Deep Convolutional Sparse Coding Networks for Interpretable Image Fusion

Zixiang Zhao[1,2] Jiangshe Zhang[1] Haowen Bai[1] Yicheng Wang[1,5] Yukun Cui[1]
Lilun Deng[1] Kai Sun[1] Chunxia Zhang[1] Junmin Liu[1] Shuang Xu[3,4*]
[1] Xi'an Jiaotong University [2] Computer Vision Lab, ETH Zürich
[3] Research and Development Institute of Northwestern Polytechnical University in Shenzhen
[4] Northwestern Polytechnical University [5] The University of Melbourne
zixiangzhao@stu.xjtu.edu.cn, xs@nwpu.edu.cn

Abstract

Image fusion is a significant problem in many fields including digital photography, computational imaging and remote sensing, to name but a few. Recently, deep learning has emerged as an important tool for image fusion. This paper presents ***CSCFuse****, which contains three deep convolutional sparse coding (CSC) networks for three kinds of image fusion tasks (i.e., infrared and visible image fusion, multi-exposure image fusion, and multi-spectral image fusion). The CSC model and the iterative shrinkage and thresholding algorithm are generalized into dictionary convolution units. As a result, all hyper-parameters are learned from data. Our extensive experiments and comprehensive comparisons reveal the superiority of CSCFuse with regard to quantitative evaluation and visual inspection.*

1. Introduction

With the development of computer vision, there has been further deepening of the understanding of scenes, which has led to higher demands for the quality of input images [21,30,51–54,63–65,82]. Image fusion is a fundamental topic in image processing [5,27,59,62,66,75–78], and it aims to generate a fusion image by combining the complementary information of source images [32, 61, 80, 88, 89]. This technique has been applied to many scenarios. For example, infrared and visible image fusion (IVF) is helpful for object detection and recognition [35, 48–50, 84]. In digital photography, high dynamic range (HDR) imaging can be solved by multi-exposure image fusion (MEF) to generate high-contrast and informative images [37, 46, 73].

Over the past a few decades, numerous image fusion algorithms have been proposed, where transform based algorithms are very popular [12, 13, 26, 28, 32, 58, 69–72, 85]. They transform source images into feature domain, detect the active levels, blend the features and at last apply the inverse transformer in order to obtain the fused image. Recently, deep neural networks have emerged as an effective tool in image fusion [22, 32]. They are divided into three groups: (1) Autoencoder-based methods. This is a deep-learning variant of transform-based algorithms. The transformers and inverse transformers are replaced by encoders and decoders, respectively [11, 18, 24, 83, 86, 87, 90]. (2) Supervised methods. For multi-focus image fusion, there are ground truth images in the synthetic datasets [31]. For MEF, Cai et al. constructed a large dataset providing the reference images by comparing 13 MEF/HDR algorithms [3]. Owing to the strong fitting ability, supervised learning networks are suitable for these tasks. (3) Human visual system-based methods. In the case without a reference image, by taking prior knowledge into account and setting proper loss functions, researchers designed regression [38,74] or adversarial [23,25,29,36] networks to make fusion images satisfy human visual systems. However, it is found that many algorithms are evaluated on a limited number of cherry-picked images. Thus, their generalizations remain unknown. It leaves room for possible improvement with reasonable and interpretable formulations.

Convolutional sparse coding (CSC) has been successfully applied to computer vision tasks on account of its high performance and robustness [4, 7, 10, 57]. The CSC model is generally solved by the iterative shrinkage and thresholding algorithm (ISTA), but the results significantly depend on hyper-parameters. To address this problem, the CSC model and ISTA are generalized into some dictionary convolutional units (DCUs) which are put in the hidden layers of neural networks. In this manner, the hyper-parameters (e.g. penalty parameters, dictionary filters and thresholding functions) in DCUs are learnable. Based on the novel unit, we design deep CSC networks for three fusion tasks, including IVF, MEF, and multi-spectral image fusion (MSF). In our experiments, we employ relatively large test datasets to make a comprehensive and convincing evaluation. Ex-

[*] Corresponding author.

perimental results show that the deep CSC networks outperform the state-of-the-art (SOTA) methods in terms of both objective metrics and visual inspection. Besides, our networks are with high reproducibility. The remainder of this paper is organized as follows. Section 2 converts the CSC and ISTA into a DCU. Then, in section 3 we design three DCU based networks for IVF, MEF and MRF tasks. The extensive experiments are reported in section 4. Section 5 concludes this paper.

2. Dictionary Convolutional Units

In dictionary learning, CSC is a typical method for image processing. Given an image $\boldsymbol{x} \in R^{c\times h\times w}$ ($c = 1$ for gray images and $c = 3$ for RGB images) and q convolutional filters $\boldsymbol{d} \in R^{q\times c\times s\times s}$, CSC can be formulated as the following problem:

$$\min_{\boldsymbol{z}} \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{d} * \boldsymbol{z}\|_2^2 + \lambda g(\boldsymbol{z}), \tag{1}$$

where λ is a hyperparameter, $*$ denotes the convolution operator, $\boldsymbol{z} \in R^{q\times h\times w}$ is the sparse feature map (or say, code) and $g(\cdot)$ is a sparse regularizer. This problem can be solved by ISTA, and it is easy to write the updating rule for feature maps as below,

$$\boldsymbol{z}^{(k+1)} \leftarrow \mathrm{prox}_{\lambda/\rho}\left(\boldsymbol{z}^{(k)} + \frac{1}{\rho}\boldsymbol{d}^T * (\boldsymbol{x} - \boldsymbol{d} * \boldsymbol{z}^{(k)})\right), \tag{2}$$

where ρ is the step size and $\boldsymbol{d}^T \in R^{c\times q\times s\times s}$ is the flipped version of $\boldsymbol{d}$ along horizontal and vertical directions. Note that $\mathrm{prox}(\cdot)$ is the proximal operator of the regularizer $g(\cdot)$. If $g(\cdot)$ is the ℓ_1-norm, its corresponding proximal operator is the soft shrinkage thresholding (SST) function defined by $\mathrm{SST}_\gamma(x) = \mathrm{sign}(x)\mathrm{ReLU}(|x| - \gamma)$, where $\mathrm{ReLU}(x) = \max(x, 0)$ is the rectified linear unit and $\mathrm{sign}(x)$ is the sign function. CSC provides a pipeline to extract features of an image, but its performance highly depends on the configuration of $\{\lambda, \rho, \boldsymbol{d}\}$. By the principle of algorithm unrolling [8, 44, 56], the ISTA of CSC can be generalized as a unit in neural networks. We employ two convolutional units, $\mathrm{Conv}_i (i = 0, 1)$, to replace $\boldsymbol{d}$ and $\boldsymbol{d}^T/\rho$, and proximal operator $\mathrm{prox}(\cdot)$ is extended to the activation function $f(\cdot)$. Hence, Eq.(2) can be rewritten as

$$\boldsymbol{z}^{(k+1)} = f\left(\mathrm{BN}\left(\boldsymbol{z}^{(k)} + \mathrm{Conv}_1(\boldsymbol{x} - \mathrm{Conv}_0(\boldsymbol{z}^{(k)}))\right)\right), \tag{3}$$

where we also take batch normalization (BN) into account. It is worth pointing out that, except for SST, the activation function can be freely set to alternatives (e.g., ReLU, parametric ReLU (PReLU) and so on) if the regularizer $g(\cdot)$ is not set to ℓ_1-norm. In what follows, Eq. (3) is called a dictionary convolutional unit (DCU). By stacking DCUs, the original CSC model can be represented as a deep CSC neural network.

In addition, stacking DCUs is interpretable to representation learning. Conv_0 serves as a decoder, since it maps $\boldsymbol{z}^{(k)}$ from feature space to image space. And Conv_1 serves as an encoder, since it maps the residual between the original image $\boldsymbol{x}$ and the reconstructed image $\mathrm{Conv}_0(\boldsymbol{z}^{(k)})$ from image space to feature space. Then, the encoded residual is added to the current code $\boldsymbol{z}^{(k)}$ for updating. Eventually, the output passes through BN and an activation function for non-linearity. This process can be regarded as an iterative auto-encoder.

3. CSCFuse

In this section, we apply deep CSC neural networks to the image fusion problem, and exhibit three paradigms of model formulation for three different image fusion tasks.

3.1. Infrared and Visible Image Fusion

By combining autoencoders and the CSC model, we propose a CSC-based IVF network (CSC-IVFN), which can be regarded as a flexible data-driven transformer. During the training phase, we train CSC-IVFN in an autoencoder fashion using all the available infrared and visible training images. In the testing phase, we use the well-trained encoder to obtain the base and detail features of the infrared and visible images. These features are then fused by an extra fusion layer and decoded by the trained decoder to produce the final fused images.

Training Phase. The architecture is displayed in Fig. 1 (a). Firstly, the input image $\boldsymbol{x}$[1] is decomposed into a base image $\boldsymbol{x}^{\mathrm{B}}$ containing low-frequency information and a detail image $\boldsymbol{x}^{\mathrm{D}}$ containing high-frequency textures. Similar to [15, 33], $\boldsymbol{x}^{\mathrm{B}}$ is obtained by applying a box-blur filter to $\boldsymbol{x}$, and as for the detail image there is $\boldsymbol{x}^{\mathrm{D}} = \boldsymbol{x} - \boldsymbol{x}^{\mathrm{B}}$. Then, the base and detail images pass through N stacked DCUs, and we will get the final feature maps, that is, $\boldsymbol{z}^{\mathrm{D}}$ and $\boldsymbol{z}^{\mathrm{B}}$. And next we feed them into a decoder to decode the base and detail images. Finally, they are combined to reconstruct the input image. Here, the output is activated by a sigmoid function to make sure that the values range from 0 to 1. The loss function is the mean squared error (MSE) plus structural similarity (SSIM) loss,

$$L^{\mathrm{IVF}} = \frac{1}{hw}\left(\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2^2 + \lambda^{\mathrm{IVF}}\frac{1 - \mathrm{SSIM}(\boldsymbol{x}, \hat{\boldsymbol{x}})}{2}\right), \tag{4}$$

where λ^{IVF} is a trade-off parameter to balance the MSE and SSIM [67]. Note that MSE is used to keep the spatial consistency and SSIM guarantees local details in terms of structure, contrast and brightness [67].

Testing Phase. After training a CSC-IVFN, there is a transformer (encoder) and an inverse transformer (decoder). In

[1] In the training phase, both infrared and visible images are indiscriminately denoted by $\boldsymbol{x}$.

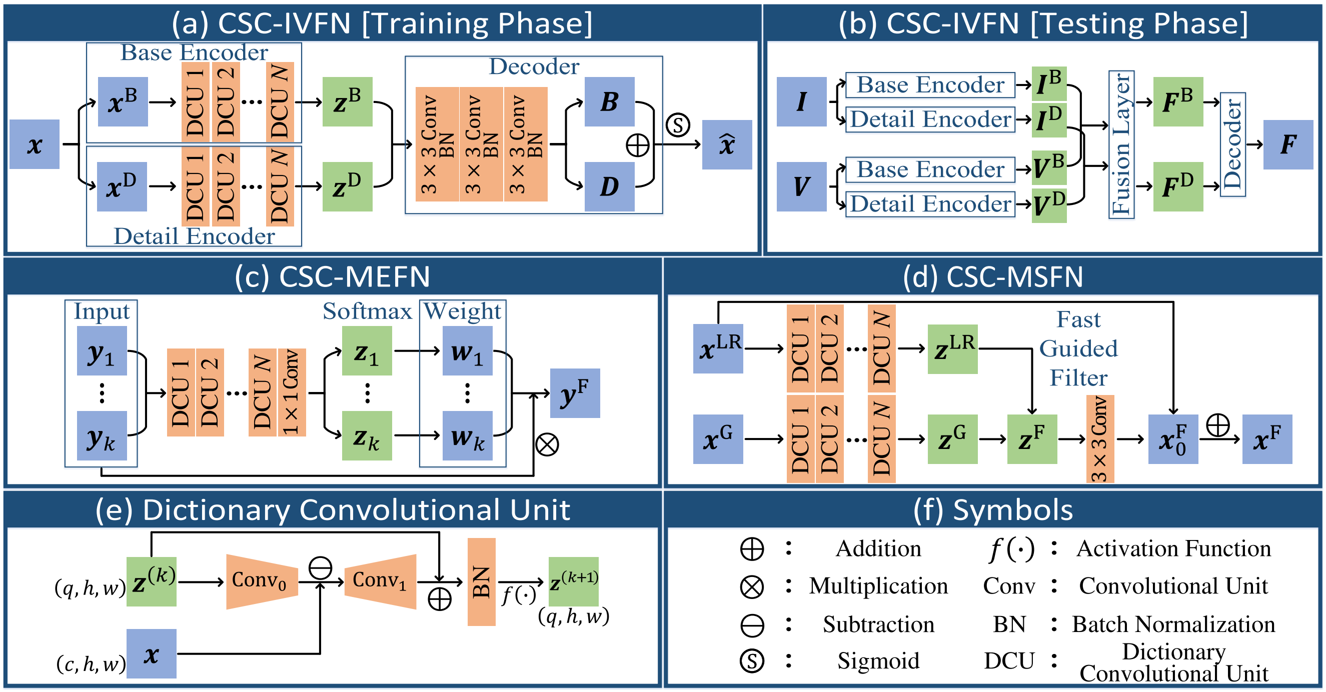
(a) CSC-IVFN [Training Phase]
Base Encoder
$\boldsymbol{x}$
$\boldsymbol{x}^{\mathrm{B}}$
DCU 1
DCU 2
DCU N
$\boldsymbol{z}^{\mathrm{B}}$
$\boldsymbol{x}^{\mathrm{D}}$
$\boldsymbol{z}^{\mathrm{D}}$
Detail Encoder
Decoder
3 × 3 Conv
BN
$\boldsymbol{B}$
$\boldsymbol{D}$
$\hat{\boldsymbol{x}}$
(b) CSC-IVFN [Testing Phase]
$\boldsymbol{I}$
$\boldsymbol{V}$
Base Encoder
Detail Encoder
$\boldsymbol{I}^{\mathrm{B}}$
$\boldsymbol{I}^{\mathrm{D}}$
$\boldsymbol{V}^{\mathrm{B}}$
$\boldsymbol{V}^{\mathrm{D}}$
Fusion Layer
$\boldsymbol{F}^{\mathrm{B}}$
$\boldsymbol{F}^{\mathrm{D}}$
Decoder
$\boldsymbol{F}$
(c) CSC-MEFN
Input
$\boldsymbol{y}_1$
$\boldsymbol{y}_k$
DCU 1
DCU 2
DCU N
1 × 1 Conv
Softmax
$\boldsymbol{z}_1$
$\boldsymbol{z}_k$
Weight
$\boldsymbol{w}_1$
$\boldsymbol{w}_k$
$\boldsymbol{y}^{\mathrm{F}}$
(d) CSC-MSFN
$\boldsymbol{x}^{\mathrm{LR}}$
$\boldsymbol{z}^{\mathrm{LR}}$
Fast Guided Filter
$\boldsymbol{x}^{\mathrm{G}}$
$\boldsymbol{z}^{\mathrm{G}}$
$\boldsymbol{z}^{\mathrm{F}}$
3 × 3 Conv
$\boldsymbol{x}_0^{\mathrm{F}}$
$\boldsymbol{x}^{\mathrm{F}}$
(e) Dictionary Convolutional Unit
(q,h,w)
$\boldsymbol{z}^{(k)}$
Conv$_0$
Conv$_1$
BN
$f(\cdot)$
$\boldsymbol{z}^{(k+1)}$
(q,h,w)
(c,h,w)
$\boldsymbol{x}$
(f) Symbols
⊕ : Addition
⊗ : Multiplication
⊖ : Subtraction
Ⓢ : Sigmoid
$f(\cdot)$: Activation Function
Conv : Convolutional Unit
BN : Batch Normalization
DCU : Dictionary Convolutional Unit

Figure 1. Network structure for our CSCFuse.

the test phase, CSC-IVFN is feed with a pair of infrared and visible images. In what follows, we use $\boldsymbol{I}^{\mathrm{B}}$, $\boldsymbol{I}^{\mathrm{D}}$, $\boldsymbol{V}^{\mathrm{B}}$ and $\boldsymbol{V}^{\mathrm{D}}$ to represent the base and detail feature maps of infrared and visible images, respectively. As exhibited in Fig. 1 (b), a fusion layer is inserted between the encoder and decoder in the test phase. It can be expressed by a unified merging operation $\mathcal{F}(\cdot)$,

$$\begin{aligned}\boldsymbol{F}^{\mathrm{B}} &= \mathcal{F}(\boldsymbol{I}^{\mathrm{B}}, \boldsymbol{V}^{\mathrm{B}}) = \boldsymbol{w}_1^{\mathrm{B}} \otimes \boldsymbol{I}^{\mathrm{B}} \oplus \boldsymbol{w}_2^{\mathrm{B}} \otimes \boldsymbol{V}^{\mathrm{B}},\\ \boldsymbol{F}^{\mathrm{D}} &= \mathcal{F}(\boldsymbol{I}^{\mathrm{D}}, \boldsymbol{V}^{\mathrm{D}}) = \boldsymbol{w}_1^{\mathrm{D}} \otimes \boldsymbol{I}^{\mathrm{D}} \oplus \boldsymbol{w}_2^{\mathrm{D}} \otimes \boldsymbol{V}^{\mathrm{D}}.\end{aligned} \tag{5}$$

Here, $\otimes$ and $\oplus$ are element-wise product and addition.

For the determination of $\{\boldsymbol{w}_1^{\mathrm{B}}, \boldsymbol{w}_2^{\mathrm{B}}\}$ and $\{\boldsymbol{w}_1^{\mathrm{D}}, \boldsymbol{w}_2^{\mathrm{D}}\}$, Here we focus on introducing the *saliency-weighted fusion strategy* [15]. Other conventional fusion strategies such as *ℓ_1-norm* and *weighted average* can refer to [18].

To highlight and retain the saliency target and information, the fusion weight of this strategy is determined by the saliency degree. We take base weights as an example. Firstly, the saliency value of $\boldsymbol{I}^{\mathrm{B}}$ at the kth pixel can be obtained by $\boldsymbol{S}_{\boldsymbol{I}}^{\mathrm{B}}(k) = \sum_{i=0}^{255} \boldsymbol{H}_{\boldsymbol{I}}^{\mathrm{B}}(i)|\boldsymbol{I}^{\mathrm{B}}(k) - i|$, where $\boldsymbol{I}^{\mathrm{B}}(k)$ is the value of the kth pixel and $\boldsymbol{H}_{\boldsymbol{I}}^{\mathrm{B}}(i)$ is the frequency of pixel value i. The initial weight at the kth pixel is $\tilde{\boldsymbol{w}}_1^{\mathrm{B}}(k) = \boldsymbol{S}_{\boldsymbol{I}}^{\mathrm{B}}(k) / \left[\boldsymbol{S}_{\boldsymbol{I}}^{\mathrm{B}}(k) + \boldsymbol{S}_{\boldsymbol{V}}^{\mathrm{B}}(k)\right]$ and $\tilde{\boldsymbol{w}}_2^{\mathrm{B}}(k) = 1 - \tilde{\boldsymbol{w}}_1^{\mathrm{B}}(k)$. To prevent region boundaries and artifacts, the weight map is refined via the guided filter $\mathcal{G}(\cdot,\cdot)$ with the guidance of base and detail feature maps:

$$\begin{aligned}&\boldsymbol{w}_1^{\mathrm{B}} = \mathcal{G}(\tilde{\boldsymbol{w}}_1^{\mathrm{B}}, \boldsymbol{I}^{\mathrm{B}}) / \left[\mathcal{G}(\tilde{\boldsymbol{w}}_1^{\mathrm{B}}, \boldsymbol{I}^{\mathrm{B}}) + \mathcal{G}(\tilde{\boldsymbol{w}}_2^{\mathrm{B}}, \boldsymbol{I}^{\mathrm{V}})\right],\\ &\boldsymbol{w}_2^{\mathrm{B}} = 1 - \boldsymbol{w}_1^{\mathrm{B}}.\end{aligned} \tag{6}$$

3.2. Multi-Exposure Image Fusion

Most of MEF algorithms fall under the umbrella of weighted summation framework, $\boldsymbol{f} = \sum_{k=1}^{K} \boldsymbol{w}_k \otimes \boldsymbol{x}_k$, where $\{\boldsymbol{x}_k\}_{k=1}^{K}$ are source images, $\{\boldsymbol{w}_k\}_{k=1}^{K}$ are the corresponding weight maps, $\boldsymbol{f}$ is the fused image and K denotes the number of exposures. We propose a CSC-based MEF network (CSC-MEFN). Different from CSC-IVFN, CSC-MEFN is an end-to-end network. Here DCUs extract feature maps, which are then used to predict weight maps to generate the fusion image. To avoid chroma distortion, the proposed CSC-MEFN works in the YCbCr space, and its channels are denoted by $\boldsymbol{y}_k, \boldsymbol{b}_k$ and $\boldsymbol{r}_k$. As shown in Fig. 1 (c), Y channels $\{\boldsymbol{y}_k\}_{k=1}^{K}$ pass through CSC-MEFN one-by-one. At first, CSC-MEFN stacks N DCUs to code the Y channels. Then, it is followed by a 1×1 convolutional unit to get the final code $\boldsymbol{z}_k$. Thereafter, the codes $\{\boldsymbol{z}_k\}_{k=1}^{K}$ are converted into weight maps $\{\boldsymbol{w}_k\}_{k=1}^{K}$ by softmax activation. At last, the fused Y channel $\boldsymbol{y}^{\mathrm{F}}$ is obtained by $\boldsymbol{y}^{\mathrm{F}} = \sum_{k=1}^{K} \boldsymbol{w}_k \otimes \boldsymbol{y}_k$. As for the Cb channels, we employ the MEF ℓ_1-norm fusion strategy, i.e., $\boldsymbol{b}^{\mathrm{F}} = \sum_{k=1}^{K} \|\boldsymbol{b}_k - 0.5\|_1 \boldsymbol{b}_k / \sum_{k=1}^{K} \|\boldsymbol{b}_k - 0.5\|_1$. So Cr channels do. After the separate fusion of three channels, the fusion image $\boldsymbol{f}$ is transformed from YCbCr to RGB space. Eventually, we apply a post-processing [20]: the values at 0.5% and 99.5% intensity levels are mapped to [0,1], and values out of this range are clipped.

CSC-MEFN is supervised by improved MEFSSIM [37]. It evaluates the similarity between source images $\{\boldsymbol{x}_k\}_{k=1}^{K}$ and the fusion image $\boldsymbol{f}$ in terms of illumination, contrast and structure. Our experimental results show that MEFSSIM often leads to haloes. Essentially, halo artifacts result from the pixel fluctuation in the illumination map (i.e., Y channel). To suppress haloes, we propose a halo loss defined by the ℓ_1-norm on gradients of the illumination map, $L_{\text{halo}} = \|\nabla \boldsymbol{y}^{\text{F}}\|_1$, where ∇ denotes the image gradient operator. In our experiments, ∇ is implemented by horizontal and vertical Sobel filters. In summary, given the penalty parameter λ^{MEF}, the loss function of CSC-MEFN is expressed by

$$L^{\text{MEF}} = \frac{1}{hw}\left(-\text{MEFSSIM} + \lambda^{\text{MEF}} L_{\text{halo}}\right). \qquad (7)$$

3.3. Multi-Spectral Image Fusion

Owing to the limitation of multi-spectral imaging devices, multi-spectral images (MS) contain enriched spectral information but with low resolution (LR). One of the promising techniques for acquiring a high-resolution (HR) MS is to fuse the LRMS with a guidance image (e.g. panchromatic or RGB images). This problem is a special MSF task. We present a CSC-based MSF network (CSC-MSFN) for the general MSF task. It is assumed that LR and guidance images are represented by $\boldsymbol{x}^{\text{LR}} = \boldsymbol{d}^{\text{LR}} * \boldsymbol{z}^{\text{LR}}$ and $\boldsymbol{x}^{\text{G}} = \boldsymbol{d}^{\text{G}} * \boldsymbol{z}^{\text{G}}$, respectively. Given the dictionary of HR images $\boldsymbol{d}^{\text{F}}$, the HR image is represented by

$$\boldsymbol{x}^{\text{F}} = \boldsymbol{d}^{\text{F}} * (\boldsymbol{z}^{\text{LR}}) \uparrow. \qquad (8)$$

The symbol $\uparrow$ denotes the upsampling operator. According to this model, CSC-MSFN separately extracts codes of $\boldsymbol{x}^{\text{LR}}$ and $\boldsymbol{x}^{\text{G}}$ by two sequences of DCUs, and we utilize the fast guidance filter to super-resolve $\boldsymbol{z}^{\text{LR}}$ with the guidance of $\boldsymbol{z}^{\text{G}}$. At last, the HR image is recovered by a 3×3 convolutional unit. The loss function is set to MSE between ground truth and fusion images.

4. Experiments

Here we elaborate the implementation and configuration details of our networks. Experiments are conducted to show the performance of our models and the rationality of network structures. For each task, our experiments utilized training, validation and test datasets. The hyperparameters are determined by validation set.

4.1. Infrared and Visible Image Fusion

Table 1. Quantitative results of the IVF task. Boldface and underline indicate the best and the second best results, respectively.

Dataset: FLIR

	DeepF	DenseF	DLF	FEZL	FGAN	SDF	TVAL	Ours
EN	$\underline{7.21}$	7.21	6.99	6.91	7.02	7.15	6.80	**7.61**
MI	2.73	2.73	$\underline{2.78}$	2.78	2.68	2.31	2.47	**3.02**
SD	$\underline{37.35}$	37.32	32.58	31.16	34.38	35.89	28.07	**55.94**
SF	15.47	15.50	14.52	14.16	11.51	$\underline{18.79}$	14.04	**21.85**
VIF	0.50	0.50	0.42	0.33	0.29	$\underline{0.50}$	0.33	**0.70**
AG	4.80	4.82	4.15	3.38	3.20	$\underline{5.57}$	3.52	**6.92**
SCD	1.72	$\underline{1.72}$	1.57	1.42	1.18	1.50	1.40	**1.80**

Dataset: NIR-Country Scene

	DeepF	DenseF	DLF	FEZL	FGAN	SDF	TVAL	Ours
EN	7.30	$\underline{7.30}$	7.22	7.19	7.06	7.30	7.13	**7.36**
MI	$\underline{4.04}$	**4.04**	3.97	3.81	3.00	3.29	3.67	3.86
SD	45.82	$\underline{45.85}$	42.31	44.44	34.91	43.74	40.47	**69.37**
SF	18.63	18.72	18.36	17.04	14.31	$\underline{20.65}$	16.69	**28.29**
VIF	0.68	0.68	0.61	0.55	0.42	$\underline{0.69}$	0.53	**1.05**
AG	6.18	6.23	5.92	5.38	4.56	$\underline{6.82}$	5.32	**9.42**
SCD	1.37	$\underline{1.37}$	1.22	1.14	0.51	1.19	1.09	**1.73**

Dataset: TNO

	DeepF	DenseF	DLF	FEZL	FGAN	SDF	TVAL	Ours
EN	$\underline{6.86}$	6.84	6.38	6.63	6.58	6.67	6.40	**6.91**
MI	2.30	2.30	2.15	2.23	$\underline{2.34}$	1.72	2.04	**2.50**
SD	$\underline{32.25}$	31.82	22.94	28.05	29.04	28.04	23.01	**46.97**
SF	11.13	11.09	9.80	9.46	8.76	$\underline{12.60}$	9.03	**12.88**
VIF	$\underline{0.58}$	0.57	0.31	0.31	0.26	0.46	0.28	**0.62**
AG	3.60	3.60	2.72	2.55	2.42	$\underline{3.98}$	2.52	**4.22**
SCD	**1.80**	$\underline{1.80}$	1.62	1.67	1.40	1.68	1.60	1.70

Table 2. Quantitative results of the MEF task. Boldface and underline indicate the best and the second best results, respectively.

	MEON	Brisque	Niqe	Piqe
EF	$\underline{8.67}$	18.83	2.91	31.06
GGIF	9.15	19.17	2.52	32.19
DenseFuse	11.85	26.44	2.58	$\underline{29.61}$
MEF-Net	9.36	19.45	2.52	32.29
FMMR	9.86	20.11	2.55	32.09
DSIFTEF	9.38	18.65	2.53	32.29
Lee18	9.81	$\underline{18.51}$	$\underline{2.46}$	32.54
Ours	**8.17**	**18.26**	**2.39**	**27.83**

Datasets, Metrics and Details. IVF experiments use three datasets (FLIR, NIR and TNO). The 180 pairs of images in FLIR compose the training set. Two subsets, *Water* (51 pairs) and *OldBuilding* (51 pairs) of NIR, are used for validation. To comprehensively evaluate the performance of different models, we employ TNO (40 pairs), NIR-Country (52 pairs) and the rest pairs of FLIR (40 pairs) as test datasets. To quantitatively measure the fusion performance, seven metrics are employed: entropy (EN), mutual information (MI), standard deviation (SD), spatial frequency (SF), visual information fidelity (VIF), average gradient (AG), and sum of the correlations of differences (SCD). Larger metrics indicate that a fusion image is better. In our experiment, the tuning parameter λ^{IVF} in Eq. (4) is set to 5. The network is optimized over 60 epochs with a learning rate of 10^{-2} in the first 30 epochs and 10^{-3} in the rest epochs. The number of DCUs, activation function and fusion strategy are reported as follows: the number of DCUs in base or detail encoder is 7; the activation functions in base and detail encoders are set as PReLU and SST, respectively; the fusion strategies for base and detail images are saliency-weighted fusion and ℓ_1-norm fusion [18], respectively. Selection of

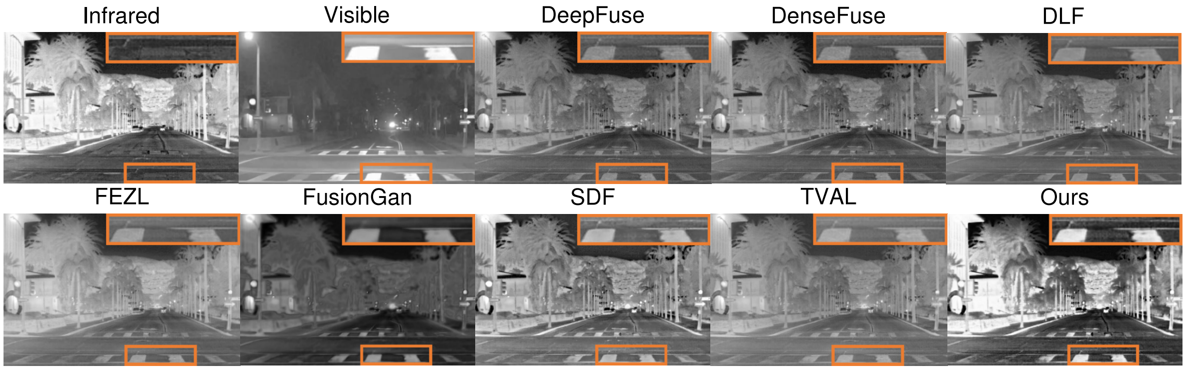

Figure 2. Visually fusion results for our CSCFuse *vs*. SOTA methods.

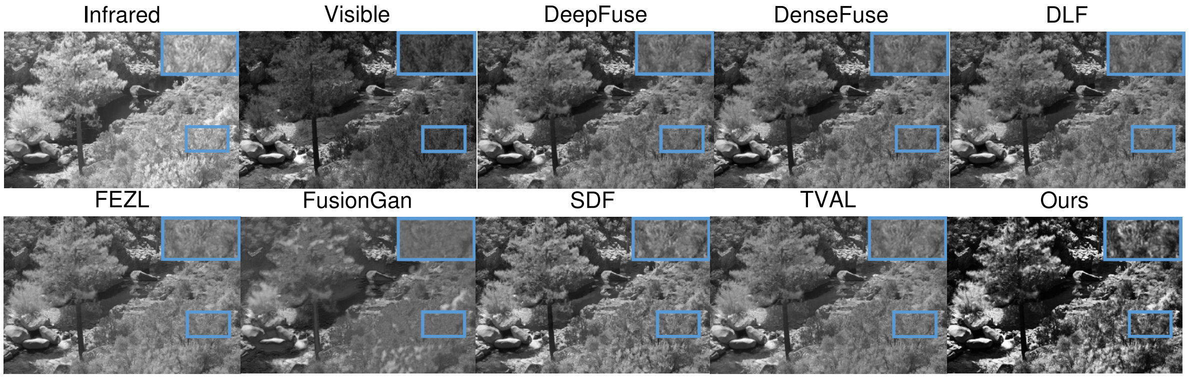

Figure 3. Visually fusion results for our CSCFuse *vs*. SOTA methods.

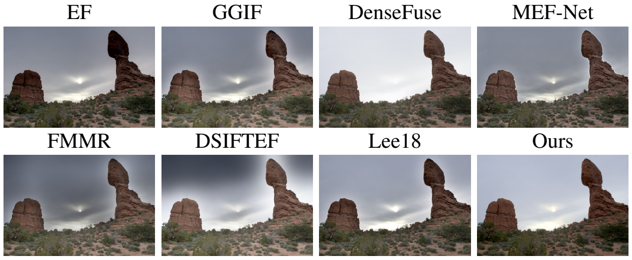

Figure 4. The fused images of *Balanced Rock*.

the above configurations was made by grid search on the validation set.

Comparison with SOTA Methods. To verify the superiority of our CSC-IVFN, we compare its fusion results with nine popular IVIF fusion methods, including ADKT [1], CSR [33], DeepFuse [47], DenseFuse [18], DLF [17], FEZL [15], FusionGAN [36], SDF [2] and TVAL [9]. Six metrics of all methods are displayed in Table 1. It is shown that our method achieves the best performance on all test sets with regard to most metrics. Therefore, our method is suitable for various scenarios with different kinds of illuminations and object categories. In contrast, the other methods (including DeepFuse, DenseFuse and SDF) can achieve good performance on certain test sets concerning a part of metrics. Besides the metric comparison, representative fusion images are displayed in Figs. 2 and 3. In the visible image, there are lots of bushes. In the infrared image, we can observe a bunker. However, it is not easy to recognize the bushes/bunker in the infrared/visible image. It is found that our fusion image keeps the details and textures of the visible image, and preserves the interest objects (i.e., the bushes and the bunker). In addition, its contrast is fairly high. In conclusion, both visible spectrum and thermal radiation information are retained in our fusion image. However, other methods cannot generate satisfactory images as good as ours.

4.2. Multi-Exposure Image Fusion

Datasets, Metrics and Details. Three datasets SICE [3], TCI2018 [37] and HDRPS[2] are employed in our experiments. HDRPS and TCI2018 are used for test and validation, respectively. SICE is a large and high-quality dataset. It is divided into two parts for training and validation. Many papers use MEFSSIM to evaluate the performance, but CSC-MEFN is supervised by MEFSSIM. Hence, it is unfair for other methods. As an alternative, we utilize four SOTA blind image quality indices , i.e., blind/referenceless image spatial quality evaluator (Brisque) [42], naturalness image quality evaluator (Niqe) [43], perception based image quality evaluator (Piqe) [45] and multi-task end-to-end optimized deep neural network (MEON) based blind image

[2] http://markfairchild.org/HDR.html

Table 3. Quantitative results of the MSF task. Boldface and underline indicate the best and the second best results, respectively.

Images	CNMF		GSA		FUSE		MAPSMM		GLPHS		PNN		PFCN		Ours	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
RF apples	34.57	0.94	32.73	0.68	38.25	0.94	41.44	0.98	43.55	0.98	39.93	0.97	41.60	0.99	**51.58**	**1.00**
RF peppers	33.13	0.93	30.96	0.70	35.77	0.92	39.56	0.97	41.60	0.98	39.48	0.97	40.47	0.98	**49.54**	**0.99**
Sponges	31.14	0.95	26.31	0.74	33.76	0.94	35.25	0.93	37.29	0.97	31.39	0.96	32.03	0.98	**43.29**	**0.99**
Stuffed toys	30.04	0.87	27.33	0.58	34.30	0.94	36.46	0.94	38.39	0.97	33.67	0.96	33.10	0.97	**44.11**	**0.99**
Superballs	21.29	0.83	32.53	0.76	36.36	0.91	27.56	0.60	39.31	0.95	36.99	0.95	36.74	0.97	**46.29**	**0.99**
Thread spools	32.37	0.89	30.66	0.66	33.96	0.91	34.92	0.94	36.36	0.96	35.81	0.95	38.82	0.98	**42.55**	**0.99**
Mean	30.42	0.90	30.09	0.69	35.40	0.93	35.87	0.89	39.42	0.97	36.21	0.96	37.13	0.98	**46.23**	**0.99**

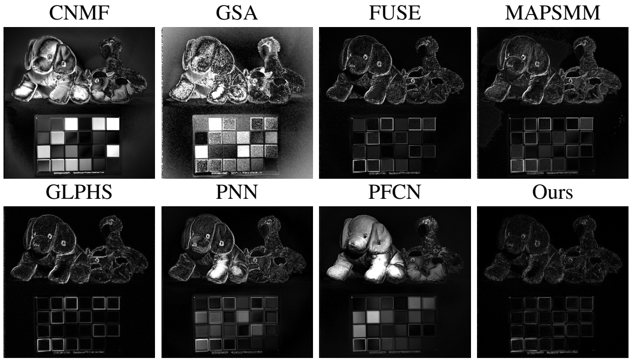

Figure 5. The error maps of *stuffed toys* (band 3). Their values are amplified 10 times for easier visual inspection. The error goes larger from black to white.

quality assessment [39]. Smaller values indicate that a fusion image is better. Experiments show that large λ^{MEF} makes training unstable, so at the ith iteration it is set to $\min\{0.25(i-1), \lambda^{\mathrm{MEF}}_{\max}\}$. We select $\lambda^{\mathrm{MEF}}_{\max} = 10$ to make halo loss and MEFSSIM loss have similar magnitudes. The network is optimized by Adam over 50 epochs with a learning rate of 5×10^{-4}. The network configuration is determined by validation datasets. We utilize $N = 3$ DCUs to extract codes and SST is employed as an activation function.

Comparison with SOTA Methods. CSC-MEFN is compared with seven classic and recent SOTA methods, including EF [41], GGIF [14], DenseFuse [18], MEF-Net [38], FMMR [19], DSIFTEF [34], Lee18 [16]. The metrics are listed in Table 2. Our network outperforms other methods. Lee18 and EF are ranked in the second and third places. Fig. 4 displays the fusion images. It is shown that GGIF, MEF-Net, FMMR, DSIFTEF and Lee18 suffer from strong halo effects around edges between the sky and rocks. For EF the right rock is too dark, and for DenseFuse the sun cannot be recognized. The contrast of local regions for both EF and DenseFuse is low. Our fusion image strikes the balance.

4.3. Multi-Spectral Image Fusion

Datasets, Metrics and Details. We employ a multi-spectral/RGB image fusion dataset, Cave [79]. It contains 32 scenes, each of which has a 31-band multi-spectral image and an RGB image. It is divided into three parts for training, testing and validation. The Wald protocol is used to construct training sets. We employ peak signal-to-noise ratio (PSNR) and SSIM as evaluation indexes. Larger PSNR and SSIM indicate that a fusion image is better. The network is optimized by Adam over 100 epochs with a learning rate of 5×10^{-4}. SST is employed as an activation function. The number of DCUs is empirically set to 4 for a speed and accuracy trade-off.

Comparison with SOTA Methods. CSC-MSFN is compared with seven classic and recent SOTA methods, including CNMF [81], GSA [60], FUSE [68], MAPSMM [6], GLPHS [55], PNN [40] and PFCN [91]. The metrics listed in Table 3 show that our network achieves the largest PSNR and SSIM. GLPHS and PFCN can be ranked in the second place in terms of PSNR and SSIM, respectively. The error maps of the third band of *stuffed toys* are displayed in Fig. 5. We found that CNMF, GSA and PFCN break down when reconstructing the color checkerboard and stuffed toys, while FUSE, MAPSMM, GLPHS and PNN perform badly at the edges. In summary, CSC-MSFN has the best performance.

5. Conclusion

Inspired by converting the ISTA and CSC models into a hidden layer of neural networks, this paper proposes three deep CSC networks for IVF, MEF and MSF tasks. Extensive experiments and comprehensive comparisons demonstrate that our networks outperform the SOTA methods. Furthermore, numerous experiments show that our networks are highly reproducible.

Acknowledgement

This work is supported by the National Key Research and Development Program of China under grant 2018AAA0102201, the National Natural Science Foundation of China under Grant 61976174, 12201497, 12201490 and 62276208, the Guangdong Basic and Applied Basic Research Foundation under Grant 2023A1515011358, the Fundamental Research Funds for the Central Universities under Grant D5000220060 and xhj032021013-03, the China Postdoctoral Science Foundation funded project under Grant 2021M702621.

References

[1] Durga Prasad Bavirisetti and Ravindra Dhuli. Fusion of infrared and visible sensor images based on anisotropic diffusion and karhunen-loeve transform. *IEEE Sensors Journal*, 16(1):203–209, 2015. 5

[2] Durga Prasad Bavirisetti and Ravindra Dhuli. Two-scale image fusion of visible and infrared images using saliency detection. *Infrared Phys. & Techn.*, 76:52–64, 2016. 5

[3] Jianrui Cai, Shuhang Gu, and Lei Zhang. Learning a deep single image contrast enhancer from multi-exposure images. *IEEE TIP*, 27(4):2049–2062, 2018. 1, 5

[4] Xin Deng and Pier Luigi Dragotti. Deep convolutional neural network for multi-modal image restoration and fusion. *IEEE TPAMI*, 43(10):3333–3348, 2021. 1

[5] Jiang Dong, Dafang Zhuang, Yaohuan Huang, and Jingying Fu. Advances in multi-sensor data fusion: Algorithms and applications. *Sensors*, 9(10):7771–7784, 2009. 1

[6] Michael T. Eismann and Russell C. Hardie. Hyperspectral resolution enhancement using high-resolution multispectral imagery with arbitrary response functions. *IEEE TGRS*, 43(3):455–465, 2005. 6

[7] Fangyuan Gao, Xin Deng, Mai Xu, Jingyi Xu, and Pier Luigi Dragotti. Multi-modal convolutional dictionary learning. *IEEE TIP*, 31:1325–1339, 2022. 1

[8] Karol Gregor and Yann LeCun. Learning fast approximations of sparse coding. In *ICML, Haifa, Israel, June 21-24*, pages 399–406, 2010. 2

[9] Hanqi Guo, Yong Ma, Xiaoguang Mei, and Jiayi Ma. Infrared and visible image fusion based on total variation and augmented lagrangian. *J. Opt. Soc. Am. A*, 34(11):1961–1968, 2017. 5

[10] Xue-mei Hu, Felix Heide, Qionghai Dai, and Gordon Wetzstein. Convolutional sparse coding for RGB+NIR imaging. *IEEE TIP*, 27(4):1611–1625, 2018. 1

[11] Zhanbo Huang, Jinyuan Liu, Xin Fan, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Reconet: Recurrent correction network for fast and efficient multi-modality image fusion. In *ECCV*, pages 539–555. Springer, 2022. 1

[12] Zhiying Jiang, Zhuoxiao Li, Shuzhou Yang, Xin Fan, and Risheng Liu. Target oriented perceptual adversarial fusion network for underwater image enhancement. *IEEE TCSVT*, 32(10):6584–6598, 2022. 1

[13] Zhiying Jiang, Zengxi Zhang, Xin Fan, and Risheng Liu. Towards all weather and unobstructed multi-spectral image stitching: Algorithm and benchmark. In *Proceedings of the 30th ACM MM*, pages 3783–3791, 2022. 1

[14] Fei Kou, Zhengguo Li, Changyun Wen, and Weihai Chen. Multi-scale exposure fusion via gradient domain guided image filtering. In *ICME, Hong Kong, China, July 10-14*, pages 1105–1110. IEEE Computer Society, 2017. 6

[15] Fayez Lahoud and Sabine Süsstrunk. Fast and efficient zero-learning image fusion. *CoRR*, abs/1905.03590, 2019. 2, 3, 5

[16] S. Lee, J. S. Park, and N. I. Cho. A multi-exposure image fusion based on the adaptive weights reflecting the relative pixel intensity and global gradient. In *ICIP, Athens, Greece, Oct. 7-10*, pages 1737–1741, 2018. 6

[17] Hui Li, Xiao-Jun Wu, and Josef Kittler. Infrared and visible image fusion using a deep learning framework. In *ICPR, Beijing, China, August 20-24, 2018*, pages 2705–2710. IEEE Computer Society, 2018. 5

[18] Hui Li and Xiao-Jun Wu. Densefuse: A fusion approach to infrared and visible images. *IEEE TIP*, 28(5):2614–2623, 2018. 1, 3, 4, 5, 6

[19] Shutao Li and Xudong Kang. Fast multi-exposure image fusion with median filter and recursive filter. *IEEE Trans. Consumer Electronics*, 58(2):626–632, 2012. 6

[20] Zhetong Liang, Jun Xu, David Zhang, Zisheng Cao, and Lei Zhang. A hybrid l1-l0 layer decomposition model for tone mapping. In *CVPR, Salt Lake City, UT, USA, June 18-22*, pages 4758–4766, 2018. 3

[21] Aishan Liu, Xianglong Liu, Jiaxin Fan, Yuqing Ma, Anlan Zhang, Huiyuan Xie, and Dacheng Tao. Perceptual-sensitive gan for generating adversarial patches. In *33rd AAAI Conference on Artificial Intelligence*, 2019. 1

[22] Aishan Liu, Xianglong Liu, Hang Yu, Chongzhi Zhang, Qiang Liu, and Dacheng Tao. Training robust deep neural networks via adversarial noise propagation. *IEEE TIP*, 2021. 1

[23] Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *CVPR*, pages 5802–5811, 2022. 1

[24] Jinyuan Liu, Xin Fan, Ji Jiang, Risheng Liu, and Zhongxuan Luo. Learning a deep multi-scale feature ensemble and an edge-attention guidance for image fusion. *IEEE TCSVT*, 32(1):105–119, 2021. 1

[25] Jinyuan Liu, Jingjie Shang, Risheng Liu, and Xin Fan. Attention-guided global-local adversarial learning for detail-preserving multi-exposure image fusion. *IEEE TCSVT*, 32(8):5026–5040, 2022. 1

[26] Jinyuan Liu, Guanyao Wu, Junsheng Luan, Zhiying Jiang, Risheng Liu, and Xin Fan. Holoco: Holistic and local contrastive learning network for multi-exposure image fusion. *Information Fusion*, 2023. 1

[27] Jinyuan Liu, Yuhui Wu, Zhanbo Huang, Risheng Liu, and Xin Fan. Smoa: Searching a modality-oriented architecture for infrared and visible image fusion. *IEEE Signal Processing Letters*, 28:1818–1822, 2021. 1

[28] Risheng Liu, Zhiying Jiang, Shuzhou Yang, and Xin Fan. Twin adversarial contrastive learning for underwater image enhancement and beyond. *IEEE TIP*, 31:4922–4936, 2022. 1

[29] Risheng Liu, Jinyuan Liu, Zhiying Jiang, Xin Fan, and Zhongxuan Luo. A bilevel integrated model with data-driven layer ensemble for multi-modality image fusion. *IEEE TIP*, 30:1261–1274, 2020. 1

[30] Risheng Liu, Zhu Liu, Jinyuan Liu, and Xin Fan. Searching a hierarchically aggregated fusion architecture for fast multi-modality image fusion. In *Proceedings of the 29th ACM MM*, pages 1600–1608, 2021. 1

[31] Yu Liu, Xun Chen, Hu Peng, and Zengfu Wang. Multi-focus image fusion with a deep convolutional neural network. *Inf. Fusion*, 36:191–207, 2017. 1

[32] Yu Liu, Xun Chen, Zengfu Wang, Z. Jane Wang, Rabab K. Ward, and Xuesong Wang. Deep learning for pixel-level image fusion: Recent advances and future prospects. *Inf. Fusion*, 42:158–173, 2018. 1

[33] Yu Liu, Xun Chen, Rabab K Ward, and Z Jane Wang. Image fusion with convolutional sparse representation. *IEEE SPL*, 23(12):1882–1886, 2016. 2, 5

[34] Yu Liu and Zengfu Wang. Dense SIFT for ghost-free multi-exposure fusion. *J. Vis. Commun. Image Represent.*, 31:208–224, 2015. 6

[35] Jiayi Ma, Yong Ma, and Chang Li. Infrared and visible image fusion methods and applications: A survey. *Inf. Fusion*, 45:153–178, 2019. 1

[36] Jiayi Ma, Wei Yu, Pengwei Liang, Chang Li, and Junjun Jiang. Fusiongan: A generative adversarial network for infrared and visible image fusion. *Inf. Fusion*, 48:11–26, 2019. 1, 5

[37] Kede Ma, Zhengfang Duanmu, Hojatollah Yeganeh, and Zhou Wang. Multi-exposure image fusion by optimizing A structural similarity index. *IEEE TCI*, 4(1):60–72, 2018. 1, 3, 5

[38] Kede Ma, Zhengfang Duanmu, Hanwei Zhu, Yuming Fang, and Zhou Wang. Deep guided learning for fast multi-exposure image fusion. *IEEE TIP*, 29:2808–2819, 2020. 1, 6

[39] Kede Ma, Wentao Liu, Kai Zhang, Zhengfang Duanmu, Zhou Wang, and Wangmeng Zuo. End-to-end blind image quality assessment using deep neural networks. *IEEE TIP*, 27(3):1202–1213, 2018. 6

[40] Giuseppe Masi, Davide Cozzolino, Luisa Verdoliva, and Giuseppe Scarpa. Pansharpening by convolutional neural networks. *Remote Sensing*, 8(7), 2016. 6

[41] Tom Mertens, Jan Kautz, and Frank Van Reeth. Exposure fusion: A simple and practical alternative to high dynamic range photography. *Comput. Graph. Forum*, 28(1):161–171, 2009. 6

[42] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE TIP*, 21(12):4695–4708, 2012. 5

[43] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a ”completely blind” image quality analyzer. *IEEE SPL*, 20(3):209–212, 2013. 5

[44] Vishal Monga, Yuelong Li, and Yonina C. Eldar. Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing. *CoRR*, abs/1912.10557, 2019. 2

[45] Venkatanath N., Praneeth D., Maruthi Chandrasekhar Bh., Sumohana S. Channappayya, and Swarup S. Medasani. Blind image quality evaluation using perception based features. In *NCC, Mumbai, India, February 27 - March 1*, pages 1–6. IEEE, 2015. 5

[46] Zhihong Pan, Baopu Li, Dongliang He, Mingde Yao, Wenhao Wu, Tianwei Lin, Xin Li, and Errui Ding. Towards bidirectional arbitrary image rescaling: Joint optimization and cycle idempotence. In *CVPR*, pages 17389–17398, 2022. 1

[47] K Ram Prabhakar, V Sai Srikar, and R Venkatesh Babu. Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs. In *ICCV*, pages 4724–4732, 2017. 5

[48] Haotong Qin, Zhongang Cai, Mingyuan Zhang, Yifu Ding, Haiyu Zhao, Shuai Yi, Xianglong Liu, and Hao Su. Bipointnet: Binary neural network for point clouds. In *ICLR*, 2021. 1

[49] Haotong Qin, Yifu Ding, Mingyuan Zhang, YAN Qinghua, Aishan Liu, Qingqing Dang, Ziwei Liu, and Xianglong Liu. Bibert: Accurate fully binarized bert. In *ICLR*, 2022. 1

[50] Haotong Qin, Ruihao Gong, Xianglong Liu, Xiao Bai, Jingkuan Song, and Nicu Sebe. Binary neural networks: A survey. *Pattern Recognition*, 105:107281, 2020. 1

[51] Haotong Qin, Ruihao Gong, Xianglong Liu, Mingzhu Shen, Ziran Wei, Fengwei Yu, and Jingkuan Song. Forward and backward information retention for accurate binary neural networks. In *CVPR*, pages 2250–2259, 2020. 1

[52] Haotong Qin, Xudong Ma, Yifu Ding, Xiaoyang Li, Yang Zhang, Zejun Ma, Jiakai Wang, Jie Luo, and Xianglong Liu. Bifsmnv2: Pushing binary neural networks for keyword spotting to real-network performance. *IEEE TNNLS*, 2023. 1

[53] Haotong Qin, Xudong Ma, Yifu Ding, Xiaoyang Li, Yang Zhang, Yao Tian, Zejun Ma, Jie Luo, and Xianglong Liu. Bifsmn: Binary neural network for keyword spotting. 1

[54] Haotong Qin, Xiangguo Zhang, Ruihao Gong, Yifu Ding, Yi Xu, and Xianglong Liu. Distribution-sensitive information retention for accurate binary neural network. *International Journal of Computer Vision*, pages 1–22, 2022. 1

[55] M. Selva, B. Aiazzi, F. Butera, L. Chiarantini, and S. Baronti. Hyper-sharpening: A first approach on SIM-GA data. *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, 8(6):3008–3024, 2015. 6

[56] Hillel Sreter and Raja Giryes. Learned convolutional sparse coding. In *ICASSP, Calgary, AB, Canada, April 15-20*, pages 2191–2195, 2018. 2

[57] Jeremias Sulam, Vardan Papyan, Yaniv Romano, and Michael Elad. Multilayer convolutional sparse modeling: Pursuit and dictionary learning. *IEEE TSP*, 66(15):4090–4104, 2018. 1

[58] Linfeng Tang, Yuxin Deng, Yong Ma, Jun Huang, and Jiayi Ma. Superfusion: A versatile image registration and fusion network with semantic awareness. *IEEE/CAA Journal of Automatica Sinica*, 9(12):2121–2137, 2022. 1

[59] Shiyu Tang, Ruihao Gong, Yan Wang, Aishan Liu, Jiakai Wang, Xinyun Chen, Fengwei Yu, Xianglong Liu, Dawn Song, Alan Yuille, Philip H.S. Torr, and Dacheng Tao. Robustart: Benchmarking robustness on architecture design and training techniques. *arXiv: 2109.05211*, 2021. 1

[60] G. Vivone, L. Alparone, J. Chanussot, M. Dalla Mura, A. Garzelli, G. A. Licciardi, R. Restaino, and L. Wald. A critical comparison among pansharpening algorithms. *IEEE TGRS*, 53(5):2565–2586, 2015. 6

[61] Jiakai Wang. Adversarial examples in physical world. In *IJCAI*, pages 4925–4926, 2021. 1

[62] Jiakai Wang, Aishan Liu, Xiao Bai, and Xianglong Liu. Universal adversarial patch attack for automatic checkout using perceptual and attentional bias. *IEEE TIP*, 31:598–611, 2021. 1

[63] Jiakai Wang, Aishan Liu, Zixin Yin, Shunchang Liu, Shiyu Tang, and Xianglong Liu. Dual attention suppression attack: Generate adversarial camouflage in physical world. In *IEEE CVPR*, 2021. 1

[64] Jiakai Wang, Aishan Liu, Zixin Yin, Shunchang Liu, Shiyu Tang, and Xianglong Liu. Dual attention suppression attack: Generate adversarial camouflage in physical world. In *CVPR*, pages 8565–8574, 2021. 1

[65] Jiakai Wang, Zixin Yin, Pengfei Hu, Aishan Liu, Renshuai Tao, Haotong Qin, Xianglong Liu, and Dacheng Tao. Defensive patches for robust recognition in the physical world. In *CVPR*, pages 2456–2465, 2022. 1

[66] Yuxuan Wang, Jiakai Wang, Zixin Yin, Ruihao Gong, Jingyi Wang, Aishan Liu, and Xianglong Liu. Generating transferable adversarial examples against vision transformers. In *ACM MM*, pages 5181–5190, 2022. 1

[67] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004. 2

[68] Q. Wei, N. Dobigeon, and J. Tourneret. Fast fusion of multi-band images based on solving a sylvester equation. *IEEE TIP*, 24(11):4109–4121, 2015. 6

[69] Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. U2fusion: A unified unsupervised image fusion network. *IEEE TPAMI*, 44(1):502–518, 2022. 1

[70] Han Xu, Jiayi Ma, Jiteng Yuan, Zhuliang Le, and Wei Liu. Rfnet: Unsupervised network for mutually reinforcing multi-modal image registration and fusion. In *CVPR*, pages 19647–19656, 2022. 1

[71] Han Xu, Xinya Wang, and Jiayi Ma. Drf: Disentangled representation for visible and infrared image fusion. *IEEE TIM*, 70:1–13, 2021. 1

[72] Han Xu, Hao Zhang, and Jiayi Ma. Classification saliency-based rule for visible and infrared image fusion. *IEEE TCI*. 1

[73] Ruikang Xu, Mingde Yao, Chang Chen, Lizhi Wang, and Zhiwei Xiong. Continuous spectral reconstruction from rgb images via implicit neural representation. In *ECCV 2022 Workshops*, pages 78–94. Springer, 2023. 1

[74] Xiang Yan, Syed Zulqarnain Gilani, Hanlin Qin, and Ajmal Mian. Unsupervised deep multi-focus image fusion. *CoRR*, abs/1806.07272, 2018. 1

[75] Zizheng Yang, Mingde Yao, Jie Huang, Man Zhou, and Feng Zhao. Sir-former: Stereo image restoration using transformer. In *Proceedings of the 30th ACM MM*, pages 6377–6385, 2022. 1

[76] Mingde Yao, Dongliang He, Xin Li, Fu Li, and Zhiwei Xiong. Towards interactive self-supervised denoising. *IEEE TCSVT*, 2023. 1

[77] Mingde Yao, Dongliang He, Xin Li, Zhihong Pan, and Zhiwei Xiong. Bidirectional translation between uhd-hdr and hd-sdr videos. *IEEE TMM*, 2023. 1

[78] Mingde Yao, Zhiwei Xiong, Lizhi Wang, Dong Liu, and Xuejin Chen. Spectral-depth imaging with deep learning based reconstruction. *Optics express*, 27(26):38312–38325, 2019. 1

[79] F. Yasuma, T. Mitsunaga, D. Iso, and S.K. Nayar. Generalized Assorted Pixel Camera: Post-Capture Control of Resolution, Dynamic Range and Spectrum. Technical report, Nov 2008. 6

[80] Zixin Yin, Jiakai Wang, Yifu Ding, Yisong Xiao, Jun Guo, Renshuai Tao, and Haotong Qin. Improving generalization of deepfake detection with domain adaptive batch normalization. In *Proceedings of the 1st International Workshop on Adversarial Learning for Multimedia*, pages 21–27, 2021. 1

[81] N. Yokoya, T. Yairi, and A. Iwasaki. Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion. *IEEE TGRS*, 50(2):528–537, 2012. 6

[82] Xiangguo Zhang, Haotong Qin, Yifu Ding, Ruihao Gong, Qinghua Yan, Renshuai Tao, Yuhang Li, Fengwei Yu, and Xianglong Liu. Diversifying sample generation for accurate data-free quantization. In *CVPR*, pages 15658–15667, 2021. 1

[83] Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Shuang Xu, Zudi Lin, Radu Timofte, and Luc Van Gool. Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. *CoRR*, abs/2211.14461, 2022. 1

[84] Zixiang Zhao, Haowen Bai, Yuanzhi Zhu, Jiangshe Zhang, Shuang Xu, Yulun Zhang, Kai Zhang, Deyu Meng, Radu Timofte, and Luc Van Gool. DDFM: denoising diffusion model for multi-modality image fusion. *CoRR*, abs/2303.06840, 2023. 1

[85] Zixiang Zhao, Shuang Xu, Chunxia Zhang, Junmin Liu, and Jiangshe Zhang. Bayesian fusion for infrared and visible images. *Signal Process.*, 177:107734, 2020. 1

[86] Zixiang Zhao, Shuang Xu, Chunxia Zhang, Junmin Liu, Jiangshe Zhang, and Pengfei Li. DIDFuse: Deep image decomposition for infrared and visible image fusion. In *IJCAI*, pages 970–976, 2020. 1

[87] Zixiang Zhao, Shuang Xu, Jiangshe Zhang, Chengyang Liang, Chunxia Zhang, and Junmin Liu. Efficient and model-based infrared and visible image fusion via algorithm unrolling. *IEEE TCSVT*, 32(3):1186–1196, 2022. 1

[88] Zixiang Zhao, Jiangshe Zhang, Xiang Gu, Chengli Tan, Shuang Xu, Yulun Zhang, Radu Timofte, and Luc Van Gool. Spherical space feature decomposition for guided depth map super-resolution. *CoRR*, abs/2303.08942, 2023. 1

[89] Zixiang Zhao, Jiangshe Zhang, Shuang Xu, Zudi Lin, and Hanspeter Pfister. Discrete cosine transform network for guided depth map super-resolution. In *CVPR*, pages 5697–5707, June 2022. 1

[90] Zixiang Zhao, Jiangshe Zhang, Shuang Xu, Kai Sun, Lu Huang, Junmin Liu, and Chunxia Zhang. FGF-GAN: A lightweight generative adversarial network for pansharpening via fast guided filter. In *ICME*, pages 1–6, 2021. 1

[91] F. Zhou, R. Hang, Q. Liu, and X. Yuan. Pyramid fully convolutional network for hyperspectral and multispectral image fusion. *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, 12(5):1549–1558, 2019. 6