7], U2Fusion, GFDF [8] and U2Fusion).

cates the similarity relationship between the source images and the fusion result. A higher similarity entails that more information in this source image is preserved in the result, thus leading to a higher information preservation degree. On the basis of these strategies, a DenseNet [10] module is trained to generate the fusion result without the need for ground truth. The characteristics and contributions of our work are summarized as follows:

- We propose a unified framework for various image fusion tasks. More concretely, we solve different fusion problems with a unified model and unified parameters. Our solution alleviates shortcomings, such as separate solutions for different problems, storage and computation issues for training, and catastrophic forgetting for continual learning.
- We develop a new unsupervised network for image fusion by constraining the similarity between the fusion image and source images to overcome the universal stumbling blocks in most image fusion problems, *i.e.*, the lack of universal ground truth and no-reference metric.
- We release a new aligned infrared and visible image dataset, *RoadScene*, to provide a new option for image fusion benchmark evaluation. It is made available at https://github.com/hanna-xu/RoadScene.
- We test the proposed method on six datasets for multi-modal, multi-exposure, and multi-focus image fusions. Qualitative and quantitative results validate the effectiveness and universality of U2Fusion.

A preliminary version of this paper appears in [11]. The new contributions are mainly from four aspects. First, the strategy for information preservation degree assignment is improved. Instead of the amount and quality of information in original source images, the information preservation degrees are assigned by the information measurement performed on extracted features. By considering additional aspects, the modified strategy provides an improved comprehensive measurement to capture the essential characteristics of source images. Second, the loss function is modified. The removal of the gradient loss alleviates the false edges, and the added pixel intensity-based loss helps reduce the luminance deviation in the fusion image. Third, we replace the first task from visible (VIS) and infrared (IR) image fusion to multi-modal image fusion where VIS-IR and medical image fusion are included. Lastly, we validate U2Fusion on additional publicly available datasets. For the ablation study, to validate the effectiveness of elastic weight consolidation (EWC) for continual learning from new tasks [12], we analyze the EWC from two additional aspects, namely, the statistical distributions of the weight for EWC and the intermediate results of all the tasks during the training phase. As for the adaptive information preservation degrees, the validation of their effectiveness is also performed.

2 Related Work

2.1 Image Fusion Methods

2.1.1 Methods Based on Traditional Fusion Framework

The traditional fusion framework can be roughly summarized as Fig. 2. As reconstruction is usually an inverse process of extraction, the key to these algorithms lies in two important factors: feature extraction and feature fusion. By modifying them, these methods can be designed for solving multi-modal, multi-exposure, or multi-focus image fusion.

To solve the issue of feature extraction, a large number of traditional methods have been proposed. The theories on which they are based can be divided into four representative categories: i) multi-scale transform, such as Laplacian pyramid (LP), ratio of low-pass pyramid (RP), gradient pyramid (GP), discrete wavelet (DWT), discrete cosine (DCT) [13], curvelet transform (CVT), shearlet, *etc.*; ii) sparse representation [14]; iii) subspace analysis, *e.g.*, independent component analysis (ICA), principal component analysis (PCA), non-negative matrix factorization (NMF), *etc.*; and iv) hybrid methods. However, these manually designed extraction approaches make fusion methods increasingly complex, thus intensifying the difficulty of designing fusion rules. The extraction methods need to be modified correspondingly to solve different fusion tasks. Furthermore, much attention needs to be given to the appropriateness of extraction methods to ensure the completeness of features. To overcome these limitations, some methods introduce convolutional neural networks (CNN) in feature extraction, either as some subparts [15], [16] or as the entire part [17], [18].

Then, the fusion rules are determined on the basis of extracted features. The commonly used rules include maximum, minimum, addition, l_1-norm, *etc*. However, the limit choices of these manually designed fusion rules produce a glass ceiling on the performance improvement even in some CNN-based methods.

0162-8828 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Carleton University. Downloaded on August 03,2020 at 08:47:48 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2020.3012548, IEEE Transactions on Pattern Analysis and Machine Intelligence

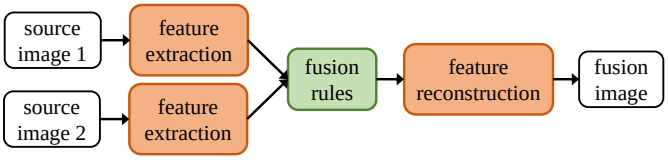


Fig. 2. Traditional image fusion framework.

Notably, there are some methods breaking away from the framework, such as the method based on gradient transfer and total variation minimization for VIS-IR image fusion [19], the multi-exposure image fusion method by optimizing a structural similarity index [20], and the method based on dense SIFT for multi-focus image fusion [21], *etc*. However, the algorithms or metrics on which these methods are based are dedicated to specific fusion tasks and may not generalize well.

2.1.2 End-to-end Models

To avoid designing fusion rules, many deep learning-based algorithms have been put forward. Unlike the methods in Sec. 2.1.1, these methods are usually end-to-end models tailored to specific fusion tasks.

Multi-modal Image Fusion. The end-to-end models for multi-modal image fusion are typically designed for VIS and IR image fusion. Ma *et al*. proposed FusionGAN [5] by establishing an adversarial game between a generator and a discriminator to preserve the pixel intensity distribution in the IR image and details in the VIS image. Later, its variant [22] was proposed to sharpen the edges of thermal targets by introducing the target-enhancement loss. DDcGAN [23], [24] enhances the prominence of thermal targets by introducing the dual-discriminator architecture. However, the unique issue in VIS and IR image fusion is the preservation of the pixel intensity distribution and details, which does not apply to other fusion tasks. In addition, ground truth is usually not present in this type of task. Thus, it is the major obstacle in utilizing supervised learning in multi-modal image fusion.

Multi-exposure Image Fusion. To solve this problem, some unsupervised methods have been put forward. Prabhakar *et al*. proposed Deepfuse [7], where the no-reference metric MEF-SSIM is adopted as the loss function. However, MEF-SSIM is especially designed for multi-exposure images by discarding the luminance component, as it is not significant in this problem. Nevertheless, it still plays an important role in other tasks. Thus, MEF-SSIM is not applicable to other problems. In some multi-exposure datasets, there are no ground truths for supervised learning.

Multi-focus Image Fusion. For this problem, Liu *et al*. put forward a network to generate the focus map [25]. The predefined labels, which indicate whether they are high-quality images or Gaussian blurred images, are used for supervised learning. Then, it was extended to a general image fusion framework [26]. Depending on the generalization, the model trained on multi-focus image fusion can be employed to solve other tasks. In addition, Guo *et al*. proposed FuseGAN [27] where the generator directly produces a binary focus mask and the discriminator attempts to distinguish the generated masks from the ground truths, which are synthesized by utilizing a normalized disk point spread function and separating the background and foreground. The focus maps/masks are significant for multi-focus image fusion, whereas they are not necessary or even not applicable in other tasks. All these methods are based on supervised learning.

Our method. By considering the abovementioned limitations, we propose a unified unsupervised image fusion network, which has the following characteristics. i) It is an end-to-end model not restricted by the limit of manually designed fusion rules. ii) It is a unified model for various fusion tasks instead of specific objectives, *e.g.*, distinctive issues, the specificity of metrics, the need of binary masks, *etc*. iii) It is an unsupervised model without the need of ground truth. iv) By continuously learning to solve new tasks without losing old capabilities, it solves multiple tasks with unified parameters.

2.2 Continual Learning

In a continual learning setting, the learning is considered as a sequence of tasks to be learned. During the training phase, the weights are adapted to new tasks without forgetting the previously learned ones. To avoid storing any training data from previously learned tasks, many algorithms based on *elastic weight consolidation* (EWC) are proposed [28], [29], which include a regularization term to force parameters to remain close to those trained for the previous tasks. These technologies have been widely applied in many practical problems, such as person reidentification [30], real-time vehicle detection [31], and emotion recognition [32], *etc*. In this study, we perform continual learning for solving multiple fusion tasks.

3 Methodology

Our system allows signals captured with different sensors and/or shooting settings from the same camera position. In this section, we provide the problem formulation, the design of loss functions, the technology of elastic weight consolidation, and the network architecture.

3.1 Problem Formulation

Focusing on the primary goal of image fusion, *i.e.*, to preserve the vital information in source images, our model is based on the measurement to determine the richness of such information. If the source image contains abundant information, it is of great importance to the fusion result, which should show a high similarity with the source image. Therefore, the key issue of our method is to explore a unified measurement to determine the information preservation degrees of source images. Rather than maximizing the similarity between the fusion result and the ground truth in supervised learning, our method depends on such degrees to preserve the adaptive similarity with source images. And, as an unsupervised model, it is applicable to multiple fusion problems where ground truth is hardly available.

For the desired measurement, a major problem is that the vital information in different types of source images varies greatly. For example, in IR and positron emission

0162-8828 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Carleton University. Downloaded on August 03,2020 at 08:47:48 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2020.3012548, IEEE Transactions on Pattern Analysis and Machine Intelligence

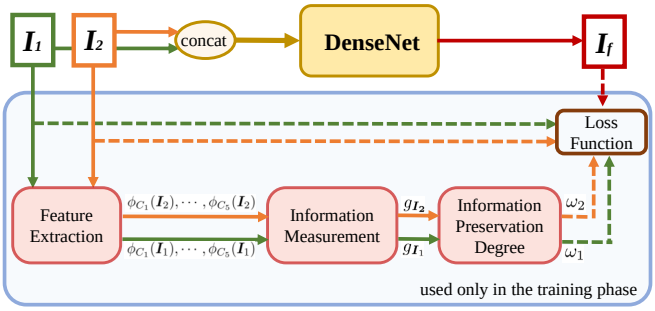


Fig. 3. Pipeline of the proposed U2Fusion. Dashed lines represent the data used in the loss function.

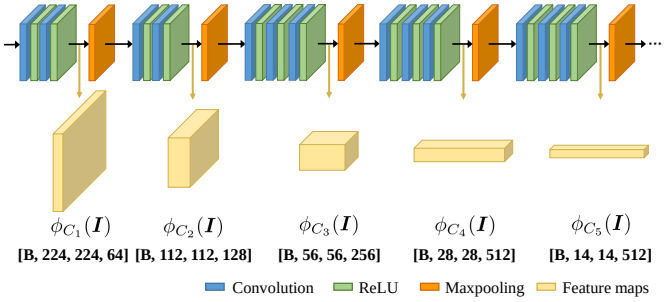


Fig. 4. Perceptual feature maps extracted by VGG-16 for input image $\boldsymbol{I}$, and $\phi_{C_j}(\boldsymbol{I})$ represents the feature map extracted by the convolutional layer before the j-th max-pooling layer. The last row is the shape of extracted feature maps in the form of [batchsize, height, width, channel].

tomography (PET) images, the vital information is the thermal radiation and functional responses that are presented as the pixel intensity distribution. In VIS and magnetic resonance imaging (MRI) images, the vital information is the reflected light and structural content represented by image gradients [19], [23]. In multi-focus images, the information to be preserved includes the objects within the depth-of-field (DoF). In multi-exposure images, the vital information concerns scene content can be enhanced. The above variability brings considerable difficulty to designing a unified information measurement, which are designed for the specific tasks cease to be effective when facing other problems. They are based on certain surface-level characteristics or specific properties while in different tasks, and are difficult to be predetermined in a unified way. We solve this problem by taking a comprehensive consideration of multifaceted properties of source images. To this end, we extract both shallow-level features (textures, local shapes, *etc*.) and deep-level features (content, spatial structures, *etc*.) for estimating the information measurement.

The pipeline of U2Fusion is summarized as Fig. 3. With source images denoted as $\boldsymbol{I}_1$ and $\boldsymbol{I}_2$, a DenseNet is trained to generate the fusion image $\boldsymbol{I}_f$. The outputs of feature extraction are the feature maps $\phi_{C_1}(\boldsymbol{I}_1),\cdots,\phi_{C_5}(\boldsymbol{I}_1)$ and $\phi_{C_1}(\boldsymbol{I}_2),\cdots,\phi_{C_5}(\boldsymbol{I}_2)$. Then the information measurement is performed on these feature maps, producing two measurements denoted by $g_{\boldsymbol{I}_1}$ and $g_{\boldsymbol{I}_2}$. With subsequent processing, the final information preservation degrees are denoted as ω_1 and ω_2. $\boldsymbol{I}_1, \boldsymbol{I}_2, \boldsymbol{I}_f, \omega_1$ and ω_2 are used in the loss function without the need for ground truth. In the training phase, ω_1 and ω_2 are measured and applied in defining the loss function. Then, a DenseNet module is optimized to minimize the loss function. In the testing phase, ω_1 and ω_2 do not need to be measured, as the DenseNet has been optimized. The detailed definitions or descriptions are given in the following subsections.

3.1.1 Feature Extraction

Compared with models trained in fusion tasks, models for other computer vision tasks are usually trained with larger and more diversified datasets. Thus, features extracted by such models are abundant and comprehensive [33], [34]. Inspired by the perceptual loss [35], [36], we adopt the pretrained VGG-16 network [37] for feature extraction, as shown in Fig. 4. The input $\boldsymbol{I}$ has been unified in a single channel in our model (we will discuss this transformation in Sec. 3.5), and we duplicate it into three channels and then feeding them into VGG-16. The outputs of the convolutional layers before max-pooling layers are feature maps for the subsequent information measurement, which are shown in Fig. 4 as $\phi_{C_1}(\boldsymbol{I}),\cdots,\phi_{C_5}(\boldsymbol{I})$ with their shapes shown below.

For intuitive analysis, some feature maps of a multi-exposure image pair are shown in Fig. 5. In the original source images, the overexposed image contains much more texture details or larger gradients than the underexposed image, as the latter suffers from much lower luminance. In Fig. 5, features in $\phi_{C_1}(\boldsymbol{I})$ and $\phi_{C_2}(\boldsymbol{I})$ are based on shallow features, such as textures and shape details. In these layers, feature maps of the overexposed image still shows more information than the underexposed one. By comparison, feature maps of higher layers, *e.g.*, $\phi_{C_4}(\boldsymbol{I})$ and $\phi_{C_5}(\boldsymbol{I})$, mainly preserve deep-level features, such as the content or spatial structures. In these layers, comparable and additional information are present in the feature maps of the underexposed image. Therefore, the combination of shallow- and deep-level features forms a comprehensive representation of the essential information that may not be easily perceived by the human visual perception system.

3.1.2 Information Measurement

To measure the information contained in the extracted feature maps, their gradients are used for evaluation. Compared with entities derived from general information theory, image gradient is a metric based on local spatial structures with small receptive fields. When used in the deep learning framework, gradients are much more efficient in both computation and storage. Thus, they are more suitable for application in CNN for information measurement. The information measurement is defined as follows:

$$g_{\boldsymbol{I}} = \frac{1}{5}\sum_{j=1}^{5}\frac{1}{H_j W_j D_j}\sum_{k=1}^{D_j}\left\|\nabla\phi_{C_j^k}(\boldsymbol{I})\right\|_F^2, \tag{1}$$

where $\phi_{C_j}(\boldsymbol{I})$ is the feature map by the convolutional layer before the j-th max-pooling layer in Fig. 4. k denotes the feature map in the k-th channel of D_j channels. $\|\cdot\|_F$ denotes the Frobenius norm, and ∇ is the Laplacian operator.

3.1.3 Information Preservation Degree

To preserve the information in source images, two adaptive weights are assigned as the information preservation degrees, which define the weights of similarities between the fusion image and the source images. The higher the weight,

0162-8828 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Carleton University. Downloaded on August 03,2020 at 08:47:48 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2020.3012548, IEEE Transactions on Pattern Analysis and Machine Intelligence

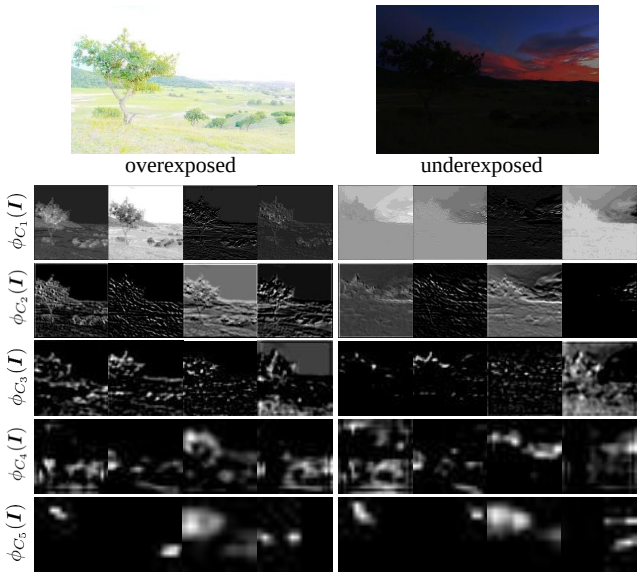


Fig. 5. Illustration of feature maps extracted by VGGNet for overexposed and underexposed images.

the higher the similarity is expected to be, and the higher the information preservation degree of the corresponding source image is.

These adaptive weights, denoted as ω_1 and ω_2, are estimated according to the information measurement results $g_{\boldsymbol{I}_1}$ and $g_{\boldsymbol{I}_2}$ obtained by Eq. (1). Given that the difference between $g_{\boldsymbol{I}_1}$ and $g_{\boldsymbol{I}_2}$ is the absolute value instead of the relative one, it may be too small compared with themselves to reflect their difference. Thus, to enhance and embody the difference in weights, a predefined positive constant c is used to scale values for better weight assignments. Thus, ω_1 and ω_2 are defined as:

$$[\omega_1, \omega_2] = \text{softmax}\left(\left[\frac{g_{\boldsymbol{I}_1}}{c}, \frac{g_{\boldsymbol{I}_2}}{c}\right]\right), \tag{2}$$

where we use the softmax function to map $\frac{g_{\boldsymbol{I}_1}}{c}, \frac{g_{\boldsymbol{I}_2}}{c}$ to real numbers between 0 and 1, and guarantee that the sum of ω_1 and ω_2 is 1. Then, ω_1 and ω_2 are employed in the loss function to control the information preservation degrees of specific source images.

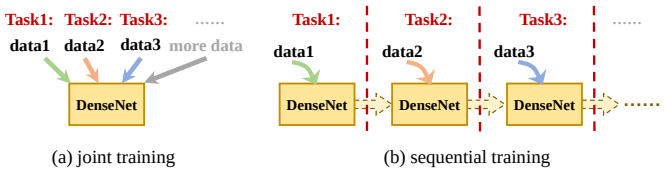
3.2 Loss Function

The loss function is mainly designed for preserving vital information and for training a single model, which is applicable for multiple tasks. It consists of two parts defined as follows:

$$\mathcal{L}(\theta, D) = \mathcal{L}_{\text{sim}}(\theta, D) + \lambda \mathcal{L}_{\text{ewc}}(\theta, D), \tag{3}$$

where θ denotes the parameters in DenseNet, and D is the training dataset. $\mathcal{L}_{\text{sim}}(\theta, D)$ is the similarity loss between the result and source images. $\mathcal{L}_{\text{ewc}}(\theta, D)$ is the item designed for continual learning, as described in next subsection. λ is a hyperparameter to control the trade-off.

We realize the similarity constraint from two aspects, *i.e.*, the structure similarity and the intensity distribution. Given that the structural similarity index measure (SSIM) is the most widely used metric that models the distortion according to similarities in the information on light, contrast, and



(a) joint training (b) sequential training

Fig. 6. Illustration of joint training and sequential training. The dashed arrow between DenseNets means that it is kept and set as the initial parameters of the next task. On this basis, these parameters are optimized according to the new objective.

structure [38], we use it to constrain the structural similarity between $\boldsymbol{I}_1$, $\boldsymbol{I}_2$, and $\boldsymbol{I}_f$. Thus, with ω_1 and ω_2 to control the information degree, the first item of $\mathcal{L}_{\text{sim}}(\theta, D)$ is formulated as:

$$\mathcal{L}_{\text{ssim}}(\theta, D) = \mathbb{E}[\omega_1 \cdot (1 - \text{S}_{\boldsymbol{I}_f, \boldsymbol{I}_1}) + \omega_2 \cdot (1 - \text{S}_{\boldsymbol{I}_f, \boldsymbol{I}_2})], \tag{4}$$

where $\text{S}_{\boldsymbol{x}, \boldsymbol{y}}$ denotes the SSIM value between two images.

While SSIM focuses on the changes of contrast and structure, it shows weaker constraints on the difference of the intensity distribution. We supplement $\mathcal{L}_{\text{ssim}}(\theta, D)$ with the second item, which is defined by the mean square error (MSE) between two images:

$$\mathcal{L}_{\text{mse}}(\theta, D) = \mathbb{E}[\omega_1 \cdot \text{MSE}_{\boldsymbol{I}_f, \boldsymbol{I}_1} + \omega_2 \cdot \text{MSE}_{\boldsymbol{I}_f, \boldsymbol{I}_2}]. \tag{5}$$

Meanwhile, the results obtained by constraining MSE suffer from relatively blurred appearance by averaging all plausible outputs, whereas SSIM can make up for this issue. Thus, these two items compensate for each other. With α controlling the trade-off, $\mathcal{L}_{\text{sim}}(\theta, D)$ is formulated as:

$$\mathcal{L}_{\text{sim}}(\theta, D) = \mathcal{L}_{\text{ssim}}(\theta, D) + \alpha \mathcal{L}_{\text{mse}}(\theta, D). \tag{6}$$

3.3 Single Model for Multi-fusion Tasks with Elastic Weight Consolidation (EWC)

Various fusion tasks usually lead to differences in feature extraction and/or fusion, as directly reflected in diverse values of DenseNet parameters. It leads to training multiple models with the same architecture but diverse parameters. However, as some parameters are redundant, the utilization of these models can be greatly improved. It motivates us to train a single model with unified parameters that integrates these models and thus become applicable for multiple tasks.

This purpose can be achieved in two ways, *i.e.*, joint training and sequential training, as shown in Fig. 6. Joint training is a simple method where all the training data are kept throughout the training process. In each batch, data from multiple tasks are randomly selected for training. Nevertheless, as the number of tasks increases, two urgent issues become difficult to solve: i) the storage issue caused by always keeping the data of previous tasks and ii) the computation issue caused by using all the data for training, in terms of both the difficulty of computation and time cost.

In sequential training, we need to change the training data for different tasks, as shown in Fig. 6(b). Thus, only the data of the current task needs to be stored in the training process, which solves storage and computation issues. However, a new problem arises when we train the model on another task for a new capability: the previous training data are unavailable [39]. As the training process

0162-8828 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Carleton University. Downloaded on August 03,2020 at 08:47:48 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2020.3012548, IEEE Transactions on Pattern Analysis and Machine Intelligence

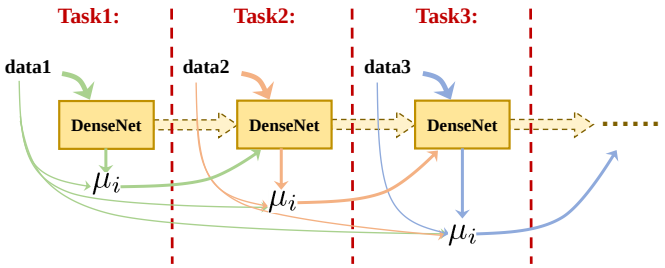


Fig. 7. Intuitive description of data flow during the process of EWC. Thin lines indicate that only a small subset of data are kept, which are merely used to calculate μ_i and not applied to train DenseNet.

continues, the parameters are optimized to solve the new problems while losing the capacity learned from previous tasks. This problem is called catastrophic forgetting. To avoid this drawback, we apply the *elastic weight consolidation* (EWC) algorithm [12] to safeguard against it.

In EWC, the squared distance between the parameter values of the current task θ and those of the previous task θ^* are weighted according to their importance to θ^*. Those important parameters are given higher weights to prevent forgetting what has been learned from old tasks, while the parameters with less importance can be modified to a greater extent to learn from the new task. In this way, the model is capable of continual learning with elastic weight consolidation. Thus, the loss for continual learning, termed as $\mathcal{L}_{\text{ewc}}(\theta, D)$, is included in the total loss function in Eq. (3). With these importance-related weights defined as μ_i, $\mathcal{L}_{\text{ewc}}(\theta, D)$ is formulated as:

$$\mathcal{L}_{\text{ewc}}(\theta, D) = \frac{1}{2}\sum_i \mu_i(\theta_i - \theta_i^*)^2, \tag{7}$$

where i represents the i-th parameter in the network and μ_i represents the weight of corresponding squared distance.

To evaluate the importance, μ_i is assigned as the diagonal terms of the Fisher information matrix and approximated by computing the square of gradients with the data in previous tasks as defined below:

$$\mu_i = \mathbb{E}\left[(\frac{\partial}{\partial\theta_i^*}\log p(D^*|\theta^*))^2|\theta^*\right], \tag{8}$$

where D^* represents the data of previous tasks. $\log p(D^*|\theta^*)$ can be approximately replaced by $-\mathcal{L}(\theta^*, D^*)$ [12]. Thus, Eq. (8) is converted to:

$$\mu_i = \mathbb{E}\left[(-\frac{\partial}{\partial\theta_i^*}\mathcal{L}(\theta^*, D^*))^2|\theta^*\right]. \tag{9}$$

Given that the Fisher information matrix can be computed before throwing away the old data D^*, the model does not require D^* for training the current task.

If several previous tasks exist, $\mathcal{L}_{\text{ewc}}(\theta, D)$ is adapted according to specific tasks and corresponding data. Then, the squares of these gradients are averaged for the final μ_i. The training process and the data flow are illustrated in Fig. 7.

In multi-task image fusion, θ is the parameters of the DenseNet. First, the DenseNet is trained to solve Task1, *i.e.*, the multi-modal image fusion problem by minimizing the similarity loss defined in Eq. (6). When adding the capacity of solving Task2, *i.e.*, the multi-exposure image fusion problem, the importance-related weights μ_i are first computed. In particular, μ_i indicates the importance of each parameter in the DenseNet to multi-modal image fusion. Then, the important parameters are consolidated to avoid catastrophic forgetting by minimizing the item $\mathcal{L}_{\text{ewc}}$ in Eq. (3); while the parameters of little significance are updated to solve the multi-exposure image fusion by minimizing the similarity loss $\mathcal{L}_{\text{sim}}$ correspondingly. Lastly, when we train the DenseNet on multi-focus image fusion, μ_i is computed according to the previous two tasks. The subsequent elastic weight consolidation strategy is the same as before. In this way, EWC can be customized to the scenario of multi-task adaptive image fusion.

3.4 Network Architecture

In our method, DenseNet is employed to generate the fusion result $\boldsymbol{I}_f$, of which the input is the concatenation of $\boldsymbol{I}_1$ and $\boldsymbol{I}_2$. Thus, it is an end-to-end model without the need for designing fusion rules. As shown in Fig. 8, the architecture of DenseNet in U2Fusion consists of 10 layers, each with a convolution followed by an activation function. The kernel size of all convolutional layers is set to 3×3 and the stride to 1. Reflection padding is employed before the convolution to reduce boundary artifacts. No pooling layer is used to avoid information loss. The activation functions in the first nine layers are LeakyReLU with the slope set to 0.2, while that of the last layer is tanh.

Moreover, research has proven that CNNs can be significantly deeper and trained efficiently if shorter connections are built between layers close to the input and those close to the output. Therefore, in the first seven layers, the densely connected blocks from densely connected CNNs [10] are employed to improve the information flow and performance. In these layers, shortcut direct connections are built between each layer and all layers in a feed-forward fashion, as shown in the concatenation operation in Fig. 8. This way, the problem of vanishing gradients can be reduced. Meanwhile, feature propagation can be further strengthened while reducing the number of parameters [40]. The channels of feature maps are all set to 44. The subsequent four layers reduce the channels of feature maps gradually until reaching a single-channel fusion result, as shown in Fig. 8.

3.5 Dealing with RGB Input

RGB inputs are first converted into the YCbCr color space. Then, the Y (luminance) channel is used for fusion, as structural details are mainly in this channel and the brightness variation in this channel is more prominent than chrominance channels. Data in the Cb and Cr (chrominance) channels are fused traditionally as:

$$\mathbf{C}_f = \frac{\mathbf{C}_1(|\mathbf{C}_1 - \tau|) + \mathbf{C}_2(|\mathbf{C}_2 - \tau|)}{|\mathbf{C}_1 - \tau| + |\mathbf{C}_2 - \tau|}, \tag{10}$$

where $\mathbf{C}_1$ and $\mathbf{C}_2$ are the Cb/Cr channel values of the first and second source image, respectively. $\mathbf{C}_f$ is the corresponding channel of the fusion result. τ is set as 128. Then, through the inverse conversion, the fusion images can be converted into the RGB space. Thus, all the problems are unified into the single-channel image fusion problem.

0162-8828 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Carleton University. Downloaded on August 03,2020 at 08:47:48 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2020.3012548, IEEE Transactions on Pattern Analysis and Machine Intelligence

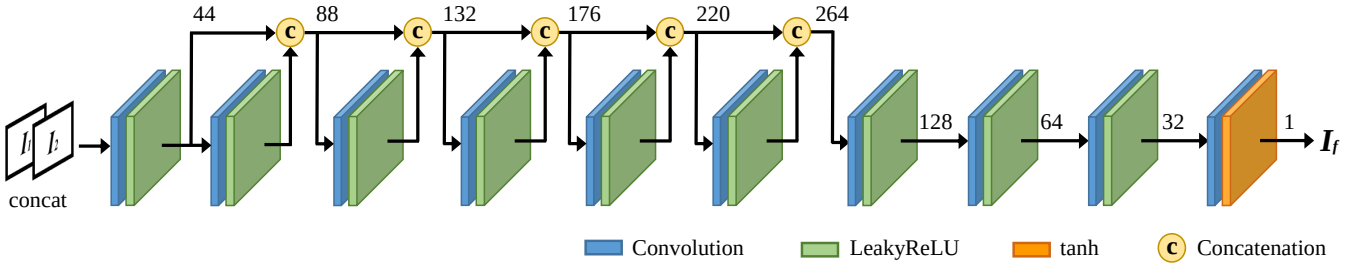


Fig. 8. Architecture of DenseNet used in our model. Numbers shown after concatenation/LeakyReLU/tanh functions are the channels of corresponding feature maps.

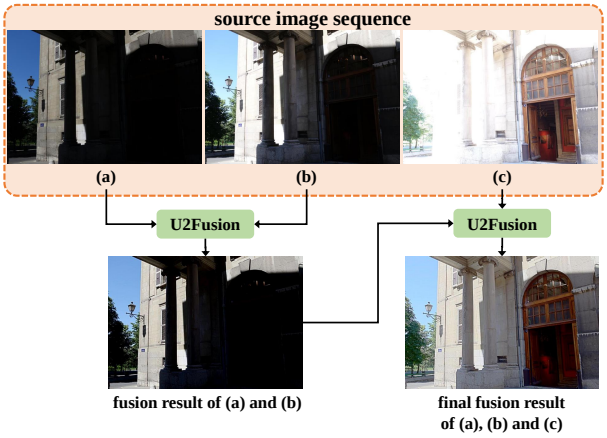


Fig. 9. U2Fusion to fuse multi-exposure image sequence.

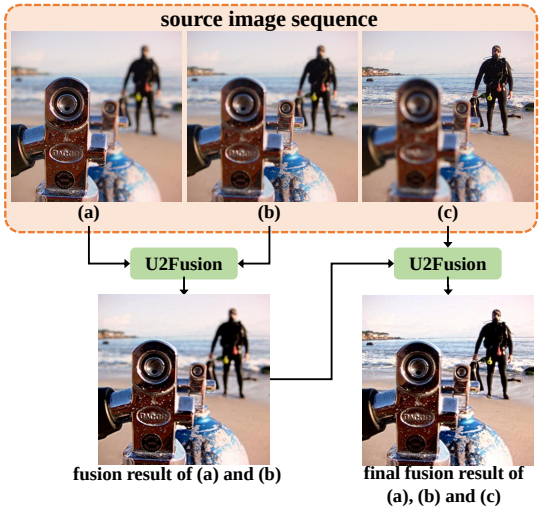


Fig. 10. U2Fusion to fuse multi-focus image sequence.

3.6 Dealing with Multiple Inputs

In multi-exposure and multi-focus fusion, we need to fuse a source image sequence, *i.e.*, more than two source images are available. In this case, these source images can be fused sequentially. As shown in Figs. 9 and 10, we initially fuse two of these source images. Then, the intermediate result is fused with another source image. In this way, U2Fusion is capable of fusing any number of inputs in theory.

4 Experimental Results and Discussions

In this section, we compare U2Fusion with several state-of-the-art methods on multiple tasks with multiple datasets by qualitative and quantitative comparisons.

4.1 Training Details

We perform U2Fusion on three types of fusion tasks: i) multi-modal image fusion, including VIS-IR and medical image (PET-MRI) fusion; ii) multi-exposure image fusion; and iii) multi-focus image fusion. Given that VIS-IR and PET-MRI fusion are similar in nature (see detailed explanation in Sec. 3.1), they are jointly seen as multi-modal image fusion (task 1). The training datasets are from four publicly available datasets: *RoadScene*[1] (VIS-IR) and *Harvard*[2] (PET-MRI) for task 1, the dataset in [41][3] for task 2, and *Lytro*[4] for task 3. To validate the universality, the test datasets also contain two additional ones: *TNO*[5] for VIS-IR image fusion and *EMPA HDR*[6] for multi-exposure image fusion.

On the basis of the FLIR video[7], we have released the *RoadScene*, which is a new aligned VIS-IR image dataset used to remedy shortcomings in existing ones. First, we select image pairs with highly repetitive scenes from the video. Second, the thermal noise in original IR images is reduced. Third, to align the image pairs accurately, we select feature points carefully and align each image pair with homography and bi-cubic interpolation. Moreover, given that some regions cannot be exactly aligned with homography because of camera distortion or imaging time elapse, we cut out the exact registration regions. *RoadScene* has 221 aligned image pairs containing rich scenes, such as roads, vehicles, and pedestrians. It solves the problems in benchmark datasets, such as few image pairs, low spatial resolution, and lack of detailed information in IR images.

Source images in all the datasets are cropped to patches of size 64×64. For multi-focus images, images are enlarged and flipped for additional training data because of the insufficient aligned image pairs. We set $\alpha = 20$ and $\lambda = 8e4$. c is set as 3e3, 3.5e3, and 1e2, and the epoches are set as 3, 2, and 2 correspondingly. The parameters are updated by RMSPropOptimizer with a learning rate 1e-4. The batch size is 18. Experiments are performed on a NVIDIA Geforce GTX Titan X GPU and 3.4 GHz Intel Core i5-7500 CPU.

1. https://github.com/hanna-xu/RoadScene
2. http://www.med.harvard.edu/AANLIB/home.html
3. https://github.com/csjcai/SICE
4. https://mansournejati.ece.iut.ac.ir/content/lytro-multi-focus-dataset
5. https://figshare.com/articles/TNOImageFusionDataset/1008029
6. http://www.empamedia.ethz.ch/hdrdatabase/index.php
7. https://www.flir.com/oem/adas/adas-dataset-form/

0162-8828 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Carleton University. Downloaded on August 03,2020 at 08:47:48 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2020.3012548, IEEE Transactions on Pattern Analysis and Machine Intelligence

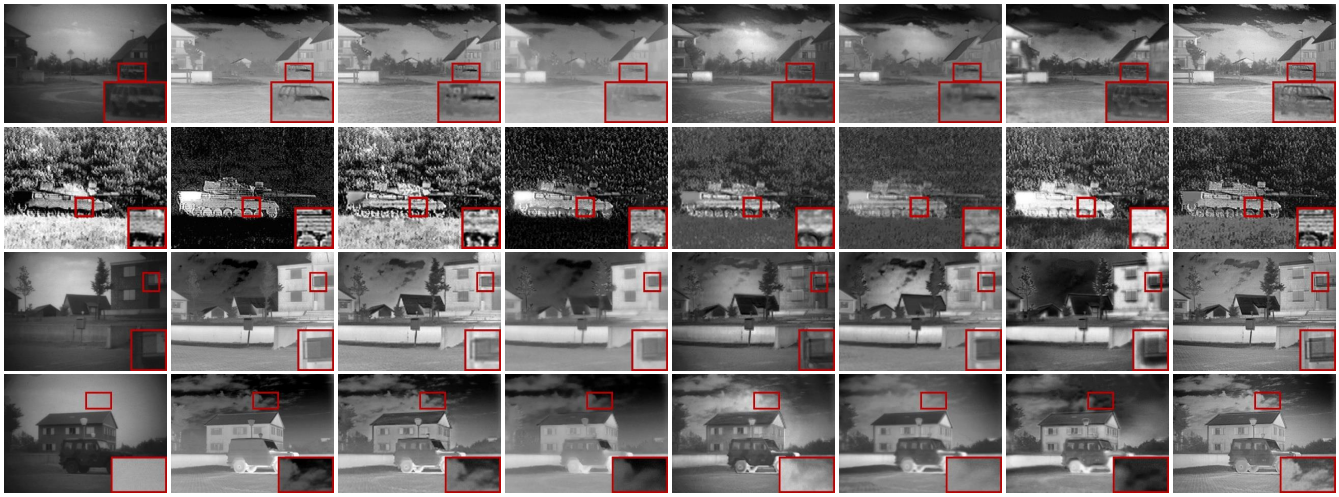


Fig. 11. Qualitative comparison of our U2Fusion with 5 state-of-the-art methods on 4 typical VIS and IR image pairs in the *TNO* dataset.

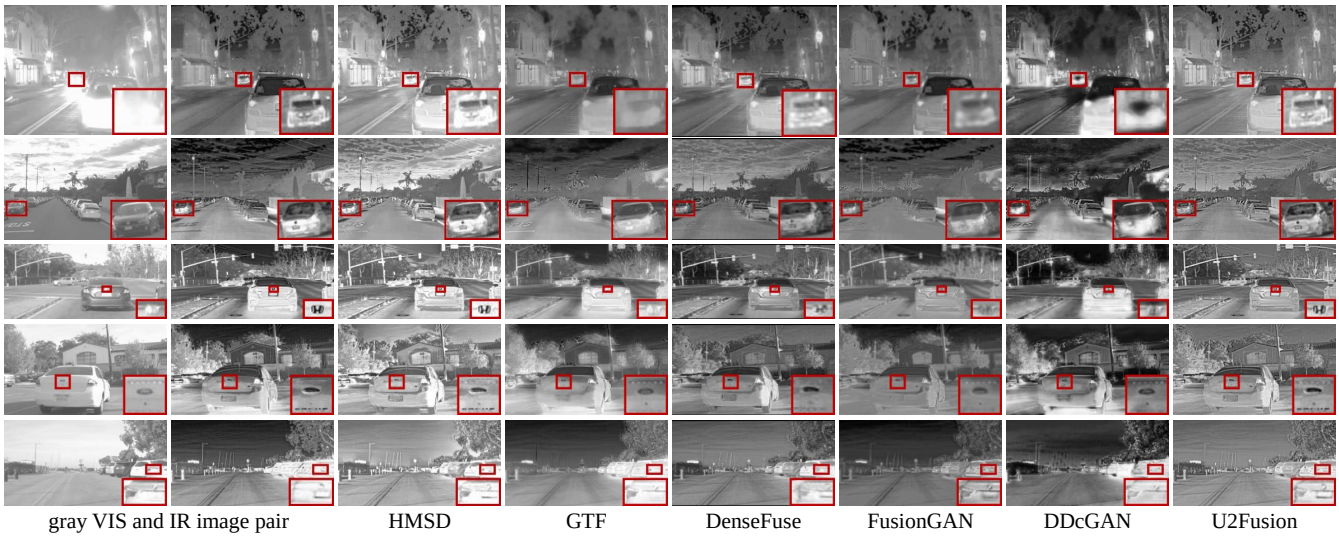


Fig. 12. Qualitative comparison of U2Fusion with 5 state-of-the-art methods on 5 typical VIS and IR image pairs in the *RoadScene* dataset.

4.2 Multi-modal Image Fusion

4.2.1 Visible and Infrared Image Fusion

We compare U2Fusion with five state-of-the-art methods: HMSD [42], GTF [19], DenseFuse [17], FusionGAN [5], and DDcGAN [24]. The qualitative results on the *TNO* and *RoadScene* datasets are shown respectively in Figs. 11 and 12. Overall, U2Fusion exhibits a sharper appearance than its competitors. As shown in the highlighted regions, the competitors lose some details, *e.g.*, cars, the logo, and the license plate. In comparison, U2Fusion alleviates this problem by presenting more details. Moreover, in the extreme case where little information is available in one of the source images, U2Fusion preserves the information in the other source image more completely in the fusion result, as shown in the last row in Fig. 11 and the first row in Fig. 12. Furthermore, U2Fusion is also applied to fuse VIS (RGB) and gray IR images in the *RoadScene*. As shown in Fig. 13, fusion results are more like VIS images enhanced by IR images for better scene representation because the fusion process is performed only on the Y channel and the chromatic information all come from VIS images.

Quantitative comparisons are performed on the remaining 20 and 45 image pairs in *TNO* and *RoadScene*. Four metrics, namely, correlation coefficient (CC), SSIM, peak signal-to-noise ratio (PSNR), and the sum of the correlations of differences (SCD) [43], are used for evaluation. CC measures the linear correlation degree between source images and the result. PSNR evaluates the distortion caused by the fusion process. SCD quantifies the quality of fusion images. As shown in Tab. 1, U2Fusion ranks first on CC, SSIM, and PSNR on both datasets. Although it ranks second on SCD, it achieves comparable results. The promising results show that U2Fusion achieves high fidelity with source images and less distortion, noise, or artifacts.

4.2.2 Medical Image Fusion

We compare U2Fusion with RPCNN [44], CNN [16], PA-PCNN [45], and NSCT [6] on the *Harvard* dataset. As shown in Fig. 14, our results have more structural (texture) information under the premise of little loss of functional (color) information. The quantitative evaluation of four

0162-8828 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Carleton University. Downloaded on August 03,2020 at 08:47:48 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2020.3012548, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 1
Mean and standard deviation of four metrics on VIS-IR image fusion on the *TNO* and *RoadScene* datasets (**red**: the best, **blue**: the second best).

Method	*TNO*				*RoadScene*			
	CC	SSIM	PSNR	SCD	CC	SSIM	PSNR	SCD
HMSD	0.464±0.13	**1.9889±0.007**	**62.687±2.67**	1.666±0.15	0.600±0.20	**1.9904±0.005**	**63.146±2.58**	1.508±0.27
GTF	0.352±0.12	1.9860±0.007	61.782±3.02	0.977±0.20	0.499±0.26	1.9863±0.009	62.013±3.26	1.007±0.17
DenseFuse	**0.533±0.10**	1.9797±0.010	59.953±1.99	1.635±0.16	0.565±0.21	1.9873±0.005	61.126±1.73	1.310±0.30
FusionGAN	0.458±0.10	1.9824±0.008	60.535±1.98	1.403±0.31	0.494±0.26	1.9850±0.009	61.341±2.59	0.844±0.52
DDcGAN	0.414±0.11	1.9824±0.006	60.248±1.49	1.269±0.18	0.506±0.20	1.9805±0.009	60.051±2.32	1.187±0.26
FusionDN	0.499±0.12	1.9875±0.004	61.691±1.29	**1.805±0.12**	**0.627±0.21**	1.9866±0.007	61.684±2.43	**1.778±0.17**
U2Fusion	**0.537±0.11**	**1.9909±0.005**	**62.914±2.07**	**1.780±0.11**	**0.635±0.20**	**1.9909±0.005**	**63.305±2.35**	**1.635±0.24**

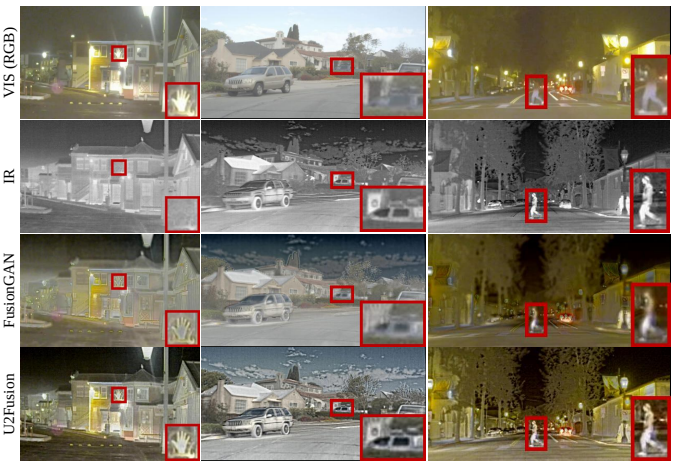


Fig. 13. Qualitative results on 3 typical VIS (RGB) and IR image pairs in the *RoadScene* dataset.

TABLE 2
Mean and standard deviation of four metrics on medical image fusion on the *Harvard* dataset.

Method	SCD	CC	SSIM	PSNR
RPCNN	**1.429±0.08**	0.785±0.09	1.9865±0.004	61.213±1.28
CNN	1.272±0.23	0.798±0.09	**1.9885±0.005**	**62.137±1.79**
PAPCNN	1.289±0.12	0.784±0.09	1.9872±0.005	61.565±1.41
NSCT	0.969±0.20	0.769±0.09	1.9875±0.005	61.695±1.49
FusionDN	0.742±0.23	**0.805±0.09**	1.9769±0.008	59.178±1.34
U2Fusion	**1.312±0.04**	**0.834±0.08**	**1.9921±0.002**	**63.458±1.15**

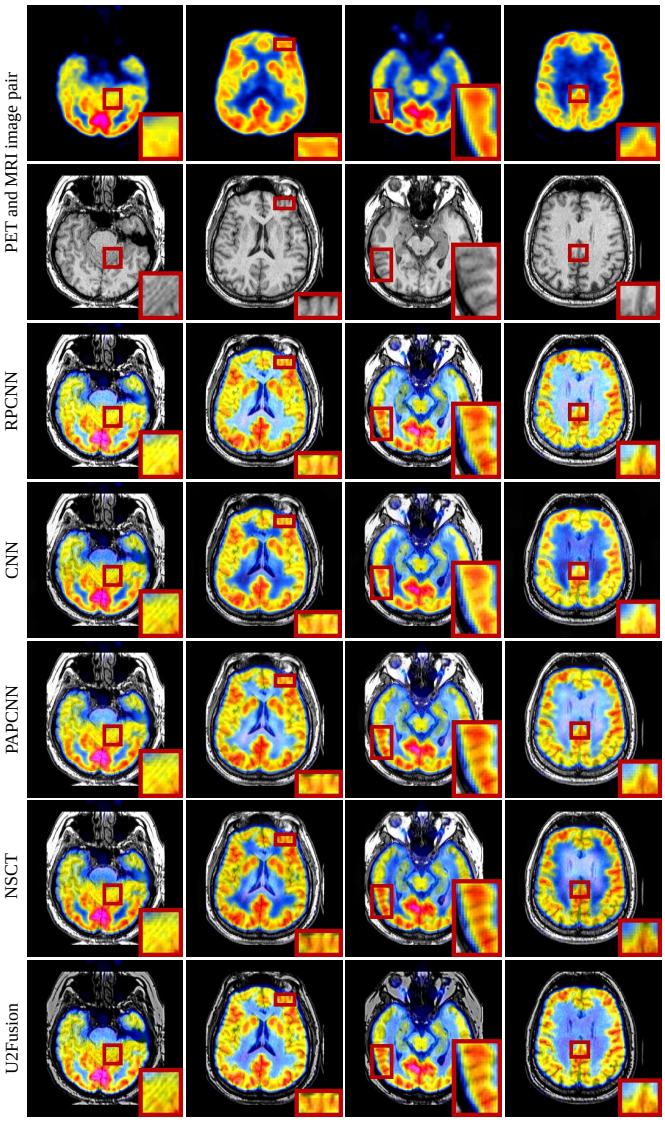


Fig. 14. Qualitative comparison of U2Fusion with 4 state-of-the-art methods on 4 typical PET and MRI image pairs in *Harvard* medical dataset.

metrics in Sec. 4.2.1 is performed on the remaining 10 test image pairs, as reported in Tab. 2. The best results on CC, SSIM, and PSNR indicate that U2Fusion achieves higher correlation and similarity with source images and produces less distortion/noise. The suboptimal result on SCD shows that U2Fusion achieves comparable correlation between the difference and source images.

4.3 Multi-exposure Image Fusion

We compare U2Fusion with GFF [46], DSIFT [47], GBM [48], Deepfuse [7], and FLER [49] on the more challenging problem where source images have a large exposure ratio and thus contain little information. Qualitative results on the dataset in [41] and the *EMPA HDR* dataset are respectively reported in Figs. 15 and 16. Given the inappropriate exposure settings in source images, the representations of the scene are weakened with poor visual perception. In

0162-8828 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Carleton University. Downloaded on August 03,2020 at 08:47:48 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2020.3012548, IEEE Transactions on Pattern Analysis and Machine Intelligence

Fig. 15. Qualitative comparison of U2Fusion with 5 state-of-the-art methods on 5 typical multi-exposure image pairs in the dataset in [41].

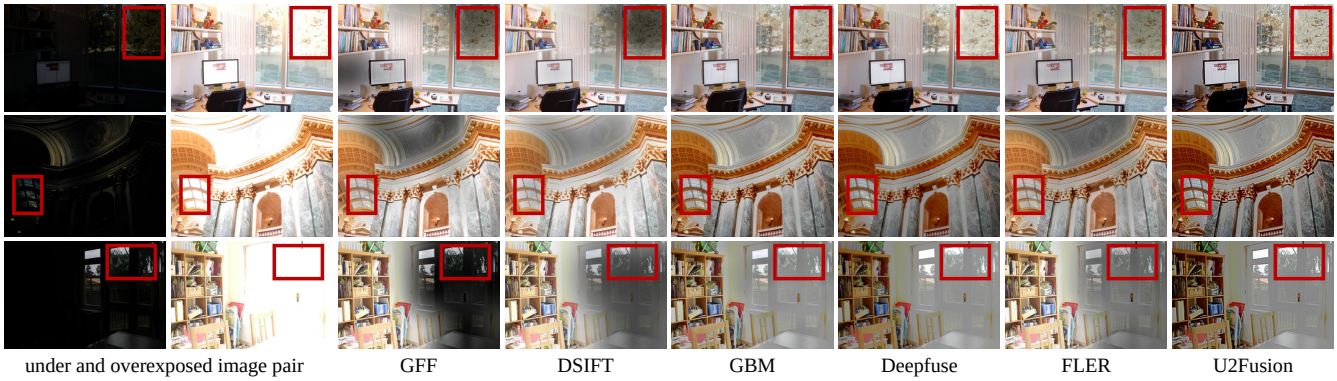


Fig. 16. Qualitative comparison of U2Fusion with 5 state-of-the-art methods on 3 typical multi-exposure image pairs in the *EMPA HDR* dataset.

our result, these representations are further enhanced with appropriate exposure. The local dark regions in GFF, DSIFT, and FLER are improved in U2Fusion. Moreover, compared with GBM and Deepfuse, our results are enriched with clearer details or higher contrast to provide better detail representation, as shown in the red boxes.

Quantitative comparisons are performed on 30 and 15 image pairs in the dataset in [41] and the *EMPA HDR* dataset, respectively. In addition to SSIM, PSNR, and CC, an additional metric, edge intensity (EI), is used for evaluation. EI reflects the gradient amplitude of edge point. The mean and standard deviation are shown in Tab. 3. On the dataset in [41], U2Fusion achieves the optimal mean on SSIM and PSNR. The results on EI and CC follow behind FusionDN and Deepfuse by 0.02 and 0.011, respectively. On the *EMPA HDR* dataset, our mean on SSIM is the best one. For other metrics, U2Fusion achieves 0.037, 0.064, and 0.009, which are close to the best values. These results show that in U2Fusion, the similarity and correlation between the fusion image and source images are higher and have less distortion and larger gradient amplitude.

4.4 Multi-focus Image Fusion

We compare our method with DSIFT [50], GBM [48], CNN [25], GFDF [8], and SESF-Fuse [18] with qualitative results shown in Fig. 17. Although U2Fusion does not use the ground truth for supervision nor does it extract and fill focused regions in fusion images, it still achieves comparable results. As shown in the first row, edges blurred at the boundary of focused and defocused regions are fused into results in the competitors. In U2Fusion, this phenomenon has been alleviated as it attempts to reconstruct the focused regions after judging their relative blurring relationship. The other difference is shown in the last two rows, in DSIFT, CNN, GFDF, and SESF-Fuse, at the boundary of focused and defocused regions. Some details in the far-focused images are lost, *e.g.*, the golf and the edge of the ear. Although GBM retains these details, noticeable brightness and color deviations can be observed in the results. By comparison, U2Fusion preserves these details to a greater extent.

Metrics for evaluation include EI, CC, visual information fidelity (VIF) [51], and mean gradient (MG). VIF measures the information fidelity by computing the distortion between source images and the fusion result. The larger MG, the more gradients the image contains and the better fusion performance. As shown in Tab. 4, U2Fusion achieves the optimal results on EI and CC. The best result on EI and the suboptimal result on MG indicate more gradients in our results for sharper appearance. The results are consistent with the qualitative results shown in Fig. 17. Moreover,

0162-8828 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Carleton University. Downloaded on August 03,2020 at 08:47:48 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2020.3012548, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 3
Mean and standard deviation of four metrics on multi-exposure image fusion on the dataset in [41] and the *EMPA HDR* dataset.

Method	dataset in [41]				*EMPA HDR*			
	SSIM	EI	PSNR	CC	SSIM	EI	PSNR	CC
GFF	1.937±0.01	0.218±0.07	54.604±0.95	0.065±0.35	1.954±0.03	**0.242±0.11**	57.108±3.59	0.423±0.36
DSIFT	1.940±0.01	0.193±0.06	54.856±0.83	0.088±0.35	1.958±0.02	0.222±0.11	57.415±3.49	0.524±0.29
GBM	1.953±0.01	0.230±0.07	55.965±0.97	0.793±0.05	1.966±0.02	0.237±0.12	58.145±2.87	0.782±0.12
Deepfuse	**1.953±0.01**	0.194±0.07	**55.992±1.09**	**0.848±0.05**	**1.967±0.02**	0.181±0.10	**58.518±3.20**	**0.840±0.10**
FLER	1.947±0.01	0.212±0.07	55.425±0.81	0.337±0.31	1.961±0.02	0.235±0.11	57.770±3.41	0.518±0.34
FusionDN	1.943±0.01	**0.285±0.12**	54.969±0.85	0.817±0.056	1.965±0.02	**0.277±0.12**	57.780±2.44	0.801±0.12
U2Fusion	**1.954±0.01**	**0.265±0.09**	**56.074±0.96**	**0.837±0.05**	**1.968±0.02**	0.240±0.13	**58.454±3.07**	**0.831±0.10**

Fig. 17. Qualitative comparison of U2Fusion with 5 state-of-the-art methods on 3 typical far-/near-focused image pairs in the *Lytro* dataset.

TABLE 4
Mean and standard deviation of four metrics on multi-focus image fusion on the *Lytro* dataset.

Method	EI	CC	VIF	MG ($\times 10^{-3}$)
DSIFT	0.300±0.11	0.969±0.01	1.140±0.05	34.636±13.59
GBM	0.294±0.09	0.927±0.02	1.203±0.26	33.575±12.25
CNN	0.297±0.11	**0.970±0.01**	1.130±0.05	34.294±13.56
GFDF	0.299±0.11	0.969±0.01	1.136±0.06	34.436±13.57
SESF-Fuse	0.300±0.11	0.969±0.01	1.145±0.06	34.568±13.58
FusionDN	**0.315±0.11**	0.969±0.01	**1.505±0.34**	**35.080±12.79**
U2Fusion	**0.316±0.11**	**0.972±0.01**	**1.466±0.20**	**34.767±12.67**

the best result on CC and the optimal result on VIF show that U2Fusion maintains the highest linear correlation with source images and achieves comparable information fidelity.

5 Ablation Experiments

5.1 Ablation Study about EWC

In U2Fusion, we use EWC to train a single model for three fusion tasks to overcome catastrophic forgetting. To validate its effectiveness, we perform a comparison experiment where tasks are sequentially trained without EWC. The effectiveness is analyzed from three aspects: i) the similarity loss, ii) statistical distributions of μ_i, and iii) intermediate fusion results during the training phase.

Changes in the similarity loss, $\mathcal{L}_{\text{sim}}(\theta, D)$ in Eq. (3), are shown in Fig. 18. The first plot is the similarity loss of each task without applying EWC, and the second plot is that with EWC. The difference is not evident between the losses of tasks 1 and 2. However, when training DenseNet on task 3 without EWC, the loss on the validation dataset of task 2 increases evidently. It indicates that the performance of the current network on multi-exposure image fusion is declining. With EWC, the similarity losses of previous tasks are basically the same as those when they were trained. Thus, by applying EWC, we can obtain a single model applicable to these tasks.

We also compare the statistical distributions of μ_i with/without EWC, as shown in Fig. 19. μ_i is computed by the similarity loss and corresponding datasets after each task has been trained. For example, the distribution after training task 3 is the statistical distribution of the mean μ_i obtained by averaging μ_i computed by the similarity loss and dataset of task 1 and those of task 2. Without EWC, not much difference is observed among the three

0162-8828 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Carleton University. Downloaded on August 03,2020 at 08:47:48 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2020.3012548, IEEE Transactions on Pattern Analysis and Machine Intelligence

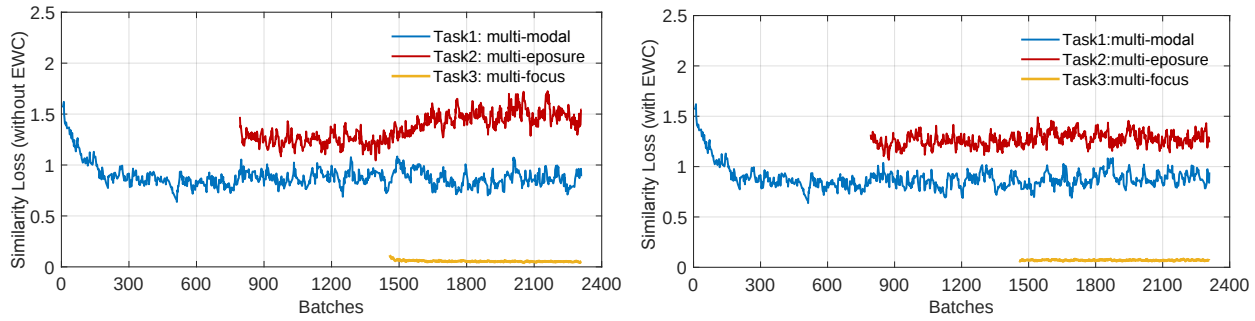


Fig. 18. Changes of the similarity loss without EWC (the first plot) or with EWC (the second plot).

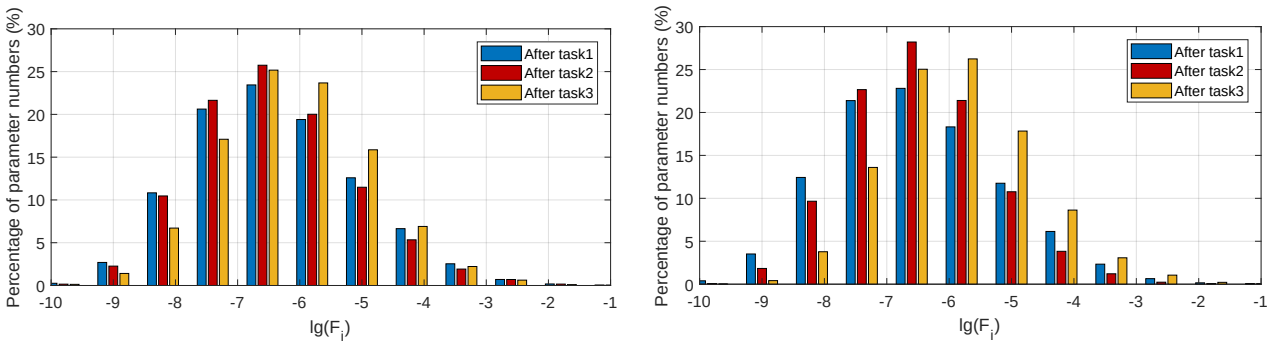


Fig. 19. Changes of statistical distributions of μ_i without EWC (the first plot) or with EWC (the second plot).

distributions of μ_i obtained after three tasks, as shown in the first plot. The parameters are only related to the current task, as μ_i only shows the importance of parameters to the current task. However, with EWC, the proportion of larger μ_i has increased significantly. This increase shows that more important parameters are present in the network. These parameters are significant not only to the current task but also to previous ones. Meanwhile, the decreased proportion of small values also shows that the redundancy of the network is decreasing. An increasing number of parameters play an import role in improving the fusion performance.

The intuitive qualitative comparisons of results with/without EWC are given in Fig. 20. After training the model on tasks 1 and 2, the models with and without EWC achieve satisfactory results on multi-modal and multi-exposure image fusion. Given that it has not been trained on task 3, the results of multi-focus image fusion show blurred edges, as shown in the results of task 3 in Figs. 20 (a), (b), and (d). However, by training the model on task 3, the results exhibit a sharper appearance, as shown in the results of task 3 in Figs. 20 (d) and (e). When the model is trained without EWC, the performance on task 2 declines, *e.g.*, the lower luminance of the whole image. Moreover, evident difference is observed between the results of task 1 in Figs. 20 (b) and (c). With EWC, these two problems have been alleviated, as shown in Figs. 20 (d) and (e).

5.2 A Unified Model for Mutual Promotion between Different Tasks

In U2Fusion, we employ EWC to learn from new tasks continuously. In this way, the unified model is capable of fusing multiple types of source images. Thus, with unified parameters, the information learned by U2Fusion from a single task can promote other tasks. For verification, we train an individual model for each task. Thus, no interaction occurs among different tasks. The fusion results are shown in Fig. 21. Although multi-modal and multi-focus image fusions are different from multi-exposure image fusion, multi-modal and multi-focus images also have overexposed regions, which can be evidently seen from the visible images in the first three columns and the far-focused image in the last column. With a unified model that has been trained for multi-exposure image fusion, U2Fusion shows better performance for these overexposed regions with clearer representation than individual models. Another instance is shown in the results of multi-exposure image fusion, *i.e.*, the sixth column. The highlighted regions in the source images are similar to multi-focus images. Given that the model has learned from multi-focus image fusion, the result of U2Fusion exhibits clearer and sharper edges than that of the individually trained model. Thus, by gathering the strengths of multiple tasks, U2Fusion obtains the strong generalization not only for multiple types of source images but also for multiple types of regions in the same type of source images. Therefore, a unified model can realize the mutual promotion of different fusion tasks.

5.3 Ablation Study about Adaptive Information Preservation Degrees

To validate the effectiveness of adaptive information preservation degrees, we perform the experiments where ω_1 and ω_2 are directly set to 0.5. The comparative results on the six datasets are shown in Fig. 22. The results in the first row are obtained when ω_1 and ω_2 are fixed to 0.5, and those in the second row are the results of U2Fusion. In multi-modal image fusion, the results without adaptive information preservation degrees show worse detail representation,

0162-8828 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Carleton University. Downloaded on August 03,2020 at 08:47:48 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2020.3012548, IEEE Transactions on Pattern Analysis and Machine Intelligence

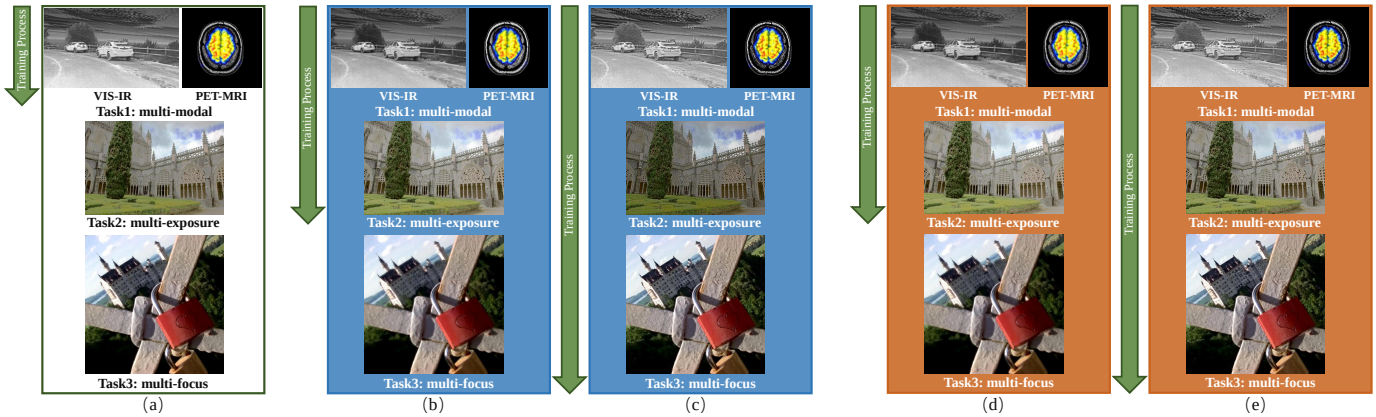


Fig. 20. Intermediate fusion results. From left to right: (a) fusion results after training the model on task 1; (b) and (c): fusion results after training the model on task 2 and task 3 without EWC; (d) and (e): fusion results after training the model on task 2 and task 3 with EWC.

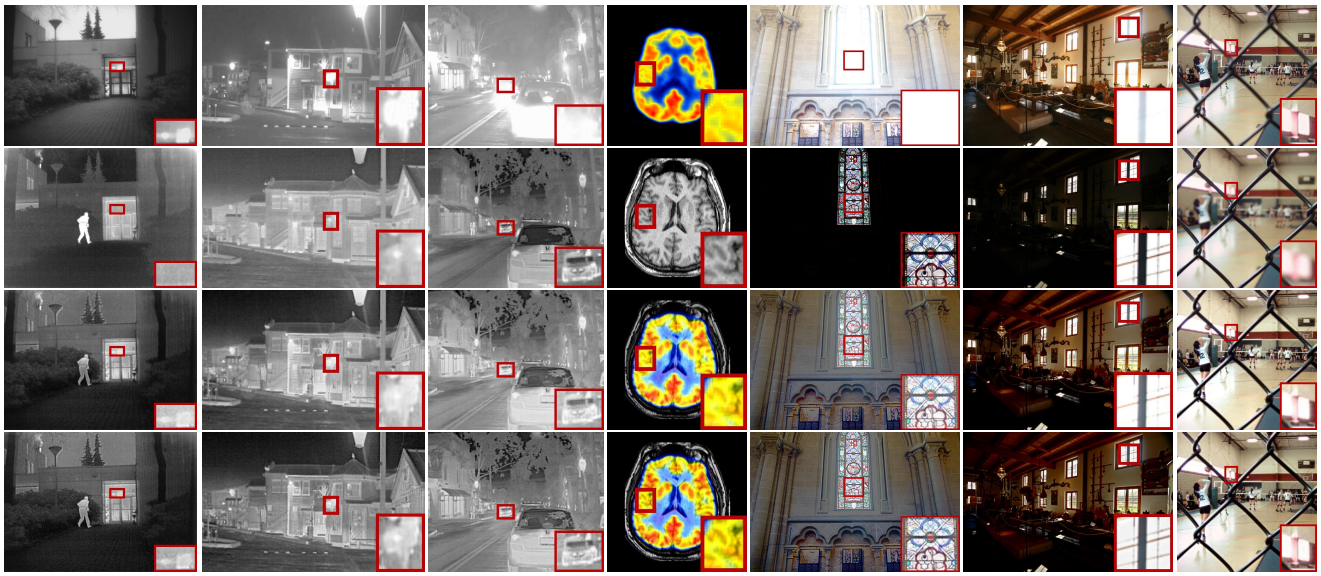


Fig. 21. Illustration of mutual promotion between different tasks in a unified model. From top to bottom: source images, fusion results by training individual models for each fusion task and fusion results of U2Fusion. From left to right: images from the *TNO*, *RoadScene* (the second and third columns), *Harvard*, the dataset in [41], *EMPA HDR* and *Lytro* datasets.

as shown in the edges of the cloud, textures of the jeep, details of the net, and the structural information. In multi-exposure image fusion, the difference is clearly seen in the overexposed regions. Without the adaptive degrees, these regions still look overexposed, such as the flower, window, and the sun. The phenomenon is most noticeable in the results of multi-focus image fusion. When ω_1 and ω_2 are directly set to 0.5, the network fails to distinguish between focused and defocused regions. Therefore, the results suffer from blurred edges, while U2Fusion generates a much sharper appearance.

5.4 Effect of Training Order

In the three fusion tasks, multi-focus image fusion is a little different from multi-modal and multi-exposure image fusion. For multi-modal and multi-exposure image patches, the fusion patch can be seen as the combination of two source images. However, for multi-focus image patches, the fusion process can be seen as the selection of the focused regions in the source images. Thus, the fusion result is expected to exhibit a high similarity with source images in the focused region. Therefore, we perform two comparison experiments in this section. For quantitative comparison, we use the correlation coefficient (CC) to measure the correlation between the result and source images and mean gradient (MG) to measure the performance of the fusion results.

On the one hand, we change the order of multi-modal and multi-exposure image fusion. The training order is reset as multi-exposure→multi-modal→multi-focus image fusion. The qualitative results are shown in Fig. 23, and the quantitative results are shown in Tab. 5. As shown in the results, the exchange of the training orders of multi-modal and multi-exposure image fusion has little effect on fusing multi-focus images. For these two tasks, the results exhibit higher brightness and mean gradient. However, the results of the original training order maintain a higher correlation with the source images.

On the other hand, considering the difference between

0162-8828 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Carleton University. Downloaded on August 03,2020 at 08:47:48 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2020.3012548, IEEE Transactions on Pattern Analysis and Machine Intelligence

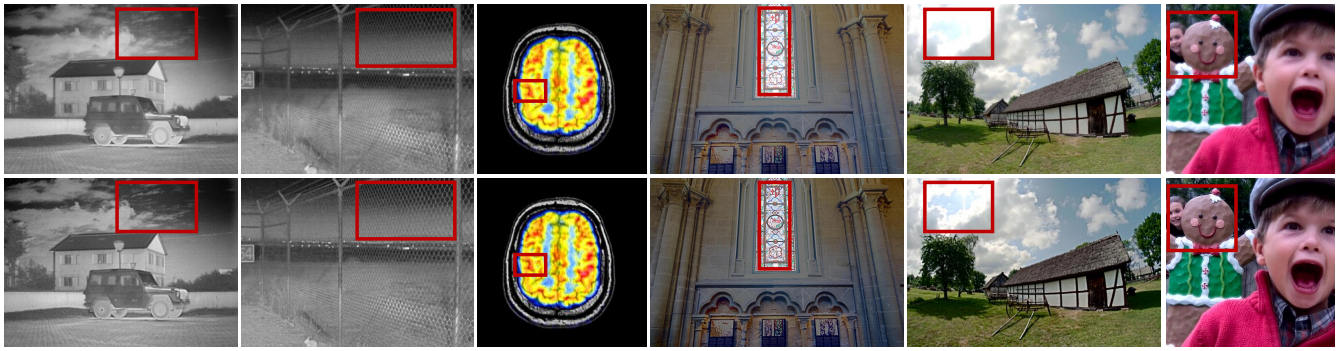


Fig. 22. Comparative qualitative results of our method without (the first row) and with (the second row) adaptive information preservation degrees. From left to right: fusion images of image pairs from the *TNO*, *RoadScene*, *Harvard*, the dataset in [41], *EMPA HDR* and *Lytro* datasets.

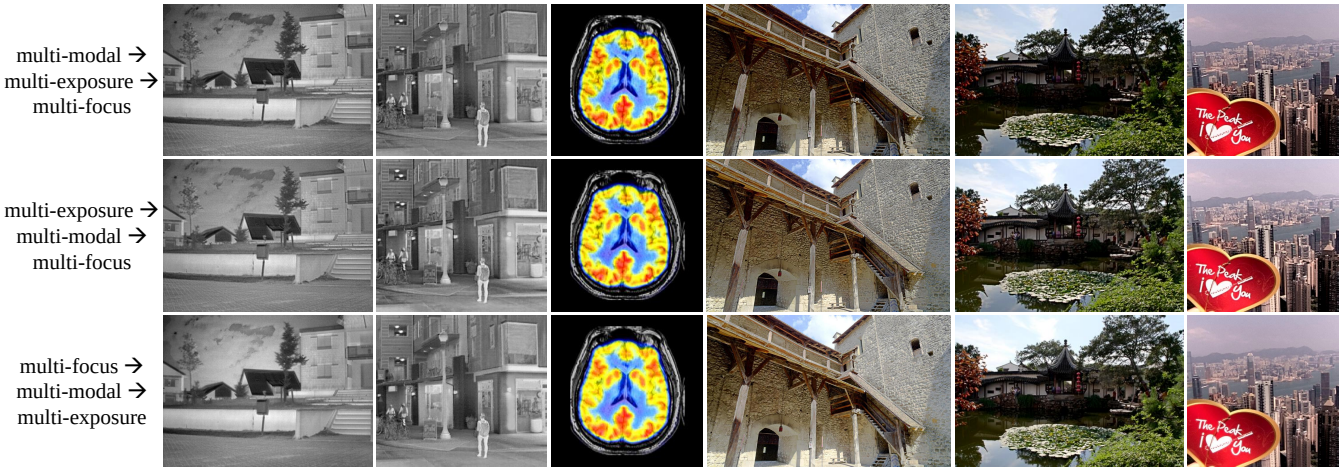


Fig. 23. Fusion results of different training orders. From left to right: results of image pairs from the *TNO*, *RoadScene*, *Harvard*, the dataset in [41], *EMPA HDR* and *Lytro* datasets.

TABLE 5
Mean of two metrics (correlation coefficient/mean gradient) of different training orders on different datasets.

Training Order	multi-modal dataset			multi-exposure dataset		multi-focus dataset
	TNO	*RoadScene*	*Harvard*	dataset in [41]	*EMPA HDR*	*Lytro*
modal→exposure →focus	**0.5370**/**0.0467**	**0.6347**/**0.0594**	**0.8442**/**0.0933**	**0.8378**/**0.0647**	**0.8158**/**0.0597**	0.9723/**0.0677**
exposure→modal →focus	**0.5359**/**0.0480**	0.6242/**0.0592**	0.8335/**0.0950**	0.8191/**0.0672**	0.8065/**0.0639**	**0.9724**/**0.0700**
focus→modal →exposure	0.5319/0.0383	**0.6613**/0.0516	**0.8406**/0.0873	**0.8322**/0.0611	**0.8081**/0.0572	**0.9803**/0.0563

the multi-focus image fusion and the two other fusion tasks, we set multi-focus image fusion as the first task. Then, the training order is reset as multi-focus→multi-modal→multi-exposure image fusion. Evidently, the result of multi-focus image fusion is more blurred than those of other orders, which can be seen from the rightmost column in Fig. 23. This phenomenon is also reflected by the substantially reduced mean gradient in Tab. 5, which drops from 0.0677 or 0.0700 to 0.0563. The ability of U2Fusion for continual learning benefits from $\mathcal{L}_{\text{ewc}}(\theta, D)$ is defined in Eq. (7). Some unimportant parameters are updated to learn from new tasks, resulting in a slight performance degradation on the previous tasks. Given the particularity of multi-focus image fusion, the performance degradation is more evident, especially reflecting in the blurring of shape edges.

Therefore, the training orders of multi-modal and multi-exposure image fusion have little effect on the fusion results, while that of multi-focus has a relatively significant effect. Comparing the quantitative results in Tab. 5, the order of multi-modal→multi-exposure→multi-focus shows the best performance. Thus, we adopt it in U2Fusion.

5.5 U2Fusion *vs.* FusionDN

The preliminary version of the proposed method is FusionDN [11], and the improvements are described in Sec. 1.

0162-8828 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Carleton University. Downloaded on August 03,2020 at 08:47:48 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2020.3012548, IEEE Transactions on Pattern Analysis and Machine Intelligence

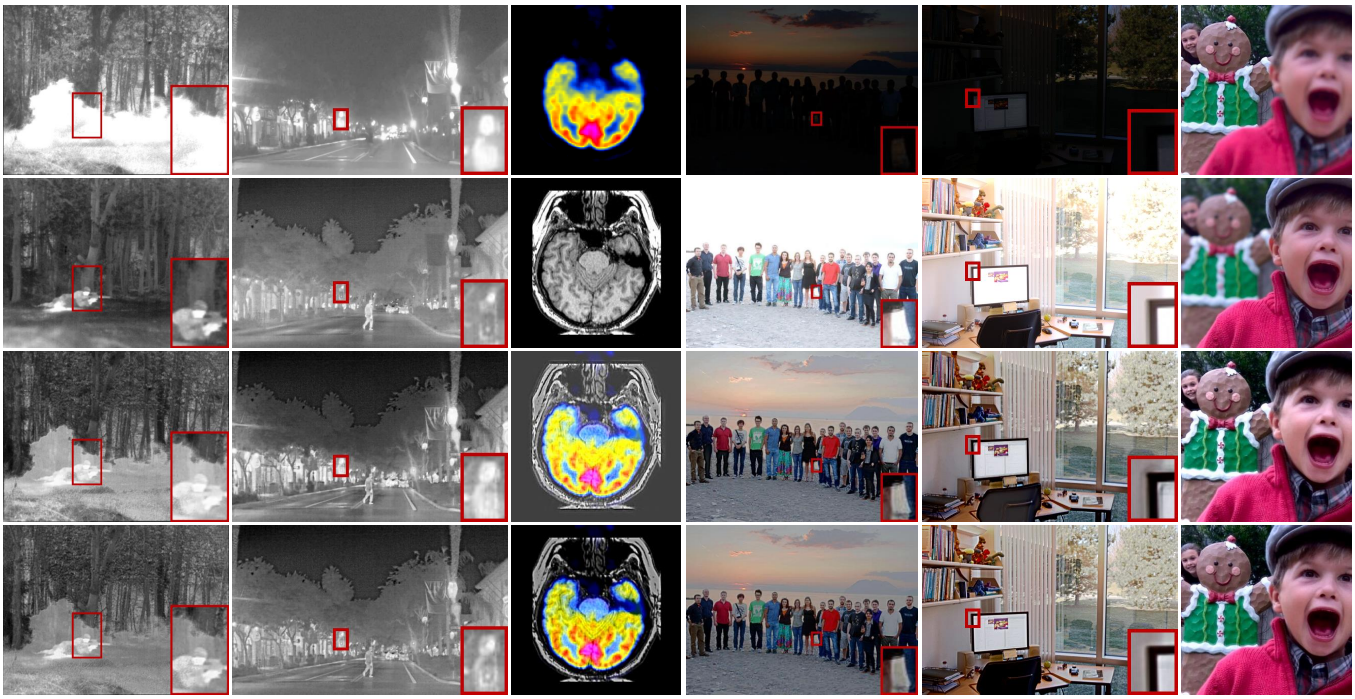


Fig. 24. Comparative results of our U2Fusion with the previous version FusionDN. From top to bottom: source images, result of FusionDN and U2Fusion. From left to right: image pairs from the *TNO*, *RoadScene*, *Harvard*, the dataset in [41], *EMPA HDR* and *Lytro* datasets.

To validate the effectiveness of these improvements, we compare the results of FusionDN and U2Fusion, as shown in Fig. 24.

First, we improve the strategy for information preservation degree assignment by modifying the amount and quality of information in source images. The effect of this improvement is shown in the first and second columns in Fig. 24. Relying on the amount and quality information in original source images, FusionDN preserves the high contrast in VIS regions, such as that between the smoke and the background. Nevertheless, a considerable amount of details in the corresponding IR regions have been lost. In U2Fusion, by considering the information in abundant extracted features, the information preservation degrees are changed, and more details in source images are preserved.

Second, we modify the loss function by removing the gradient loss and adding the MSE loss. In FusionDN, the gradient loss is introduced to preserve more gradients. However, it causes some false edges, as in the results of FusionDN in the fourth and fifth columns. By removing it, we rely on SSIM and the improved information preservation degree assignment strategy to preserve the structural information. The results still show sharp appearance and alleviate false edges. Moreover, given that the intensity distribution is preserved solely by SSIM, the luminance component of the result shows slight deviation from source images, as shown in the result of FusionDN in the last column. In U2Fusion, to overcome the luminance deviation, we add the MSE loss. As in the last column, the intensity of U2Fusion is more similar to that of source images.

Lastly, we replace the first fusion task from VIS-IR image fusion to multi-modal image fusion. In this task, VIS-IR and PET-MRI image fusion are included. As the model in FusionDN has not been trained on the medical dataset, the result seems unsatisfactory with weak edges and gray background, as shown in the third column.

6 Conclusion

In this study, we propose a novel unified and unsupervised end-to-end image fusion network, termed as *U2Fusion*, to solve multiple fusion problems. First, adaptive information preservation degrees are obtained as the measurement of the amount of information contained in source images. Thus, different tasks are solved under a unified framework. In particular, the adaptive degrees allows the network to be trained to preserve the adaptive similarity between the fusion result and source images. Consequently, the ground truth is not required. Moreover, we solve the catastrophic forgetting problem as well as the storage and computation issues to train a single model applicable to multiple problems. This single model is capable of solving multi-modal, multi-exposure, and multi-focus image fusion problems with high-quality results. The qualitative and quantitative results validate the effectiveness and universality of U2Fusion. Moreover, we release a new aligned infrared and visible image dataset RoadScene on the basis of FLIR video to provide a new option for image fusion benchmark evaluation.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant nos. 61773295, 61971165 and 61772512, the U.S. National Science Foundation under Grant no. 1618398, and the Natural Science Foundation of Hubei Province under Grant no. 2019CFA037.

0162-8828 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Carleton University. Downloaded on August 03,2020 at 08:47:48 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2020.3012548, IEEE Transactions on Pattern Analysis and Machine Intelligence

REFERENCES

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

[2] J. Ma, Y. Ma, and C. Li, "Infrared and visible image fusion methods and applications: A survey," *Inf. Fusion*, vol. 45, pp. 153–178, 2019.

[3] Y. Zhang and Q. Ji, "Active and dynamic information fusion for facial expression understanding from image sequences," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 699–714, 2005.

[4] Z. Liu, E. Blasch, Z. Xue, J. Zhao, R. Laganiere, and W. Wu, "Objective assessment of multiresolution image fusion algorithms for context enhancement in night vision: a comparative study," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 94–109, 2011.

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

[6] Z. Zhu, M. Zheng, G. Qi, D. Wang, and Y. Xiang, "A phase congruency and local laplacian energy based multi-modality medical image fusion method in nsct domain," *IEEE Access*, vol. 7, pp. 20 811–20 824, 2019.

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

[8] X. Qiu, M. Li, L. Zhang, and X. Yuan, "Guided filter-based multi-focus image fusion through focus region detection," *Signal Process. Image Commun.*, vol. 72, pp. 35–46, 2019.

[9] Y. Liu, X. Chen, Z. Wang, Z. J. Wang, R. K. Ward, and X. Wang, "Deep learning for pixel-level image fusion: Recent advances and future prospects," *Inf. Fusion*, vol. 42, pp. 158–173, 2018.

[10] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4700–4708.

[11] H. Xu, J. Ma, Z. Le, J. Jiang, and X. Guo, "Fusiondn: A unified densely connected network for image fusion," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12 484–12 491.

[12] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska *et al.*, "Overcoming catastrophic forgetting in neural networks," *Proc. Natl. Acad. Sci.*, vol. 114, no. 13, pp. 3521–3526, 2017.

[13] L. Cao, L. Jin, H. Tao, G. Li, Z. Zhuang, and Y. Zhang, "Multi-focus image fusion based on spatial frequency in discrete cosine transform domain," *IEEE Signal Process. Lett.*, vol. 22, no. 2, pp. 220–224, 2014.

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

[15] Y. Liu, X. Chen, R. K. Ward, and Z. J. Wang, "Image fusion with convolutional sparse representation," *IEEE Signal Process. Lett.*, vol. 23, no. 12, pp. 1882–1886, 2016.

[16] Y. Liu, X. Chen, J. Cheng, and H. Peng, "A medical image fusion method based on convolutional neural networks," in *Proc. Int. Conf. Inf. Fusion*, 2017, pp. 1–7.

[17] H. Li and X.-J. Wu, "Densefuse: A fusion approach to infrared and visible images," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2614–2623, 2018.

[18] B. Ma, X. Ban, H. Huang, and Y. Zhu, "Sesf-fuse: An unsupervised deep model for multi-focus image fusion," *arXiv preprint arXiv:1908.01703*, 2019.

[19] J. Ma, C. Chen, C. Li, and J. Huang, "Infrared and visible image fusion via gradient transfer and total variation minimization," *Inf. Fusion*, vol. 31, pp. 100–109, 2016.

[20] K. Ma, Z. Duanmu, H. Yeganeh, and Z. Wang, "Multi-exposure image fusion by optimizing a structural similarity index," *IEEE Transactions on Computational Imaging*, vol. 4, no. 1, pp. 60–72, 2017.

[21] Y. Liu, S. Liu, and Z. Wang, "Multi-focus image fusion with dense sift," *Inf. Fusion*, vol. 23, pp. 139–155, 2015.

[22] J. Ma, P. Liang, W. Yu, C. Chen, X. Guo, J. Wu, and J. Jiang, "Infrared and visible image fusion via detail preserving adversarial learning," *Inf. Fusion*, vol. 54, pp. 85–98, 2020.

[23] H. Xu, P. Liang, W. Yu, J. Jiang, and J. Ma, "Learning a generative model for fusing infrared and visible images via conditional generative adversarial network with dual discriminators," in *Proc. Int. Joint Conf. Artif. Intell.*, 2019, pp. 3954–3960.

[24] J. Ma, H. Xu, J. Jiang, X. Mei, and X.-P. Zhang, "Ddcgan: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 4980–4995, 2020.

[25] Y. Liu, X. Chen, H. Peng, and Z. Wang, "Multi-focus image fusion with a deep convolutional neural network," *Inf. Fusion*, vol. 36, pp. 191–207, 2017.

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

[27] X. Guo, R. Nie, J. Cao, D. Zhou, L. Mei, and K. He, "Fusegan: Learning to fuse multi-focus image via conditional generative adversarial network," *IEEE Trans. Multimed.*, vol. 21, no. 8, pp. 1982–1996, 2019.

[28] S.-W. Lee, J.-H. Kim, J. Jun, J.-W. Ha, and B.-T. Zhang, "Overcoming catastrophic forgetting by incremental moment matching," in *Advances in Neural Information Processing Systems*, 2017, pp. 4652–4662.

[29] F. Zenke, B. Poole, and S. Ganguli, "Continual learning through synaptic intelligence," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3987–3995.

[30] N. Sugianto, D. Tjondronegoro, G. Sorwar, P. Chakraborty, and E. I. Yuwono, "Continuous learning without forgetting for person re-identification," in *Proc. IEEE Int. Conf. Advanced Video and Signal Based Surveillance*, 2019, pp. 1–8.

[31] J. A. Eichel, A. Mishra, N. Miller, N. Jankovic, M. A. Thomas, T. Abbott, D. Swanson, and J. Keller, "Diverse large-scale its dataset created from continuous learning for real-time vehicle detection," *arXiv preprint arXiv:1510.02055*, 2015.

[32] R. Xia and Y. Liu, "A multi-task learning framework for emotion recognition using 2d continuous space," *IEEE Trans. Affect. Comput.*, vol. 8, no. 1, pp. 3–14, 2015.

[33] Y. Wei, W. Xia, M. Lin, J. Huang, B. Ni, J. Dong, Y. Zhao, and S. Yan, "Hcp: A flexible cnn framework for multi-label image classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 9, pp. 1901–1907, 2015.

[34] Y. Qi, S. Zhang, L. Qin, Q. Huang, H. Yao, J. Lim, and M.-H. Yang, "Hedging deep features for visual tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 5, pp. 1116–1130, 2018.

[35] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Europ. Conf. Comput. Vis.*, 2016, pp. 694–711.

[36] G. Mai, K. Cao, P. C. Yuen, and A. K. Jain, "On the reconstruction of face images from deep face templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 5, pp. 1188–1202, 2018.

[37] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[38] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli *et al.*, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.

[39] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 2935–2947, 2017.

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

[41] J. Cai, S. Gu, and L. Zhang, "Learning a deep single image contrast enhancer from multi-exposure images," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 2049–2062, 2018.

[42] Z. Zhou, B. Wang, S. Li, and M. Dong, "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] V. Aslantas and E. Bendes, "A new image quality metric for image fusion: the sum of the correlations of differences," *Aeu-Int. J. Electr. Commun.*, vol. 69, no. 12, pp. 1890–1896, 2015.

[44] S. Das and M. K. Kundu, "A neuro-fuzzy approach for medical image fusion," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 12, pp. 3347–3353, 2013.

[45] M. Yin, X. Liu, Y. Liu, and X. Chen, "Medical image fusion with parameter-adaptive pulse coupled neural network in nonsubsampled shearlet transform domain," *IEEE Trans. Instrum. Meas.*, vol. 68, no. 1, pp. 49–64, 2018.

[46] S. Li, X. Kang, and J. Hu, "Image fusion with guided filtering," *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2864–2875, 2013.

0162-8828 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Carleton University. Downloaded on August 03,2020 at 08:47:48 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2020.3012548, IEEE Transactions on Pattern Analysis and Machine Intelligence

[47] Y. Liu and Z. Wang, "Dense sift for ghost-free multi-exposure fusion," *J. Visual Commun. Image Represent.*, vol. 31, pp. 208–224, 2015.
[48] S. Paul, I. S. Sevcenco, and P. Agathoklis, "Multi-exposure and multi-focus image fusion in gradient domain," *J. Circuit. Syst. Comp.*, vol. 25, no. 10, p. 1650123, 2016.
[49] Y. Yang, W. Cao, S. Wu, and Z. Li, "Multi-scale fusion of two large-exposure-ratio images," *IEEE Signal Process. Lett.*, vol. 25, no. 12, pp. 1885–1889, 2018.
[50] Y. Liu, S. Liu, and Z. Wang, "Multi-focus image fusion with dense sift," *Inf. Fusion*, vol. 23, pp. 139–155, 2015.
[51] Y. Han, Y. Cai, Y. Cao, and X. Xu, "A new image fusion performance metric based on visual information fidelity," *Inf. Fusion*, vol. 14, no. 2, pp. 127–135, 2013.

Han Xu received the B.S. degree from the Electronic Information School, Wuhan University, Wuhan, China, in 2018. She is currently a Ph.D. student in the Multi-spectral Vision Processing Lab, Electronic Information School, Wuhan University, Wuhan. Her current research interests include computer vision and pattern recognition.

Jiayi Ma received the B.S. degree in information and computing science and the Ph.D. degree in control science and engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2008 and 2014, respectively. He is currently a Professor with the Electronic Information School, Wuhan University. He has authored or co-authored more than 130 refereed journal and conference papers, including IEEE TPAMI/TIP, IJCV, CVPR, ICCV, ECCV, *etc*. His research interests include computer vision, machine learning, and pattern recognition. Dr. Ma has been identified in the 2019 Highly Cited Researchers list from the Web of Science Group. He is an Area Editor of *Information Fusion*, an Associate Editor of *Neurocomputing*, and a Guest Editor of *Remote Sensing*.

Junjun Jiang received the B.S. degree in Information and Computing Science from School of Mathematical Sciences, Huaqiao University, Quanzhou, China, in 2009, and the Ph.D. degree in communication and information system from School of Computer, Wuhan University, Wuhan, China, in 2014. He is currently a Professor with the School of Computer Science and Technology, Harbin Institute of Technology. His research interests include applications of image processing and pattern recognition in video surveillance, image super-resolution, image interpolation, and face recognition.

Xiaojie Guo is currently a tenured Associate Professor with the college of intelligence and computing, Tianjin University. He was a recipient of the Piero Zamperoni Best Student Paper Award in the International Conference on Pattern Recognition, in 2010, and a recipient of the IEEE ICME best student paper runner-up award, in 2018.

Haibin Ling received the B.S. and M.S. degrees from Peking University in 1997 and 2000, respectively, and the Ph.D. degree from the University of Maryland, College Park, in 2006. From 2000 to 2001, he was an assistant researcher at Microsoft Research Asia. From 2006 to 2007, he worked as a postdoctoral scientist at the University of California Los Angeles. In 2007, he joined Siemens Corporate Research as a research scientist; then, from 2008 to 2019, he worked as a faculty member of the Department of Computer Sciences at Temple University. In fall 2019, he joined Stony Brook University as a SUNY Empire Innovation Professor in the Department of Computer Science. His research interests include computer vision, augmented reality, medical image analysis, and human computer interaction. He received Best Student Paper Award at ACM UIST (2003), NSF CAREER Award (2014), Yahoo Faculty Research Award (2019), and Amazon AWS Machine Learning Research Award (2019). He serves as Associate Editors for several journals including IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI), Pattern Recognition (PR), and Computer Vision and Image Understanding (CVIU), and has served as Area Chairs various times for CVPR and ECCV.

0162-8828 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Carleton University. Downloaded on August 03,2020 at 08:47:48 UTC from IEEE Xplore. Restrictions apply.