Fast and Efficient Zero-Learning Image Fusion

Fayez Lahoud, *Student Member, IEEE,* Sabine Süsstrunk, *Fellow, IEEE*

***Abstract*—We propose a real-time image fusion method using pre-trained neural networks. Our method generates a single image containing features from multiple sources. We first decompose images into a base layer representing large scale intensity variations, and a detail layer containing small scale changes. We use visual saliency to fuse the base layers, and deep feature maps extracted from a pre-trained neural network to fuse the detail layers. We conduct ablation studies to analyze our method's parameters such as decomposition filters, weight construction methods, and network depth and architecture. Then, we validate its effectiveness and speed on thermal, medical, and multi-focus fusion. We also apply it to multiple image inputs such as multi-exposure sequences. The experimental results demonstrate that our technique achieves state-of-the-art performance in visual quality, objective assessment, and runtime efficiency.**

***Index Terms*—Image fusion, visual saliency, two-scale decomposition, neural networks, real-time**

I. Introduction

IMAGE fusion plays an important role in multiple image processing and computer vision applications. In fact, any procedure requiring the analysis of two or more images of the same scene benefits from image fusion [1]. For instance, image fusion between visible and infrared bands [2] is used in night-time surveillance, image dehazing [3], face recognition [4], military reconnaissance missions, and firefighting [5]. In medical imaging applications, images can come from multiple modalities such as magnetic resonance imaging (MRI), positron emission tomography (PET), and computed tomography (CT). The fusion of these images helps physicians provide reliable and accurate medical diagnosis to their patients, and navigate otherwise impossible surgeries [6]. In photography, the fusion of images taken with different focal settings allows photographers to obtain a single all-in-focus image [7], [8]. Finally, a stack of low dynamic range images with varying exposure levels can be fused into a single high dynamic range image [9], [10] that usually looks better.

A large number of image fusion techniques have been proposed in the literature [1], [2], [6], [7]. Classical approaches often rely on multi-scale fusion techniques [11], [12]. These methods excel at preserving details from different source images. However, they may produce brightness and color distortions since they do not consider spatial consistency in their fusion process. Other approaches use sparse representation [13], [14], or pulse-coupled neural networks [15], [16], and rely on optimization techniques with multiple iterations. Although they can produce better fusion results, they are computationally inefficient. Finally, all these methods rely on manually crafted fusion rules that might not apply for all source images.

Recently, convolutional neural networks (CNN) have been successfully used to design robust and efficient fusion rules for multiple fusion tasks [10], [17], [18], with state-of-the-art performance. In fact, neural networks can be considered as feature extractors, where intermediate maps represent salient features that can be used to generate fusion weight maps. Unlike classical methods, which are based on a single fusion rule, CNNs are trained on large datasets and can model a variety of image features, allowing them to obtain more general fusion rules. However, training these networks requires the construction of large datasets, which is expensive and time-consuming. They are also specialized to single tasks and do not perform as well on tasks they were not trained for. Finally, they require large amounts of memory, time, and energy to run.

To solve the aforementioned problems, we propose a novel and fast image fusion based on *pre-trained* convolutional neural networks for image fusion, illustrated in Fig. 1. We first decompose the source images into base and detail layers. We use visual saliency to fuse the base layers, and deep feature maps from a pre-trained CNN to fuse the detail layers. The weights maps are aligned with the source images through a guided filter. Then, both fused base and detail levels are joined to reconstruct the fused image.

Our approach is based on pre-trained neural networks, which have learned a multitude of example images and incorporated a large set of image priors. This, contrasted with MST-, SR-, and PCNN-based methods, allows our technique to generalize better on various images and image fusion tasks. Additionally, unlike prior CNN-based fusion techniques, our method requires neither a specific dataset nor a particular network architecture since it leverages pre-trained network models. Therefore, our method alleviates the need to collect training data, and generalizes well to any fusion problem.

We conduct experiments to analyze the proposed method and its parameters. First, we study the effect of the decomposition filter on the fusion quality and its runtime. Second, we run ablation studies to understand the impact of the weight construction techniques, the guided filter parameters, and the depth and architecture of the pre-trained CNNs on the fusion performance. The ablation studies show the advantages of our fusion rules and their robustness to changes in parametrization. We compare our approach with state-of-the-art methods on thermal, medical, and multi-focus fusion. Our technique demonstrates advantages in visual quality, objective assessment, and runtime efficiency. Finally, we show that our method is applicable to any number of source images by displaying multi-exposure sequence fusion results.

Our main contribution is a novel technique to integrate pre-trained neural networks in an image fusion pipeline. Our

The authors are with the School of Computer and Communication Sciences, École polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland. E-mail: (fayez.lahoud@epfl.ch;sabine.susstrunk@epfl.ch).

Code available at https://github.com/IVRL/Fast-Zero-Learning-Fusion

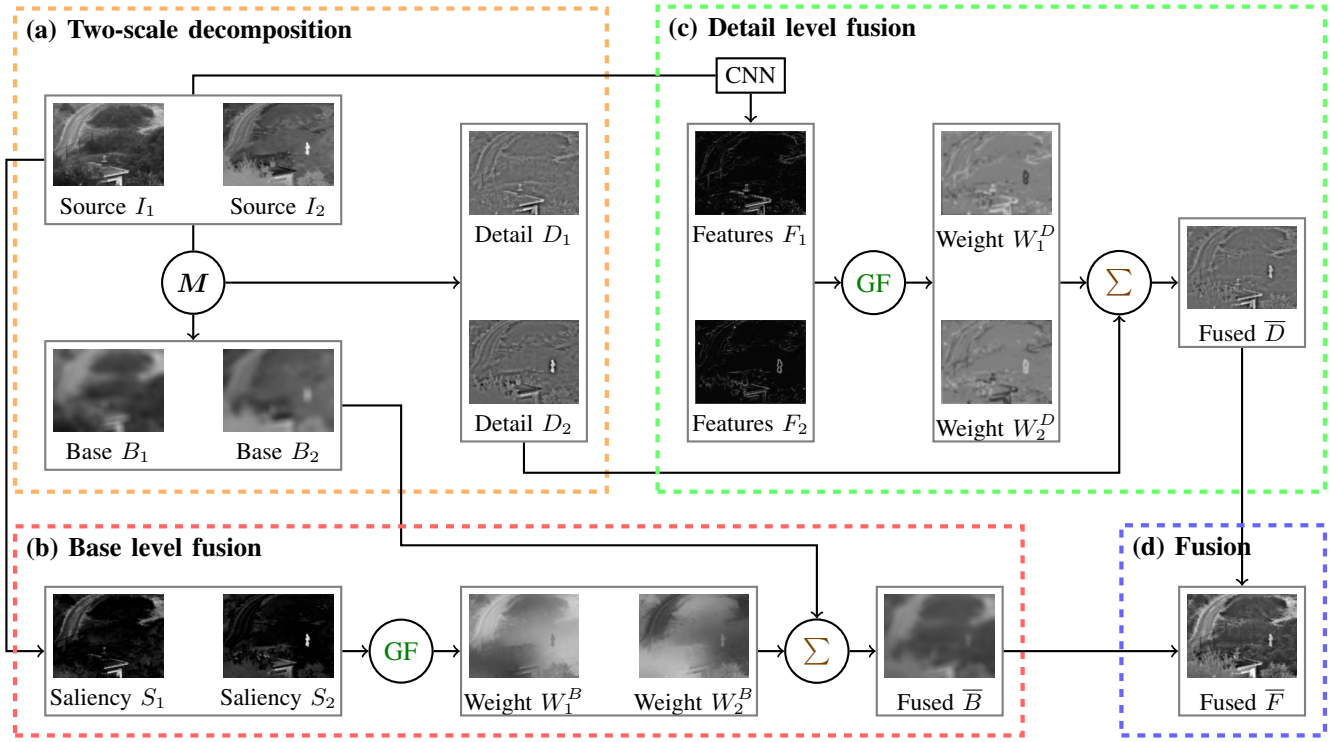

Fig. 1. Schematic diagram of our proposed fusion method. The two-scale decomposition transforms the image into base and detail layers. The base layers are fused based on a saliency measure comparison, and the detail layers based on CNN intermediate feature maps. The guided filter is used to smooth the weight maps and enforce consistency with the source images. In the last stage, the fused base and detail layers are joined to obtain the final fusion. M, GF, and $\sum$ indicate the two-scale decomposition filter, the input-guided filter, and the weighted sum operator, respectively.

method shows state-of-the-art performance and runs in real time. The low computational costs make it very beneficial for embedded systems applications, such as thermal fusion for firefighters where security restrictions enforce hardware limitations, medical fusion for continuous monitoring systems and efficient diagnosis, or multi-focus and multi-exposure fusion integrated in camera pipelines.

II. Related Work

We present a two-scale image fusion method based on saliency and pre-trained convolutional neural networks. Consequently, we review classical and CNN-based image fusion approaches. We discuss two-scale decompositions, visual saliency, and edge-preserving filters in the context of image fusion. Finally, we address pre-trained CNNs and their ability to represent images in a space well-suited for comparison.

A. Image fusion

Classical fusion algorithms use multi-scale transforms (MST) to obtain perceptually good results. Frequently used MSTs include pyramids [11], [19], wavelets [12], [20], filters [21], [22], and geometrical transforms like contourlets and shearlets [15], [23], [24]. Another popular approach for modeling images is sparse representation (SR) [13], [14], [25], [26]. Under SR, images are encoded as a sparse linear combination of samples representing salient features. This encoding is used to fuse the source images. Pulse-coupled neural networks (PCNN) are another set of models that are commonly used in image fusion [15], [16]. They are inspired by the cat's visual cortex, and are two-dimensional, with each neuron corresponding to a single pixel in an image. MST, SR, and PCNN approaches enforce a specific relation between inputs and outputs, i.e., they operate on a rule which fits a limited set of predefined image priors. Our approach is based on pre-trained CNNs which learn by example. Having been trained on a large amount of examples, these networks are able to construct a more complete understanding of image priors, and as such generalize better on image fusion tasks.

The earliest fusion work involving neural networks poses multi-focus fusion as a classification task [27]. Three focus measures define the input features to a shallow network which outputs the weight maps corresponding to the source images. Due to architectural constraints, the method can only run on image patches, and generates boundary artifacts. More recently, convolutional neural networks have been trained to generate decision maps for multi-focus [17], multi-exposure [10], medical [28], and thermal fusion [18]. Although these approaches often achieve better performance than their classical counterparts, they still have major drawbacks. First, they require large specialized datasets for training. Second, deep networks often overfit the datasets they are trained on, e.g., a network trained for multi-focus image fusion will only be suitable for that task. Finally, CNNs require large amounts of memory and are computationally expensive both in time and energy. In contrast, our method requires no training, which alleviates the necessity of collecting data. Moreover, we use a single pre-trained network as feature extractor for

any image fusion task, so our technique generalizes well to any fusion problem. Finally, by evaluating the impact of depth and architecture, we can select a memory and time efficient neural network for our fusion method.

B. Two-scale decomposition

Two-scale decompositions are a subset of MSTs that are commonly used for image enhancement and tone mapping [29]–[31]. The two-scale approach avoids mixing low and high frequency information and reduces halo artifacts. It operates in the spatial domain, and preserves source intensity values and spatial constraints. The base layer contains large scale intensity variations and is obtained by applying a smoothing filter on the image. The detail layer is the difference between the original image and the base layer and contains small scale information. Multiple decomposition filters have been proposed and used for image fusion such as the average filter [21], [32], bilateral filter [22], [33], [34], and local extrema [30], [35]. We conduct ablation studies to choose the most appropriate filter for our method.

C. Visual saliency

Visual saliency detects perceptually salient visual structures, regions or objects in an image [36]–[38]. These locations stand out from their neighborhood, and attract the human visual system's attention. Saliency is widely used in multiple computer vision tasks such as object recognition [39] and person re-identification [40]. As for image fusion, saliency maps commonly serve as activity level measurements, since they reflect the important features of the source images [26], [41]. We use saliency to fuse the base layers as it correlates with the contrast between different low-frequency image regions. We use the method presented in [36] because it can produce saliency maps in $O(N)$ time with low memory requirements.

D. Edge-preserving filters

Edge-preserving filters such as the bilateral filter [33], guided filter [42], and cross-bilateral filter [43] smooth image details while preserving strong edges and avoiding ringing artifacts. Due to these qualities, they are used in image fusion either as multi-level decomposition methods [22], [34], [44] or consistency verification schemes [17], [21]. Among these filters, the guided filter [42] is a time-efficient method that runs in linear time, independently of the filter size. It is used in multiple applications such as detail enhancement and no-flash denoising [42], and in image fusion methods [21], [44]. We use it as a consistency verification model to ensure that the generated weight maps adhere to the region boundaries defined by the source images.

E. Pre-trained models

Pre-trained neural networks are networks that have been previously trained on a specific task and dataset. While these networks are specifically built for that task in mind, they can be used to generate image representations for other tasks. A simple example would be to use the neural networks as feature extractors instead of manually designing shape, color, and texture features for classification [45]. Additionally, neural style transfer compares deep features from a pre-trained network to transfer style statistics from a target image to a source image [46]. Deep features can also be stacked together to form hypercolumns for semantic image segmentation [47]. These works show that deep features are well-suited to compare texture, style, and semantics between images without any additional training or data collecting costs. Similarly, we use pre-trained neural networks to generate detail fusion weights.

III. CNN-BASED IMAGE FUSION

The schematic in Fig. 1 summarizes the main sections of the proposed two-scale decomposition fusion method. First, an image filter is applied to divide the image into base and detail levels. Then the base and detail layers are fused based on saliency measures and neural network activations, respectively. At the end, the fused representations are joined together to reconstruct the final fused image.

A. Two-scale decomposition

Suppose there are K pre-registered source images denoted as $I_k|k \in \{1, 2, \ldots, K\}$. As seen in Fig. 1(a), each source images I_k is decomposed into two-scale levels using a smoothing image filter. The base layer of source image I_k is the result of the smoothing operation using filter M

$$B_k = M(I_k). \tag{1}$$

Given the base layer B_k, the detail layer is obtained by subtracting it from the source image I_k as

$$D_k = I_k - B_k. \tag{2}$$

This two-level decomposition separates the image into two components, the first containing large-scale intensity variations and smooth regions and another containing small-scale edges and image details such as textures. The choice of smoothing filter and its effect on the quality and runtime of the fusion are discussed in the experimental evaluations.

B. Base layer fusion

In order to fuse the base layers $B_k|k \in 1, 2, \ldots, K$, we first compute the visual saliency map of each image I_k, as illustrated in Fig. 1(b). Let $I(p)$ denote the intensity value of pixel p in a given image I. The saliency value $S(p)$ at pixel p is defined as

$$\begin{aligned} S(p) &= |I(p) - I(1)| + |I(p) - I(2)| + \cdots + |I(p) - I(N)| \\ &= \sum_{i=0}^{255} M(i)|I(p) - i|, \end{aligned} \tag{3}$$

where N is the total number of pixels in the image, and $M(i)$ is the frequency of pixels whose value is i, i.e., the histogram of the image at i. The histogram can be computed in $O(N)$ time order, and the distance measures $|I(p) - i|$ can be calculated for all $i \in [0, 255]$ in constant time. By pre-computing

all color distances, a map $D(x, y) = |I(x) - I(y)|$ can be constructed for all color value pairs. Given the histogram $M(.)$ and the color distance matrix $D(.,.)$, the saliency for pixel p is simply computed as

$$S(p) = \sum_{i=0}^{255} M(i)D(p, i), \tag{4}$$

which also completes in constant time. The saliency maps are then normalized to the range $[0, 1]$. For further details, please refer to the complete implementation in [36].

Following the computation of saliency maps S_k for each image I_k, we generate the base layer weight maps W_k^b as follows

$$W_k^b = \frac{S_k}{\sum_{j \in K} S_j}. \tag{5}$$

The weight maps W_k^b obtained are generally noisy and do not completely respect region boundaries, which could further produce artifacts in the fused image. Enforcing spatial consistency via guided filtering solves this problem: the filtering process ensures that adjacent pixels with similar intensity values obtain similar weights. This leads to smooth weight maps for the base layers without introducing artificial edges. The smoothed weights $\overline{W_k^b}$ are thus obtained as

$$\overline{W_k^b} = G_{r_b, \epsilon_b}(W_k^b, I_k), \tag{6}$$

where $G_{r,\epsilon}(.,.)$ is the guided filter, and r and ϵ control the radius and degree of blur, respectively. After obtaining the K weight maps, they are all normalized such that they sum up to one at each pixel p

$$\overline{W_k^b} = \frac{\overline{W_k^b}}{\sum_{j \in K} \overline{W_j^b}} \ . \tag{7}$$

Following that, the fused base layer is constructed as a weighted-sum of all base layers given the weight maps

$$\overline{B} = \sum_{k \in K} \overline{W_k^b} B_k. \tag{8}$$

C. Detail layer fusion

We propose a novel fusion strategy using convolutional neural networks to extract deep features and generate weight maps. In this section, we detail how we obtain the weight maps, and construct the fused image. Fig. 1(c) illustrates the process of taking two source images, extracting their representative feature maps, generating smoothed fusion weights according to their activation levels, and finally using them to fuse the detail layers.

Recall that there are K pre-registered source images denoted as $I_k | k \in \{1, 2, \cdots, K\}$. Additionally, suppose there is a pre-trained convolutional neural network with $\mathcal{L}$ layers, with C_l output channels per layer l. We denote $F_k^{c,l}$ as the c-th feature map of the k-th image extracted at the l-th layer of the network (taken after ReLU operation), the feature map is computed as

$$F_k^l = \max(0, \mathcal{F}_l(I_k)), \tag{9}$$

where $\mathcal{F}_l(.)$ is the application of the network function to the input image up to layer l. The $\max(0, .)$ function denotes the ReLU operation.

For every feature map, we denote $\hat{F}_k^l$ as the l_1-norm computed over the C_l channels of the feature maps of layer l as follows

$$\hat{F}_k^l = \sum_{c=0}^{C_l} ||F_k^{c,l}||_1 \ . \tag{10}$$

This constitutes a measure of the activity corresponding to the input image at layer l where high pixel values correspond to high activity levels.

The feature maps are extracted for $\mathcal{L}$ layers , so we obtain, per image k, a set of features maps $\{\hat{F}_k^l | l \in \mathcal{L}\}$. For every layer l, the K feature maps are used to generate K weight maps that indicate the amount of contribution of each image at a specific pixel. For our method, we use the softmax operator to generate said maps as follows

$$W_k^{d,l} = \frac{e^{\hat{F}_k^l}}{\sum_{j=1}^{K} e^{\hat{F}_j^l}} \ , \tag{11}$$

where $e^{(.)}$ is the exponentiation with base e.

In order to account for small mis-registrations, and remove undesirable artifacts around the edges of both modalities, we apply a guided filter to the weight maps. This correctly aligns the weights with the source edges while preserving the sharpness of their details. After smoothing the weight maps using the guided filter as

$$\overline{W_k^{d,l}} = G_{r_d, \epsilon_d}(W_k^{d,l}, I_k), \tag{12}$$

the map are normalized such that they sum up to one at each pixel p, similar to Eq. (7).

At layer l, we have a set of weights $\overline{W_k^{d,l}} | k \in \{1, 2, \cdots, K\}$. Using these weight maps, the image fusion at layer l is computed as

$$\overline{D^l} = \sum_{k=1}^{K} \overline{W_k^{d,l}} D_k. \tag{13}$$

Finally, some neural networks contain convolution layers with large strides or pooling layers, and thus generate deep feature maps with smaller spatial resolution than the input images. In these cases, we simply upsample the feature maps using nearest neighbors. The subsequent filtering step ensures the weight maps remain smooth and adhere to the edges of the source images even after upsampling.

D. Two-scale reconstruction

Having obtained both fused base $\overline{B}$ and detail $\overline{D}$ layers, the fused image $\overline{F}$ is the pixel by pixel combination of those layers, as illustrated in Fig. 1(d):

$$\overline{F} = \overline{B} + \overline{D}. \tag{14}$$

$\overline{F}$ is clipped to the appropriate range to remove any out of range values. Finally, if desired, a tone mapping function could be applied to the result.

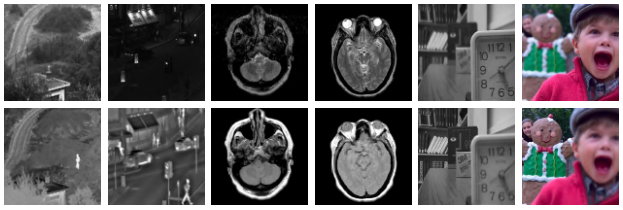

Fig. 2. Pairs of testing images from the datasets used in the experiments. From left to right, shown are two pairs of visible and infrared images, two pairs of MRI-T1 and MRI-T2 images, and two pairs of multi-focus images.

IV. Experiments

A. Experimental setup

We perform evaluations on three fusion tasks using different datasets. The first dataset is the TNO set containing 21 pairs of visible and infrared images representing natural scenes [48]. The second dataset is extracted from the Whole Brain Atlas [49], and contains 97 pairs of computed tomography (CT) and magnetic resonance imaging (MRI) images, and 24 pairs of T1-T2 weighted MRI images. CT is sensitive to dense structures and bones while MRI captures softer organ tissues. T1 and T2 are relaxation times that characterize tissue, and thus capture different properties of the same substance under MRI. The third dataset consists of 20 commonly used images in multi-focus fusion ('Book', 'Clock', 'Desk', etc.) in addition to the multi-focus color images from the Lytro dataset [50]. The images from these datasets have been used in many related papers: TNO [4], [25], [41], Whole Brain Atlas [21], [24], [28], multi-focus images [17], [51], [52], and Lytro [1], [17]. All the images are pre-registered. Fig. 2 shows sample pairs taken from these datasets.

In the following experiments, most of the source images are grayscale and the fusion is as presented in Sec. III. However, for the Lytro dataset, color images are fused via their luma components. The sources are transformed from RGB color space to YCbCr. Then the Y components are fused together while the Cb and Cr components are averaged. The final fused image is reconstructed by converting the intermediate YCbCr fused image back to RGB.

B. Image fusion metrics

In order to quantitatively evaluate the performance of different fusion methods, we use eight commonly adopted objective fusion metrics. They are entropy EN, mutual information MI, visual information fidelity of fusion VIFF [53], mutual information index Q_{MI} [54], edge information index Q_G [55], structural similarity index Q_Y [56], Cvejic's metric Q_C [57], and phase congruence index Q_P [58]. EN reflects the amount of information present in the fused image while MI estimates the amount of information transferred from the inputs images to the fused image. VIFF highly correlates with the human visual system. Q_{MI} measures how well information from every source image is preserved. Q_G measures the amount of edge information transferred. Q_Y measures how well structural information is preserved in the fusion process. Q_C evaluates the method's ability in transferring information while reducing distortions. Finally, Q_P reflects how well salient features are preserved. For all these metrics, a higher value suggests a better performance. Values for VIFF, Q_G, Q_Y, Q_C, and Q_P are normalized values and bounded between $[0, 1]$. EN, MI, and Q_{MI} are bounded by the number of bits used to represent pixels based on the Shannon Entropy rule, e.g., EN is bounded between $[0, 8]$ for images using 8-bit pixel representation $[0 - 255]$.

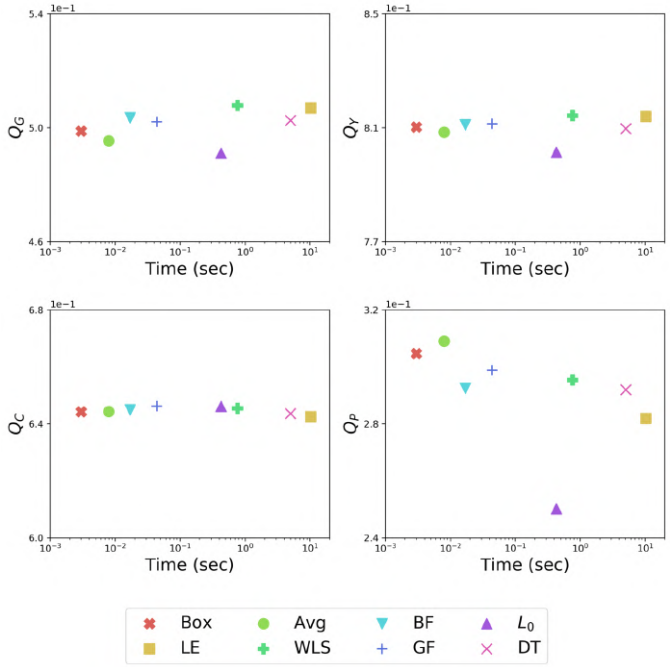

Fig. 3. Fusion quality with respect to two-scale decomposition runtime.

C. Ablation studies

In this subsection, we study the influence of different parameters on our fusion model. First, we conduct a study on two-scale decomposition filters, and compare their impact on the quality and runtime of our method. Then, we compare our weight generation techniques using saliency and CNNs with the max and average fusion rules commonly found in the literature [18], [24], [41]. As we are using a parameterizable guided filter function, we also analyze the effect of its free parameters on both base and detail layer fusion. Finally, we study the effect of network architecture and layer depth on the fusion in terms of performance, runtime, and memory consumption. The ablation studies are performed on the TNO dataset. Here, we only present plots for Q_G, Q_Y, Q_C, and Q_Y. Please refer to the supplemental material for additional results.

1) Two-scale decompositions: In this experiment, we compare commonly used two-scale decomposition filters for their impact on the proposed method. The compared filters are box (Box), average (Avg), bilateral (BF) [33], L_0-minimization (L_0) [59], local extrema (LE) [30], weighted least squares (WLS) [29], guided filter (GF) [42], and domain transform (DT) [31]. Fig. 3 shows the fusion quality with respect to the two-scale decomposition runtime of each evaluated filter. The

TABLE I
QUANTITATIVE ASSESSMENT OF DIFFERENT WEIGHTING SCHEMES FOR BOTH BASE AND DETAIL LAYERS.

Metric	AVG-MAX	AVG-CNN	S-MAX	S-CNN
EN	6.639	6.618	7.016	**7.126**
MI	13.277	13.236	14.032	**14.252**
VIFF	0.562	0.625	0.610	**0.690**
Q_{MI}	2.006	1.997	2.063	**2.064**
Q_G	0.444	0.498	0.440	**0.500**
Q_Y	0.768	0.807	0.775	**0.816**
Q_C	0.614	0.623	0.600	**0.649**
Q_P	0.251	0.294	0.270	**0.315**

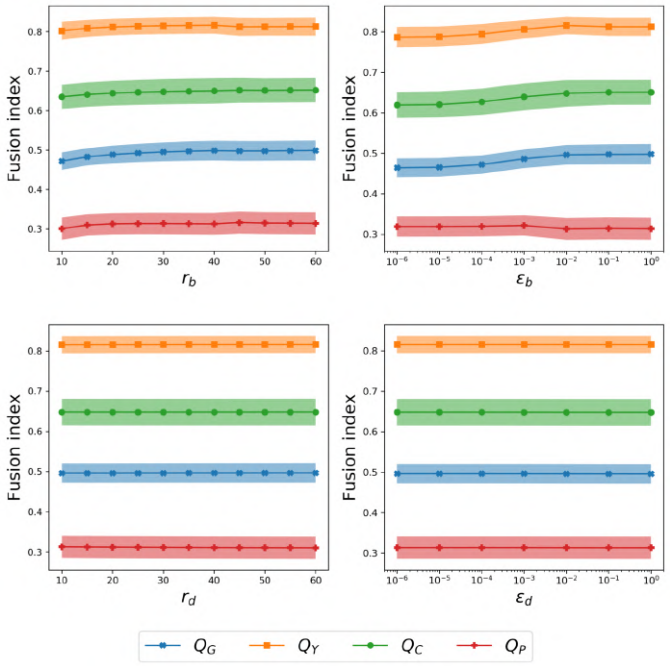

Fig. 4. Fusion quality with respect to the guided filter parameters. The shaded regions indicate the range of values obtained on the evaluated images.

quality of the resulting fusion is not heavily affected by the choice of filter, with the maximum difference across all metrics being 0.062 for Q_P between the L_0 and Avg filters on a scale from $[0, 1]$. However, the runtime of these decompositions varies between 0.001 seconds for the box filter and 10.312 seconds for the local extrema filter. The plots demonstrate that a simple filter for decomposition is as good as the more complicated edge-aware or minimization-based filters while being orders of magnitude faster.

2) Weight construction methods: We compare our fusion rules to the commonly used weight construction approaches of averaging (AVG) for base layers, and maximum for detail layers (MAX). In comparison, our fusion rules are saliency (S) for base layers, and (CNN) for detail layers. We compare the following weight construction approaches: AVG-MAX, AVG-CNN, S-MAX, and S-CNN. Table I shows the average values of the fusion metrics on the TNO dataset. EN, MI, VIFF, and Q_P significantly improve between AVG-MAX and S-MAX, showing the impact of saliency based fusion on the base layers. Additionally, Q_G, Q_Y, and Q_C show small differences

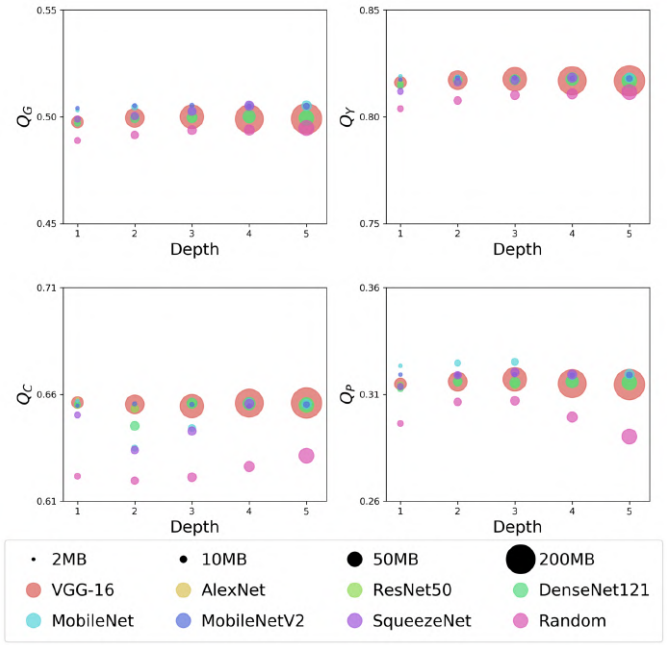

Fig. 5. Fusion performance with respect to network architecture and depth. The depth indicates the convolutional block at which the deep feature maps are compared. The disk size indicates the memory requirements for the evaluation up-to the specified depth.

between AVG-CNN and S-CNN and large differences against their AVG-MAX and S-MAX counterparts, confirming the benefits of using pre-trained deep network features to fuse the detail layers. Together, the saliency and CNN-based fusion weights generate the highest quality fusion results on all evaluated metrics, validating our choices for fusion rules.

3) Guided filter parameters: We use two different guided filters $G_{r_b,\epsilon_b}(.,.)$ and $G_{r_d,\epsilon_d}(.,.)$, and as such have four different parameters to tune. In order to study the impact of each parameter, we freeze the remaining ones. The base guided filter is frozen to a large size and blur degree to ensure smooth base weight maps. In contrast, the detail guided filter is frozen to a small size and blur degree to preserve sharpness in the detail weight maps. As such, when a parameter is frozen it is set to one of $r_b = 35$, $\epsilon_b = 0.01$, $r_d = 7$, and $\epsilon_d = 10^{-6}$. A similar approach was also employed in [21]. The filter size r is changed linearly from 10 to 60 in steps of 5, and the blur amount ϵ is changed logarithmically from 10^{-6} to 1 in steps of 10.

Fig. 4 shows the changes in objective fusion quality with respect to changes in the guided filter parameters. The plots for r_b and ϵ_b show that larger values are preferred to obtain a higher quality base layer fusion. However, the choice of r_d and ϵ_d do not significantly affect the performance, even though higher values lead to slightly lower quality.

4) Architecture and depth impact: In this experiment, we study the effect of the neural network architecture on the fusion quality. We consider seven models pre-trained on ImageNet [60] VGG-16 [61], AlexNet [62], ResNet50 [63], DenseNet121 [64], MobileNet [65], MobileNetV2 [66], and SqueezeNet [67]. VGG-16 and AlexNet are feed-forward convolutional neural networks, while ResNet50 and DenseNet121

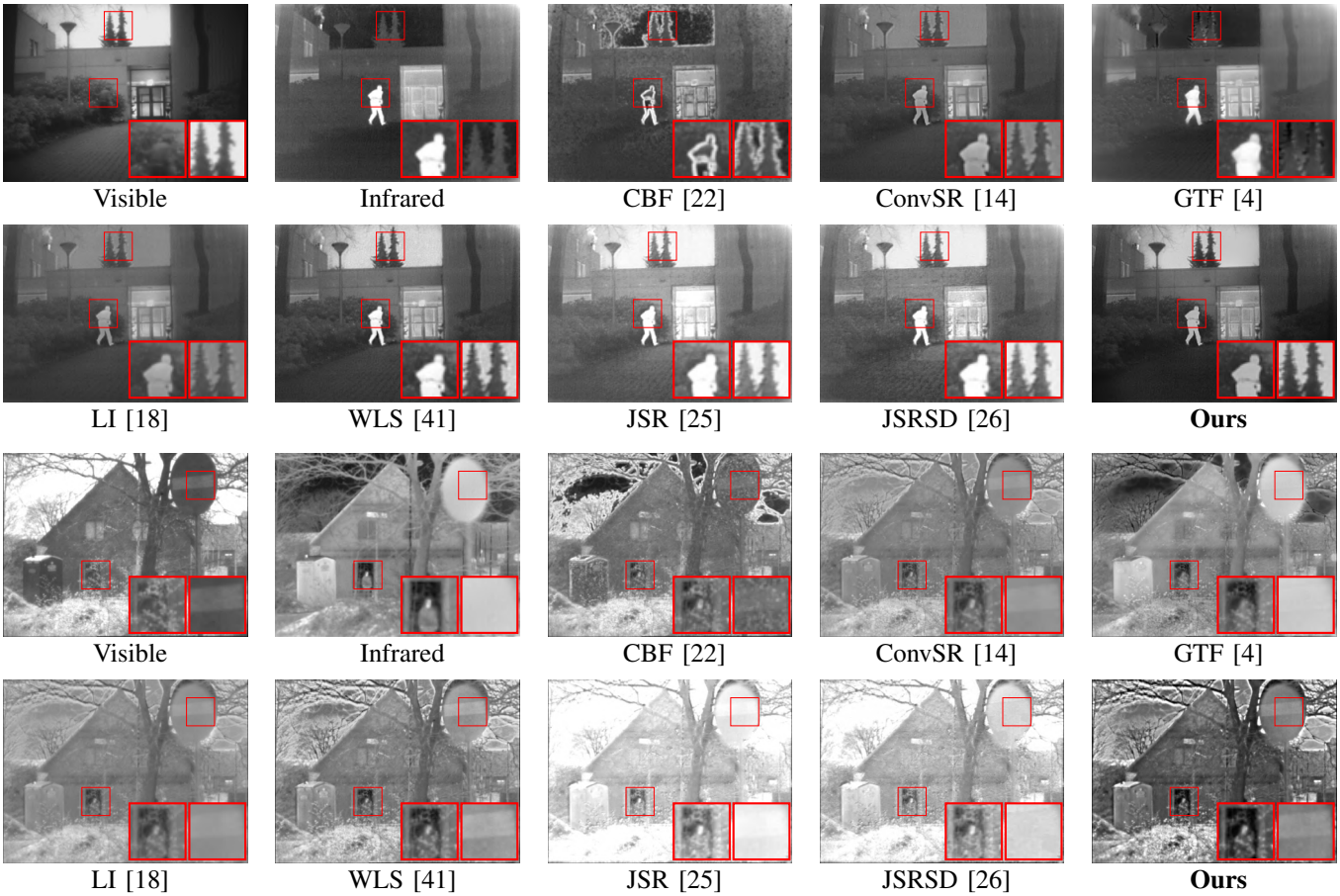

Fig. 6. Visible and infrared source images with the fusion results obtained by different methods. Insets are magnified ×2. Best viewed on screen.

TABLE II
OBJECTIVE ASSESSMENT OF DIFFERENT METHODS ON INFRARED AND VISIBLE IMAGE FUSION. **BOLD** AND UNDERLINED VALUES INDICATE THE **BEST** AND SECOND-BEST SCORES, RESPECTIVELY. TIME WAS COMPUTED ON IMAGES WITH AN AVERAGE SIZE OF 460×610 ($\pm 115 \times 155$).

Metric	CBF [22]	ConvSR [14]	GTF [4]	LI [18]	WLS [41]	JSR [25]	JSRSD [26]	Ours
EN	6.857	6.259	6.635	6.182	6.638	6.363	6.693	**7.126**
MI	13.714	12.517	13.271	12.364	13.276	12.727	13.386	**14.252**
VIFF	0.265	0.272	0.188	0.259	0.444	0.363	0.292	**0.690**
Q_{MI}	2.039	1.946	2.006	1.934	2.006	1.963	2.014	**2.064**
Q_G	0.378	0.491	0.421	0.364	**0.509**	0.308	0.265	0.500
Q_Y	0.643	0.802	0.726	0.702	0.805	0.588	0.501	**0.816**
Q_C	0.486	0.594	0.468	0.606	0.601	0.467	0.427	**0.649**
Q_P	0.147	**0.355**	0.205	0.297	0.309	0.173	0.118	0.315
Time	15.28	86.8	2.57	6.20	1.28	346.18	396.74	**0.16**
Std σ	6.69	37.44	1.31	2.79	0.65	151.25	178.73	**0.08**

contain residual connections, i.e. direct links from earlier to later stages of the model. Finally, MobileNet, MobileNetV2, and SqueezeNet are smaller and more efficient models designed for limited hardware devices and embedded systems. Finally, we also consider a non trained ResNet50 network, i.e., its parameters are randomly initialized.

Deep neural networks consist of multiple layers, and it is important for our application to understand the effect of depth on the output fusion. To that end, we compute the image fusion at multiple depth levels of the network and compare them. Here, we define network depth as the layer at the end of a convolution block. For example, the VGG-16 architecture consists of 5 convolution blocks, and we take the features at the end of each one. We proceed similarly for the other architectures. For more information on their structure, please refer to the original works [61]–[67].

Fig. 5 shows the average fusion metrics on images generated using the compared network architectures at different depths. All pre-trained networks perform similarly on the evaluated metrics. The standard deviations of the mean metric across different pre-trained networks are 2×10^{-3} for Q_G, 1×10^{-3} for Q_Y, 5×10^{-3} for Q_C, and 3×10^{-3} for Q_P, showing the similarity between those results. These experiments highlight that our method is stable under network and depth variations,

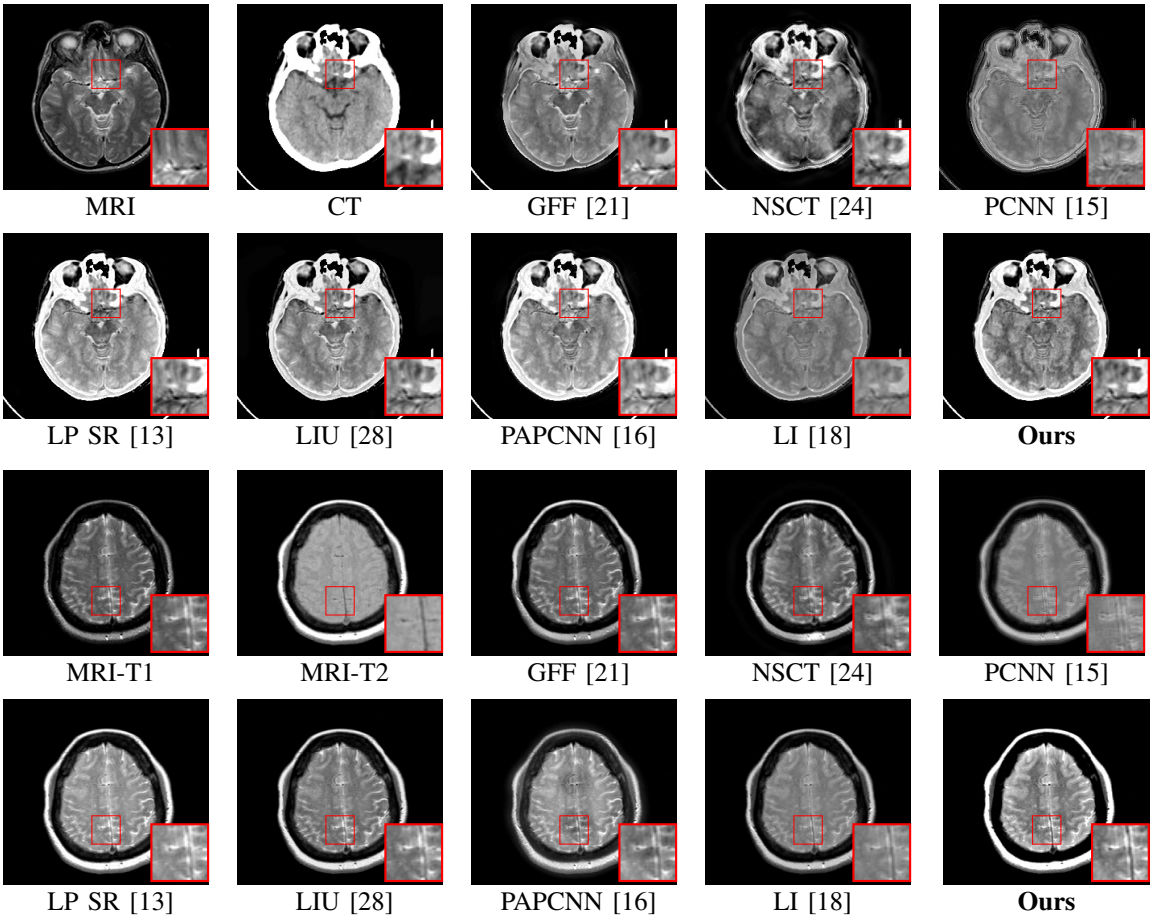

Fig. 7. MRI-T1 and MRI-T2 source images with the fusion results obtained by different methods. Insets are magnified $\times 2$. Best viewed on screen.

TABLE III
OBJECTIVE ASSESSMENT OF DIFFERENT METHODS ON MEDICAL IMAGE FUSION (MRI-CT AND T1-T2). **BOLD** AND UNDERLINED VALUES INDICATE THE **BEST** AND SECOND-BEST SCORES, RESPECTIVELY. TIME WAS COMPUTED ON IMAGES WITH A SIZE OF 256×256.

Metric	GFF [21]	NSCT [24]	PCNN [15]	LP SR [13]	LIU [28]	PAPCNN [16]	LI [18]	Ours
EN	3.322	3.579	3.481	3.424	3.610	3.969	3.329	**4.470**
MI	6.644	7.158	6.963	6.849	7.219	7.937	6.657	**8.939**
VIFF	0.110	0.110	0.085	0.164	0.144	0.112	0.108	**0.752**
Q_{MI}	0.953	0.694	0.787	0.976	0.899	0.761	**1.022**	0.882
Q_G	0.488	0.399	0.447	0.496	0.437	0.435	0.481	**0.710**
Q_Y	0.539	0.405	0.447	0.565	0.531	0.513	0.534	**0.826**
Q_C	0.508	0.370	0.412	0.522	0.494	0.477	0.491	**0.744**
Q_P	**0.720**	0.312	0.105	0.624	0.638	0.524	0.589	0.639
Time	0.05	3.49	0.46	0.04	13.12	6.92	1.65	**0.03**
Std σ	0.02	1.90	0.10	0.008	1.91	2.46	0.22	**0.001**

and can be applied using a variety of networks. Additionally, comparing the quality of pre-trained and non trained ResNet50 on all depth levels shows the importance of using pre-trained models that have learned natural image representations. Having similar performance than the feed-forward architectures, the residual networks offer lower memory constraints. Additionally, MobileNet, MobileNetV2, and SqueezeNet perform as well as the other architectures, and can be adopted on low-energy and limited hardware systems.

D. Comparison with other methods

We compare our method with different fusion algorithms on thermal, medical, and multi-focus fusion. We select commonly used fusion methods pertaining to each fusion task. The parameters of all the evaluated methods are set to the default values from their publicly available code. Following the observations from the ablation studies, we set the following configuration for our method. We use a box filter for the two-scale decomposition and set the guided filter parameters to $r_b = 45$, $\epsilon_b = 0.1$, $r_d = 7$, and $\epsilon_d = 10^{-6}$. We generate the detail weight maps using a ResNet50 network, with feature

maps extracted at depth $l = 3$.

1) Infrared and visible fusion: The methods we compare against are CBF [22], ConvSR [14], GTF [4], WLS [41], JSR [25], JSRSD [26], and LI [18]. CBF uses a cross bilateral filter to extract detail layers of source images, and use those layers to generate weight maps to combine the sources. ConvSR uses convolutional sparse representation to fuse the detail layers of two-scale decomposed images, while averaging the base layers. GTF poses the infrared and visible image fusion as total variation minimization problem in the infrared domain constrained to the gradients of the visible image. WLS uses visual saliency and weighted least-squares to fuse two-scale image decompositions. JSR and JSRSD use dictionaries to learn sparse image representations for weight computation, JSRSD additionally uses saliency information to guide the fusion. LI uses a neural network to directly fuse the high frequencies and averages the low frequencies.

Fig. 6 shows two pairs of visible and infrared images, and the fusion results from the compared methods. In comparison with CBF and ConvSR, our method has less edge and ringing artifacts. Additionally, JSR and JSRSD produce over-exposed images while LI generates under-exposed results. In GTF, it's very difficult to notice the trees, as seen in the inset image. In contrast, our method better respects the intensity values present in the source images. Additionally, in the second pair of images, the person and the sign post are most noticeable in our fusion among all the compared methods.

Additionally, Table II summarizes the average objective metrics on infrared and visible fusion of all images in the TNO dataset. Our method consistently records the highest or second-highest performance on all the evaluated metrics, achieving state-of-the-art performance on thermal fusion.

2) Medical image fusion: For this task, the compared methods are GFF [21], NSCT [24], PCNN [15], LP SR [13], LIU [28], PAPCNN [16], and LI [18]. GFF is based on a base and detail decomposition with guided filters to generate weights. NSCT and PCNN both use the non-subsampled contourlet transform to decompose the images into low and high frequencies, and apply different fusion rules depending on the frequency. LP SR uses Laplacian pyramids to decompose the images and applies sparse representation for low frequency fusion and maximum selection rule for high frequency fusion. LIU employs Siamese Neural Networks [68] to predict the fusion weights with input images in the spatial domain. PAPCNN uses the non-subsampled shearlet transform for image decomposition, with an energy based fusion for low frequencies, and an adaptive PCNN for high frequencies. LI is the same method presented in the previous section.

Fig. 7 shows two pairs of images corresponding to grayscale fusion of MRI-CT and T1-T2 images, respectively. GFF, NSCT and PAPCNN are not able to capture complementary information in all the image regions, which is reflected in the stark contrast inside regions that should have similar intensities. This can be seen for GFF in the MRI-CT fusion, where some edges exhibit a single intensity in both sources but present different levels in the fusion. Similar artifacts can be spotted for NSCT in MRI-CT and PAPCNN in T1-T2 fusion. In contrast, our fusion approach relies on neural networks that have been trained on various intensity levels and thus respects the intensity differences inside and between salient regions leading to a more intensity consistent fusion that is easier to interpret. Additionally, LIU and LI generate weights that are highly biased towards the highest intensity signal. This results in low contrast images, where darker details from sources do not appear in the fusion. In comparison, our method better represents edges between neighboring regions while respecting their saliency, resulting in higher contrast, noise-free fusions. Finally, Table III shows the performance of different methods averaged over both the MRI-CT and T1-T2 fusion sets. The evaluations show that our method pushes state-of-the-art performance on all metrics except in Q_{MI} where it still obtains competitive results.

3) Multi-focus fusion: We compare against the previously mentioned ConvSR [14], CBF [22], GFF [21], and NSCT [24]. We also compare against LIU-M [17], DSIFT [51], and BF (refered to as boundary finding fusion BFF, to reduce confusion with the bilateral filter) [52]. LIU-M uses convolutional neural networks to compare image patches and propose weight maps. Then, the weight map boundaries are refined using morphological operators and guided filters. DSIFT uses local feature descriptors to generate an initial decision map and refines it via local matching. Finally, BFF finds and classifies the boundaries between focused and defocused image regions, then generates the fused image by combining the focused parts of the source images. The other fusion methods do not use any boundary adherence checks.

Fig. 8 shows two pairs of multi-focus source images, and their corresponding fusions using the evaluated methods. In the first pair of images, notice that ConvSR, CBF, and NSCT generate artifacts that were not present in any source image. Additionally, BFF is not able to properly estimate the boundaries as can be seen in the inset. GFF, LI, DSIFT, and our method better estimate the boundaries between the focused and unfocused images. Similar observations can be made for the second row, with the flower petal unfocused in the BFF result, while more properly focused in ours. Table IV summarizes the performance of the evaluated methods on the multi-focus and Lytro datasets. Our method is competitive with DSIFT and BFF, which are the state-of-the-art approaches for multi-focus fusion. While BFF is able to obtain better performance on some metrics, it is unable to always correctly estimate the focus boundaries in the qualitative evaluation.

E. Computational costs

In addition to the quantitative and qualitative experiments, we conduct runtime evaluations for every compared method. The last two rows of Tables II, III, and IV show the average runtime of each method with its standard deviation on images taken from the different datasets. The experiments were run on an Intel Core i7-7700HQ CPU (2.8GHz), and a GeForce GTX 1050 GPU (2Gb) for those methods requiring it ([17], [18], [28]). All our computations are done on CPU, except the forward pass through the convolutional neural network which is run on GPU. However, thanks to the robustness of our method to different network architectures, we use a

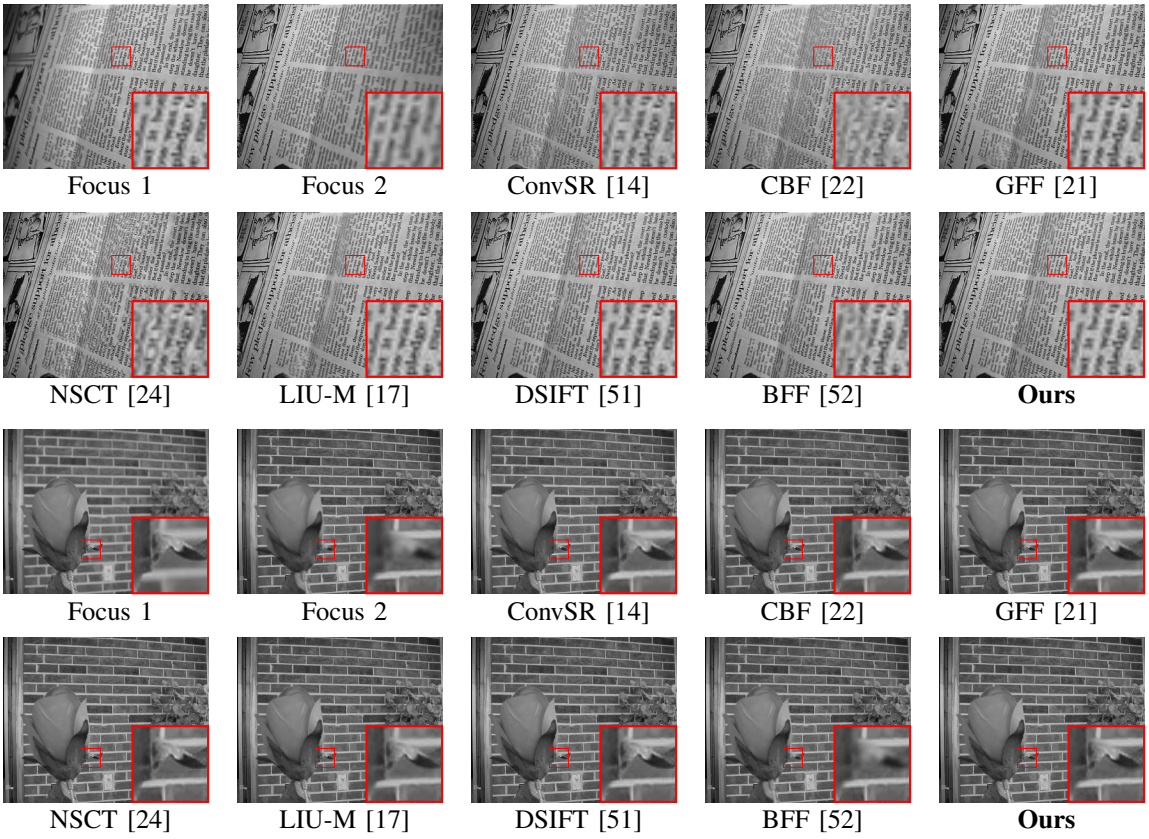

Fig. 8. Multi-focus source images with the fusion results obtained by different methods. Insets are magnified $\times 4$. Best viewed on screen.

TABLE IV
OBJECTIVE ASSESSMENT OF DIFFERENT METHODS ON MULTI-FOCUS IMAGE FUSION. **BOLD** AND UNDERLINED VALUES INDICATE THE **BEST** AND SECOND-BEST SCORES, RESPECTIVELY. TIME WAS COMPUTED ON IMAGES WITH AN AVERAGE SIZE OF 360×450 ($\pm 135 \times 155$).

Metric	ConvSR [14]	CBF [22]	GFF [21]	NSCT [24]	LIU-M [17]	DSIFT [51]	BFF [52]	Ours
EN	7.261	7.281	7.286	**7.297**	7.279	7.280	7.275	7.287
MI	14.522	14.562	14.573	**14.595**	14.559	14.560	14.551	14.575
VIFF	0.840	0.876	0.886	0.896	0.885	0.887	0.881	**0.899**
Q_{MI}	0.869	0.996	1.048	0.931	1.118	1.148	1.143	**1.155**
Q_G	0.588	0.677	0.703	0.673	**0.710**	0.709	**0.710**	0.708
Q_y	0.888	0.950	0.975	0.955	0.985	0.982	**0.987**	0.985
Q_c	0.639	**0.659**	0.642	0.650	0.635	0.629	0.629	0.646
Q_p	0.794	0.810	0.834	0.806	0.839	0.836	0.836	**0.841**
Time	62.70	10.33	0.14	45.87	80.79	4.24	0.58	**0.08**
Std σ	40.07	6.58	0.09	26.31	38.08	4.01	0.77	**0.05**

small network (ResNet50) whose runtime slows down by only $2\%(\pm 0.3)$ when moving its forward pass from GPU to CPU, barely impacting the overall runtime of the proposed pipeline.

Across the three evaluated fusion tasks, and the various source image dimensions, our method has the fastest run time while still obtaining state-of-the-art performance. This is because both saliency and guided filter algorithms run in $O(N)$ time complexity, and they can be even further sped up using GPU implementations [69]–[71]. These qualities make our approach suitable for deployment on limited hardware architectures.

F. Extension to multiple images

In this section, we demonstrate the applicability of our method to fusion problems with more than 2 inputs. For instance, multi-exposure fusion typically require multiple images in the exposure stack to capture the whole dynamic range and minimize the exposure bias difference between consecutive images [10]. Fig. 9 shows two multi-exposure image sequences taken from [72] and their respective fusion results using our proposed method. The weight maps show how the base layers preserve the intensity levels while respecting regions boundaries. The detail weight maps also show the consistency with the edges present in the sources, allowing for all details to be preserved in the result, as can be seen on the well-defined

Fig. 9. Multi-input fusion examples. On the left side, the images are organized as sources (rows 1 and 4), base weights (rows 2 and 5), and detail weights (rows 3 and 6). On the right side, the resulting image fusion is shown for both sequences. Best viewed on screen.

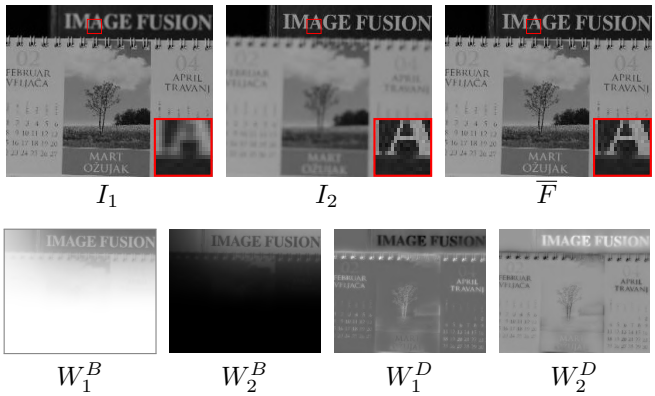

Fig. 10. Example of imperfect boundary in the base weight maps. Insets are magnified $\times 4$. Best viewed on screen.

leaf edges in the first example.

This example illustrates how the two-scale decomposition affects the fusion, but more importantly that our method works well even on sequences of input images. Note that the feed-forward pass of neural networks allows batch-processing of inputs. This means that the computational speed is barely affected by the number of images in the input sequence.

G. Limitations

In the multi-focus fusion task, we do not rely on any boundary detection method to refine the weight maps. The guided filter provides a good estimate of the boundaries between the multi-focus images, but does not always find the perfect boundaries. Fig. 10 shows an example where the base weight maps do not accurately represent the focus regions, leading to imperfectly focused fusion results. LIU-M [17] proposes the use of morphological operators to reshape the boundary conditions and fill holes in the multi-focus weight maps generated by their neural networks. Such adjustments to the focus weight maps generated by our method could further improve its multi-focus fusion weights, and consequently the resulting all-in-focus images.

V. Conclusion

We present a novel image fusion algorithm based on saliency and pre-trained neural networks. We first decompose source images into a base and a detail layer. Then, visual saliency maps and deep feature maps are used to compute base and detail fusion weights, respectively. Unlike typical neural network based techniques, our method requires no prior training on the image modalities and generalizes well to cover different fusion tasks. We demonstrate the robustness of our technique to the choices of decomposition filter, network architecture and feature depth.

Due to its robustness, we can configure our method to generate extremely fast and high-quality images, obtaining state-of-the-art results. We also demonstrate its applicability to diverse image fusion tasks, namely thermal, medical, and multi-focus fusion. Additionally, we show that our method can be extended to any number of input images. In conclusion, our method is a lightweight and high-quality technique with

promising applications in real time systems and on low-energy hardware.

ACKNOWLEDGMENT

We thankfully acknowledge the support of the Hasler Foundation (grant no. 16076, S.A.V.E.) for this work.

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," *Information Fusion*, vol. 33, 2017.

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

[3] F. Dümbgen, M. E. Helou, N. Gucevska, and S. Süsstrunk, "Near-infrared fusion for photorealistic image dehazing," *Electronic Imaging*, no. 16, 2018.

[4] J. Ma, C. Chen, C. Li, and J. Huang, "Infrared and visible image fusion via gradient transfer and total variation minimization," *Information Fusion*, vol. 31, 2016.

[5] F. Lahoud and S. Susstrunk, "Ar in vr: Simulating infrared augmented vision," in *International Conference on Image Processing*. IEEE, 2018.

[6] J. Du, W. Li, K. Lu, and B. Xiao, "An overview of multi-modal medical image fusion," *Neurocomputing*, vol. 215, 2016.

[7] R. Garg, P. Gupta, and H. Kaur, "Survey on multi-focus image fusion algorithms," in *Recent Advances in Engineering and Computational Sciences*. IEEE, 2014.

[8] M. El Helou, Z. Sadeghipoor, and S. Süsstrunk, "Correlation-based deblurring leveraging multispectral chromatic aberration in color and near-infrared joint acquisition," in *International Conference on Image Processing*. IEEE, 2017.

[9] K. Ma, H. Li, H. Yong, Z. Wang, D. Meng, and L. Zhang, "Robust multi-exposure image fusion: a structural patch decomposition approach," *Transactions on Image Processing*, vol. 26, no. 5, 2017.

[10] K. Ram Prabhakar, V. Sai Srikar, and R. Venkatesh Babu, "Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs," in *IEEE International Conference on Computer Vision*, 2017.

[11] A. Toet, "A morphological pyramidal image decomposition," *Pattern Recognition Letters*, vol. 9, no. 4, 1989.

[12] R. Singh and A. Khare, "Fusion of multimodal medical images using daubechies complex wavelet transform–a multiresolution approach," *Information Fusion*, vol. 19, 2014.

[13] Y. Liu, S. Liu, and Z. Wang, "A general framework for image fusion based on multi-scale transform and sparse representation," *Information Fusion*, vol. 24, 2015.

[14] Y. Liu, X. Chen, R. K. Ward, and Z. J. Wang, "Image fusion with convolutional sparse representation," *IEEE Signal Processing Letters*, vol. 23, no. 12, 2016.

[15] N. Wang, Y. Ma, K. Zhan, and M. Yuan, "Multimodal medical image fusion framework based on simplified pcnn in nonsubsampled contourlet transform domain," *Journal of Multimedia*, vol. 8, no. 3, 2013.

[16] 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 Transactions on Instrumentation and Measurement*, no. 99, 2018.

[17] Y. Liu, X. Chen, H. Peng, and Z. Wang, "Multi-focus image fusion with a deep convolutional neural network," *Information Fusion*, vol. 36, 2017.

[18] H. Li, X.-J. Wu, and J. Kittler, "Infrared and visible image fusion using a deep learning framework," in *IEEE International Conference on Pattern Recognition*, 2018.

[19] J. Du, W. Li, B. Xiao, and Q. Nawaz, "Union laplacian pyramid with multiple features for medical image fusion," *Neurocomputing*, vol. 194, 2016.

[20] G. Qu, D. Zhang, and P. Yan, "Medical image fusion by wavelet transform modulus maxima," *Optics Express*, vol. 9, no. 4, 2001.

[21] S. Li, X. Kang, and J. Hu, "Image fusion with guided filtering," *IEEE Transactions on Image Processing*, vol. 22, no. 7, 2013.

[22] B. S. Kumar, "Image fusion based on pixel significance using cross bilateral filter," *Signal, Image and Video Processing*, vol. 9, no. 5, 2015.

[23] S. Das and M. K. Kundu, "Nsct-based multimodal medical image fusion using pulse-coupled neural network and modified spatial frequency," *Medical & Biological Engineering & Computing*, vol. 50, no. 10, 2012.

[24] G. Bhatnagar, Q. J. Wu, and Z. Liu, "Directive contrast based multimodal medical image fusion in nsct domain," *IEEE Transactions on Multimedia*, vol. 15, 2013.

[25] Q. Zhang, Y. Fu, H. Li, and J. Zou, "Dictionary learning method for joint sparse representation-based image fusion," *Optical Engineering*, vol. 52, no. 5, 2013.

[26] C. Liu, Y. Qi, and W. Ding, "Infrared and visible image fusion method based on saliency detection in sparse domain," *Infrared Physics & Technology*, vol. 83, 2017.

[27] S. Li, J. T. Kwok, and Y. Wang, "Multifocus image fusion using artificial neural networks," *Pattern Recognition Letters*, vol. 23, no. 8, 2002.

[28] Y. Liu, X. Chen, J. Cheng, and H. Peng, "A medical image fusion method based on convolutional neural networks," in *IEEE Fusion*, 2017.

[29] Z. Farbman, R. Fattal, D. Lischinski, and R. Szeliski, "Edge-preserving decompositions for multi-scale tone and detail manipulation," in *Transactions on Graphics*, vol. 27, no. 3. ACM, 2008.

[30] K. Subr, C. Soler, and F. Durand, "Edge-preserving multiscale image decomposition based on local extrema," in *Transactions on Graphics*, vol. 28, no. 5. ACM, 2009.

[31] E. S. Gastal and M. M. Oliveira, "Domain transform for edge-aware image and video processing," in *Transactions on Graphics*, vol. 30, no. 4. ACM, 2011.

[32] D. P. Bavirisetti and R. Dhuli, "Two-scale image fusion of visible and infrared images using saliency detection," *Infrared Physics & Technology*, vol. 76, pp. 52–64, 2016.

[33] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images." in *International Conference on Computer Vision*, vol. 98, no. 1. IEEE, 1998.

[34] J. Hu and S. Li, "The multiscale directional bilateral filter and its application to multisensor image fusion," *Information Fusion*, vol. 13, no. 3, 2012.

[35] Z. Xu, "Medical image fusion using multi-level local extrema," *Information Fusion*, vol. 19, 2014.

[36] Y. Zhai and M. Shah, "Visual attention detection in video sequences using spatiotemporal cues," in *International Conference on Multimedia*. ACM, 2006.

[37] R. Achanta, S. Hemami, F. Estrada, and S. Süsstrunk, "Frequency-tuned salient region detection," in *International conference on computer vision and pattern recognition*. IEEE, 2009.

[38] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S.-M. Hu, "Global contrast based salient region detection," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, 2015.

[39] R. Wu, Y. Yu, and W. Wang, "Scale: Supervised and cascaded laplacian eigenmaps for visual object recognition based on nearest neighbors," in *Conference on Computer Vision and Pattern Recognition*. IEEE, 2013.

[40] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *Conference on Computer Vision and Pattern Recognition*. IEEE, 2013.

[41] J. Ma, Z. Zhou, B. Wang, and H. Zong, "Infrared and visible image fusion based on visual saliency map and weighted least square optimization," *Infrared Physics & Technology*, vol. 82, 2017.

[42] K. He, J. Sun, and X. Tang, "Guided image filtering," in *European conference on Computer Vision*. Springer, 2010, pp. 1–14.

[43] G. Petschnigg, R. Szeliski, M. Agrawala, M. Cohen, H. Hoppe, and K. Toyama, "Digital photography with flash and no-flash image pairs," *ACM Transactions on Graphics*, vol. 23, no. 3, 2004.

[44] Z. Li, J. Zheng, Z. Zhu, W. Yao, and S. Wu, "Weighted guided image filtering," *IEEE Transactions on Image Processing*, vol. 24, no. 1, 2015.

[45] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *arXiv preprint arXiv:1312.6229*, 2013.

[46] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[47] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Hypercolumns for object segmentation and fine-grained localization," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[48] A. Toet *et al.*, "Tno image fusion dataset," *Figshare. data*, 2014.

[49] D. Summers, "Harvard whole brain atlas: www.med.harvard.edu/aanlib/home.html," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 74, no. 3, 2003.

[50] M. Nejati, S. Samavi, and S. Shirani, "Multi-focus image fusion using dictionary-based sparse representation," *Information Fusion*, vol. 25, 2015.

[51] Y. Liu, S. Liu, and Z. Wang, "Multi-focus image fusion with dense sift," *Information Fusion*, vol. 23, 2015.

[52] Y. Zhang, X. Bai, and T. Wang, "Boundary finding based multi-focus image fusion through multi-scale morphological focus-measure," *Information fusion*, vol. 35, 2017.

[53] Y. Han, Y. Cai, Y. Cao, and X. Xu, "A new image fusion performance metric based on visual information fidelity," *Information fusion*, vol. 14, no. 2, 2013.

[54] M. Hossny, S. Nahavandi, and D. Creighton, "Comments on information measure for performance of image fusion," *Electronics letters*, vol. 44, no. 18, 2008.

[55] C. Xydeas, , and V. Petrovic, "Objective image fusion performance measure," *Electronics letters*, vol. 36, no. 4, 2000.

[56] C. Yang, J.-Q. Zhang, X.-R. Wang, and X. Liu, "A novel similarity based quality metric for image fusion," *Information Fusion*, vol. 9, no. 2, 2008.

[57] N. Cvejic, A. Loza, D. Bull, and N. Canagarajah, "A similarity metric for assessment of image fusion algorithms," *International Journal of Signal Processing*, vol. 2, no. 3, 2005.

[58] J. Zhao, R. Laganiere, and Z. Liu, "Performance assessment of combinative pixel-level image fusion based on an absolute feature measurement," *International Journal of Innovative Computing, Information and Control*, vol. 3, no. 6, 2007.

[59] L. Xu, C. Lu, Y. Xu, and J. Jia, "Image smoothing via l 0 gradient minimization," in *Transactions on Graphics*, vol. 30, no. 6. ACM, 2011.

[60] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE conference on computer vision and pattern recognition*. IEEE, 2009.

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

[62] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012.

[63] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[64] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Conference on Computer Vision and Pattern Recognition*, 2017.

[65] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[66] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Conference on Computer Vision and Pattern Recognition*. IEEE, 2018.

[67] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size," *arXiv preprint arXiv:1602.07360*, 2016.

[68] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a siamese time delay neural network," in *Advances in Neural Information Processing Systems*, 1994.

[69] A. Rahman, D. Houzet, D. Pellerin, S. Marat, and N. Guyader, "Parallel implementation of a spatio-temporal visual saliency model," *Journal of Real-Time Image Processing*, vol. 6, no. 1, 2011.

[70] L. Dai, M. Yuan, Z. Li, X. Zhang, and J. Tang, "Hardware-efficient guided image filtering for multi-label problem," in *Conference on Computer Vision and Pattern Recognition*. IEEE, 2017.

[71] H. Wu, S. Zheng, J. Zhang, and K. Huang, "Fast end-to-end trainable guided filter," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2018.

[72] 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, 2018.

Fayez Lahoud received his B.E. in Computer and Communication Engineering and his minor in Mathematics from the American University of Beirut, Lebanon. He received his M.S. in Computer Science at École Polytechnique Fédérale de Lausanne where he is currently pursuing his Ph.D. at the Image and Visual Representation Laboratory. His work focuses on the development of computational and visual tools to help firefighters accomplish their tasks more efficiently.

Sabine Süsstrunk leads the Images and Visual Representation Lab (IVRL) at EPFL, Switzerland. Her research areas are in computational photography, color computer vision and color image processing, image quality, and computational aesthetics. She has published over 150 scientific papers, of which 7 have received best paper/demos awards, and holds 10 patents. She received the IS&T/SPIE 2013 Electronic Imaging Scientist of the Year Award and IS&Ts 2018 Raymond C. Bowman Award. She is a Fellow of IEEE and IS&T.