A Convolutional Neural Network-based Conditional Random Field model for Structured Multi-Focus Image Fusion Robust to Noise

Odysseas Bouzos, Ioannis Andreadis, and Nikolaos Mitianoudis

***Abstract*—The limited depth of field of optical lenses, makes multi-focus image fusion (MFIF) algorithms of vital importance. Lately, Convolutional Neural Networks (CNN) have been widely adopted in MFIF methods, however their predictions mostly lack structure and are limited by the size of the receptive field. Moreover, since images have noise due to various sources, the development of MFIF methods robust to image noise is required. A novel robust to noise Convolutional Neural Network-based Conditional Random Field (mf-CNNCRF) model is introduced. The model takes advantage of the powerful mapping between input and output of CNN networks and the long range interactions of the CRF models in order to reach structured inference. Rich priors for both unary and smoothness terms are learned by training CNN networks. The α-expansion graph-cut algorithm is used to reach structured inference for MFIF. A new dataset, which includes clean and noisy image pairs, is introduced and is used to train the networks of both CRF terms. A low-light MFIF dataset is also developed to demonstrate real-life noise introduced by the camera sensor. Qualitative and quantitative evaluation prove that mf-CNNCRF outperforms state-of-the-art MFIF methods for clean and noisy input images, while being more robust to different noise types without requiring prior knowledge of noise.**

***Index Terms*—Convolutional Neural Network, Conditional Random Field (CRF), multi-focus image fusion, Energy minimization**

I. Introduction

THE creation of all-in-focus images is of great importance for both human visual perception and computer vision tasks. However, due to the limited Depth-of-Field of optical lenses, only objects within a certain distance from the camera sensor can be well focused each time. The parts of the scene that lie outside the focal plane of the camera sensor remain out-of-focus or blurred. Multi-focus image fusion (MFIF) algorithms can be used to cope with the problem of finite depth of field of optical lenses, by merging multiple input images, which are captured with different focal settings, in a single fused image with extended depth of field. The fused image should have higher visual quality than each one of the input images without introducing artifacts during fusion. Moreover, since real world images contain noise, such as sensor noise and quantization noise, MFIF methods robust to different noise types are important. Lately, a great amount of MFIF-methods has been developed, which according to the recent survey of Liu et al. [1] can be classified in four major categories: *transform domain methods*, *spatial domain methods*, *combined methods* and *deep learning methods*.

Transform domain-based MFIF methods use a forward transform to decompose input images to their respective transform domain representations, which are then fused with custom hand-crafted and predefined fusion rules. Finally, the inverse transform is applied to the fused transform coefficients in order to obtain the final fused image. Since the quality of the fused images, is highly affected by the transform domain selection and the manual design of the fusion rules, a great number of transform domain-based MFIF algorithms has been introduced. Popular transform domain MFIF algorithms include: multi-scale decomposition-based methods [2]–[6], sparse representation-based methods [7]–[9], gradient domain-based methods [10], [11], methods based on other transform domains [12], [13] and methods that combine different transforms [14], [15]. Li et al. [16] introduced a multi-focus image fusion method that used nonsubsampled contourlet transform and residual removal. Their method outperformed some state-of-the-art methods. Dictionary-based domain methods, that incorporate a coefficient shrinkage strategy, such as [17], are robust against Gaussian noise. However, explicit knowledge about the noise characteristics is required in order to successfully use the coefficient shrinkage strategy, such as [17]. This limits the generalization capabilities of transform domain-based methods to effectively fuse real input images, which contain noise but prior information of noise characteristics is unavailable. The imperfect forward-backward transforms, result to blocking and ringing artifacts, due to the Gibbs phenomenon. Finally, since most of the transform domain-based MFIF methods are not shift-invariant, possible mis-registration found in the input images, due to dynamic scene or camera shake, will result to visible artifacts in the fused images.

In spatial domain MFIF methods, the fused image is estimated as the weighted average of the input images. Based on the adopted activity level estimations, weight maps are constructed and are used to fuse the input images. Spatial domain methods, can be categorized according to the method used for activity level estimation, as block-based, region-based and pixel-based. In block-based methods, the input images are decomposed in blocks of fixed size and activity level estimations of the whole blocks are used to construct the weight maps. The size of the block greatly affects the quality of the fused images. Region-based methods, provide

The authors are with the Department of Electrical and Computer Engineering, Democritus University of Thrace, 67100 Xanthi, Greece.
E-mail: obouzos@ee.duth.gr, iandread@ee.duth.gr, nmitiano@ee.duth.gr
Manuscript received 6.6.2022

higher flexibility than block-based methods, because they estimate the activity measurement for a region of irregular size. However, since blocks and regions are likely to simultaneously contain both well-focused and out-of-focus pixels, artifacts near boundaries between well-focused and blurred regions are likely to appear in both cases. In order to cope with this problem, pixel-based methods have become more popular, due to the use of pixel-level activity estimation. Pixel-based methods have higher accuracy near the boundaries of well-focused and out-of-focus pixels, however they are likely to produce noisy weight maps, which also deteriorate the quality of the fused images. A major drawback of spatial domain-based MFIF methods is their sensitivity to image noise.

Representative spatial domain-based MFIF methods include: Dense-SIFT (DSIFT) [18], Image matting (IM) [19], Guided filtering (GF) [20], Quadtree-based [21], Boundary Finding (BF) [22] and Cross Bilateral Filtering (CBF) [23].

In order to preserve advantages of both transform and spatial domains, combined-based MFIF methods have emerged. Bouzos et al. [24] proposed a Conditional Random Field model, which combined the advantages of both the transform domain Independent Component Analysis (ICA) and the spatial domain introducing mf-CRF. Yang et al. [25] combined the advantages of both nonsubsampled contourlet transform (NSCT) and spatial domain. Wang et al. [26] used a pulse coupled neural network (PCNN) and a guided filter in order to solve the MFIF problem. Li et al. [27] introduced a joint image fusion and denoising method that combined image decomposition and sparse representation. Chen et al. [28] combined an Image Matting strategy with top-hat and bottom-hat transforms in order to develop their model MGIM. Combined methods are likely to perform better than transform domain and spatial domain methods.

The performance of conventional MFIF methods, is limited by the hand-crafted features and the manual design of fusion rules, that can not fully model the complexity of the MFIF problem. This led to the increased popularity of deep learning-based methods for MFIF. The fact that deep learning-based methods do not require the hand-crafted design of features for focus measurement or the manual design of fusion rules, makes them likely to produce fused images of higher quality than conventional MFIF methods [29].

In [29], Zhang made an extensive study of deep learning methods for MFIF and classified them in two major categories: *Decision map-based methods* and *End-to-end methods*.

1) Decision-map based methods: In *decision map-based methods*, the network predicts a decision map according to the activity levels of the input images. Then, post processing methods are usually applied to refine the predicted decision maps. Lastly, the final decision map is used to guide the fusion of the input images.

Liu et al. [30] were the first to propose a CNN-based network for MFIF. More precisely, they trained a siamese architecture to classify pairs of image patches, as well-focused or out-of-focus. This block-based classification lead to the prediction of the labels of the decision map, which were used to guide the fusion of the MFIF input image. Since then, various decision-map based deep learning MFIF methods have been proposed in order to improve the prediction accuracy of the decision map and thus the performance of MFIF. Typical decision-map based methods include: P-CNN [31], Ensemble-CNN [32], MSCNN [33] and fully convolutional network FCN [34]. In order to improve the quality of the predicted decision map, Li et al. [35] used the complimentary information of the input image pairs and introduced DRPL which is a pixel-based approach. Ma et al. [36] proposed MMF-Net, which consists of two networks. The first network is used to extract an initial prediction, while the second is used to improve the decision boundary between well-focused and out-of-focus pixels. Xiao et al. [37] introduced GEU-Net, which used a U-Net architecture in order to estimate the focus maps as a global two-class classification problem. Decision-map methods have two major issues: 1) the block-based approaches like CNN-Fusion [30] have lower accuracy near the boundaries between well-focused and out-of-focus pixels, since blocks may contain simultaneously both well focused and out-of-focus pixels. 2) pixel based classification networks, such as DRLP [35], have better accuracy near boundaries, however are likely to produce noisy decision maps.

Decision map-based MFIF deep learning methods usually adopt post-processing steps in order to refine the predicted decision maps and remove noisy regions. CNN-Fusion [30] used Guided filter [38] to refine the decision maps. FCN [34] and GEU-Net [37] used fully connected Conditional Random Fields [39] to refine the predicted decision maps. Morphological operations and consistency verification methods are also widely used as post-processing steps. Although post-processing steps may remove noisy predicted regions from the decision map, they are also likely to decrease the quality of the final fused images [35]. Liu et al. [40] introduced the MSFIN network, which is a multiscale feature interactive network and they used a fully-connected conditional random field as post processing in order to refine their decision map.

2) End-to-end: In the end-to-end deep learning MFIF methods, the network is trained with regression optimization in order to learn the mapping between the input images and the target image, without the intermediate step of predicting the decision map.

Xu et al. [41] introduced the unified Image Fusion Network (U2Fusion), which used the DenseNet architecture and was trained with unsupervised learning. Li et al. [42] used a U-net architecture to directly predict the fused image from the input images. DenseFuse [43] and IFCNN [44] firstly trained an autoencoder. The encoder was then applied to both input images followed by hand-crafted fusion rules, in order to combine the deep feature coefficients. Lastly a decoder was applied to return the fused image. In [45], the authors trained an autoencoder network and introduced multiple image fusion strategies. The use of hand-crafted fusion rules is likely to reduce the image quality of the fused image. Zhao et al. [46] proposed MLCNN, an end-to-end deep learning network for MFIF with enhancement. Ma et al. [47] introduced an end-to-end image fusion framework based on the Swin Transformer. Cheng et al. [48] proposed a MUFusion architecture, which is based on a memory unit architecture. This unit utilized intermediate fused results during training. The imperfect

forward and backward transforms between spatial and deep feature domains are likely to result to fused images of lower quality, due to the Gibbs phenomenon. Moreover, end-to-end deep learning MFIF methods are not likely to preserve accurately the intensity and the contrast of the input images in the final fused image. Other issues rely on the sensitivity to mis-registration and the sensitivity to noise.

Some of the issues of deep learning based MFIF methods include sensitivity to noise, since noise scenarios are not studied and sensitivity of methods to possible mis-registration found in the input images. These problems are not widely considered in the deep learning based MFIF methods. Other problems include the lack of structure in the predicted decision map and the prediction being limited by the size of the receptive field. Lastly, since most deep learning based MFIF methods consider only pairs of input images, hierarchical fusion is needed to fuse more than two input images, which increases the complexity. To cope with the aforementioned issues, we propose mf-CNNCRF, which is a novel CNN-based CRF model. The proposed mf-CNNCRF combines advantages of CNN networks to learn rich priors and the long range interactions of CRF models in order to reach structured inference for MFIF. The use of efficient architectures allows arbitrary N images to be processed in parallel, making mf-CNNCRF computationally efficient. Lastly the proposed framework is robust to different types of noise for MFIF.

More specifically, since the proposed method combines advantages of both convolutional neural networks and energy minimization through graph-cut optimization, some recent related work that lies in these categories is briefly described.

Graphical models have been successfully applied for inference in MFIF methods, Their success lies mostly in the long range interactions and their close-to-global-optimum solution. Graphical models for MFIF were either used with explicitly defined priors or were used as a post processing step. More precisely, Sun et al. [11] and Bouzos et al. [24] introduced graphical models for MFIF, that used hand-crafted unary and smoothness priors. Thus, their performance was limited by the complexity of their hand-crafted priors. On the other hand GEU-Net [37] and FCN [34] used the fully connected CRF model [39] as a post processing step, in order to refine the binary decision maps, since they were predicted by their respective deep learning-based architectures. Hand crafted priors were also used to refine the decision maps. This approach limited the capabilities of the fully connected CRF models and there was error propagation from the CNN prediction to the final decision map. The application of graphical models in MFIF remains limited by the hand crafted priors.

Zagoruyko et al. [49] introduced three neural network architectures in order to learn similarity functions and compare input image patches, including siamese architectures. The success of siamese network architectures to compare input image patches, led us to the development of siamese architectures for both unary and smoothness terms. In order to cope with the fusion of an arbitrary number of N input images for MFIF, we introduce efficient siamese architectures to estimate both terms. The proposed architectures allow both terms to be efficiently estimated with low computational cost, since N input images can be processed in parallel. Details of the proposed efficient siamese networks 'UnaryNet' and 'SmoothnessNet' that estimate the Unary potential U and the Smoothness potential V respectively, are described in the following sub-sections. However, the prediction of the siamese architectures is limited by the size of the receptive field of each network.

In order to combine the advantages of both deep learning-based methods and CRF graphs, while overcoming their individual limitations, we introduce the CNN-based CRF model, named mf-CNNCRF for MFIF.

Compared to the aforementioned applications of CRF models, the proposed mf-CNNCRF framework is very different. Most importantly, both unary and smoothness priors are estimated through CNN architectures that are trained end-to-end. Thus, rich priors that better describe the mapping between input images and target image, more suitable for MFIF are developed. The predicted unary and smoothness priors provide complimentary information to the CRF model, in order to reach global optimal or close to optimal structured inference for MFIF, through solving the energy mininisation problem with α-expansion [50], which is based on Graph cuts. The proposed mf-CNNCRF method is a decision-map based method.

The main contributions of this manuscript can be summarized as follows:

1) The proposed mf-CNNCRF model, combines advantages of the complex mapping between input-output of CNN networks and the long range interactions of CRF model in order to reach structured inference for MFIF, without requiring further post-processing. Both UnaryNet and SmoothnessNet are efficient siamese networks trained end-to-end with CNN architectures of low complexity in order to learn rich-complex priors for MFIF. Both networks provide considerable speedup for handling arbitrary N input images and support commutativity of the input images.
2) The proposed mf-CNNCRF model was trained on a new synthetic MFIF dataset, which contained both clean images and images contaminated with Gaussian noise, Salt and Pepper noise and Poisson noise. The proposed loss functions along with the proposed dataset, which were used to train both Unary and Smoothness terms, make mf-CNNCRF robust to different noise types without requiring prior knowledge about the noise characteristics of the input images. Thus, mf-CNNCRF has great generalization capabilities for both clean input images and input images that contain different types of noise.
3) The major novelty of the proposed energy minimization approach compared to our previous work mf-CRF [24], lies in the use of CNN networks in order to model the complex relations of inputs/outputs, while previous work used carefully handcrafted priors to model the unary and smoothness terms in order to solve the multi-focus image fusion problem.
4) Since SmoothnessNet leads to pairwise smoothness priors, efficient pairwise solvers can be used, instead

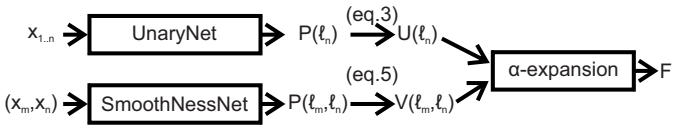

Fig. 1. Framework of mf-CNNCRF for the estimation of fused image F for N input images $x_1, \ldots, x_N$ and for their concatenations (x_m, x_n)

of high order CRF solvers that would have increased complexity.

To the best of our knowledge, training of inference techniques with CNNs has not yet been demonstrated for multi-focus image fusion. This is the first time that Convolutional Neural Networks are used to train the Unary and Smoothness terms of inference method based on graph cuts in image fusion, in order to achieve structured inference for multi-focus image fusion.

II. Energy Minimization

The proposed MFIF framework is a decision-map based framework. For each location in the decision map D, we create one node in the CRF graph. The pairwise connections in the CRF graph are formed by connecting each node to the respective nodes that lie in the $N8$-neighborhood. Both the unary potential function U and the pairwise potential function V are estimated with CNNs that are trained end-to-end. More precisely, the 'UnaryNet', which learns the unary potential function U, and the 'SmoothnessNet', which learns the pairwise potential function V, using efficient siamese architectures. These networks allow an arbitrary number of input images N to be processed in parallel, providing high computational efficiency and acceleration for the estimation of both terms for MFIF. UnaryNet and SmoothnessNet support commutativity of the input images to the framework, since the branches of each network share the same architecture and the same weights. The use of CNN networks allows rich unary and smoothness potentials to be learned. The proposed energy minimization approach takes advantage of the rich potentials learned through the CNN networks and the long range interactions of the CRF model, in order to reach structured inference with global or close to global MFIF solution.

In order to estimate the labels ℓ of the decision map, we use the following energy minimization:

$$\ell_{1..N} = \underset{\ell_{1..N}}{\operatorname{argmin}} \left[\sum_{n=1}^{N_s} U_n(\ell_n) + \sum_{(m,n)\,\epsilon\, C} V_{m,n}(\ell_m, \ell_n) \right] \quad (1)$$

where m, n are adjacent pixels in clique C, which is equal to a $N-8$ neighborhood and N_s is the total number of spatial locations. The whole procedure is summarised in Fig. 1.

The unary potential term U is used to estimate the significant contribution of each of the input images to the fused image. The smoothness potential term V is used to compute the label compatibility between adjacent pixels m, n of the decision map in the $N8$-neighborhood. Both the unary term U and the smoothness term V are trained end-to-end with efficient siamese CNN architectures of low complexity. The

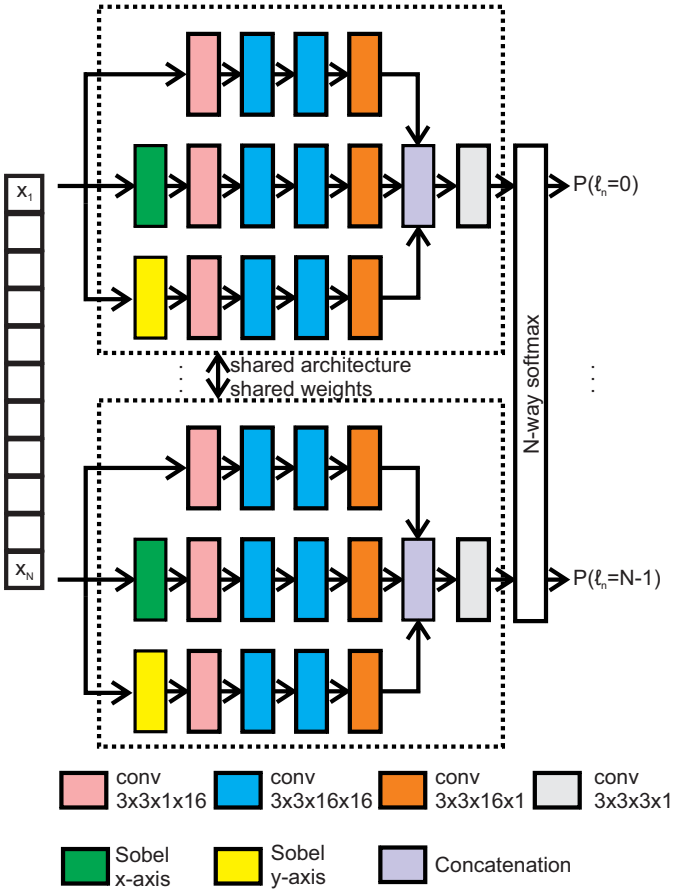

Fig. 2. Architecture of UnaryNet for N input images $x_1 \ldots x_N$

energy minimization problem is solved efficiently with the α-expansion algorithm [50], which is based on graph-cuts. Details of the proposed framework can be found in the following sub-sections.

Finally, for N input images and labels ℓ of the decision map, the fused image F is estimated as:

$$F_n = \begin{cases} x_1(n) & , \text{ if } \ell_n = 0 \\ \ldots & \ldots \\ x_N(n) & , \text{ if } \ell_n = N-1 \end{cases} \quad (2)$$

where n defines each spatial location.

A. Unary term and Unary network

A CNN network called 'UnaryNet' is trained through data in order to learn the probabilities $P(\ell_n)$, that each one of the input images ℓ_n should contribute to the final fused image at spatial location n. Fig. 2 demonstrates the architecture of the siamese network for N input images, used to estimate the probabilities $P(\ell_n)$. 'UnaryNet' is an efficient siamese architecture that uses N branches equal to the number of N input images. Each branch takes as input one of the N input images, while the output of each branch is passed to the decision layer. A N-way softmax function is used as the decision layer, in order to predict the probabilities $P(\ell_n)$. The efficient siamese architecture of 'UnaryNet' supports N images

that are processed in parallel and the probabilities $P(\ell_n)$ are estimated computationally efficiently. Finally, the Unary potential $U(\ell_n)$ is estimated as the negative log-likelihood of the predicted probabilities $P(\ell_n)$.

$$U(\ell_n) = -\log(P(\ell_n)) \tag{3}$$

The siamese branches of UnaryNet share the same architecture and the same weights. Each siamese branch is a ConvNet architecture with three branches. The first branch of the ConvNet consists of a convolutional layer with filter size $[3 \times 3 \times 1 \times 16]$ followed by ReLu. The next 2 convolutional layers have filter size $[3 \times 3 \times 16 \times 16]$ and each one is followed by ReLu. The last of the ConvNet branch convolutional layer has filter size $[3 \times 3 \times 16 \times 1]$. In the second branch the first convolutional layer is the Sobel gradient convolution filter 3×3 in horizontal axis, followed by a convolutional layer with filter size $[3 \times 3 \times 1 \times 16]$ followed by ReLu, 2 convolutional layer with filter size $[3 \times 3 \times 16 \times 16]$ followed by ReLu and a convolutional layer with filter size $[3 \times 3 \times 16 \times 1]$. In the third branch, the first convolutional layer is the Sobel gradient convolution filter 3×3 in vertical axis, followed by a convolutional layer with filter size $[3 \times 3 \times 1 \times 16]$ followed by ReLu, 2 convolutional layer with filter size $[3 \times 3 \times 16 \times 16]$ followed by ReLu and a convolutional layer with filter size $[3 \times 3 \times 16 \times 1]$. Finally, a depth concatenation layer is applied to collect the outputs of all three branches and a convolutional layer with filters $[3 \times 3 \times 3 \times 1]$ is applied to extract the final output for the given input image.

UnaryNet is trained through data in order to learn the probabilities efficiently through loss function minimization. Two branches of the UnaryNet are used to train the weights of the network on pairs of input images. The two branches share same architecture and same weights. The 'UnaryNet' learns probabilities that depend on the distance of pixel intensity and the distance of pixel features between the input images and the target images. Higher probabilities are learned for the pixels and edges of the input images that are similar to the respective ones of the target images, while lower probabilities are learned for the pixels and edges that are different from the respective target ones. A major advantage of 'UnaryNet' is that the probabilities are not hand-crafted but learned by the network, which is trained end-to-end.

For two input images x_1, x_2 and target image y the loss function $\mathcal{L}_U$ used to train UnaryNet is formulated as:

$$\begin{aligned}\mathcal{L}_U = &-\log(P(\ell=0))\left[\frac{|y-x_1|+c}{|y-x_0|+c}+\frac{|y_{mag}-x_{1mag}|+c}{|y_{mag}-x_{0mag}|+c}\right]- \\ &-\log(P(\ell=1))\left[\frac{|y-x_0|+c}{|y-x_1|+c}+\frac{|y_{mag}-x_{0mag}|+c}{|y_{mag}-x_{1mag}|+c}\right]\end{aligned} \tag{4}$$

where $x_{1mag}, x_{2mag}, y_{mag}$ are the Sobel magnitude of first, second and target images respectively and $c = 0.01$.

The proposed loss function allows the network to learn probabilities that are proportional to the distance of the pixel intensity and the distance of sobel magnitude between the input images and the target image. The loss function is pixel based in order to have better accuracy near the boundary of focused and out-of-focus pixels.

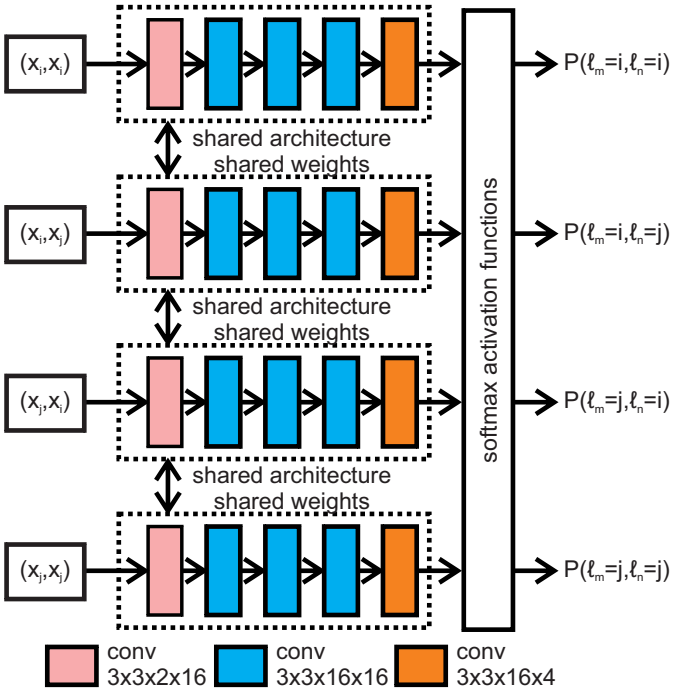

Fig. 3. Architecture of SmoothnessNet for input images $i, j \epsilon [1, N]$. (x_i, x_j) are depth concatenated input image pairs.

B. Smoothness term and Smoothness network

The goal of the smoothness term V_{pq} is to assign lower pairwise cost between adjacent pixels p, q that belong to the boundary of focused-defocused pixels, thus to the graph cut solution and higher pairwise cost between adjacent pixels p, q that are likely to belong to the same input image. In order to predict the label compatibility between adjacent pixels p, q in the N8-neighbourhood of the decision map, 'SmoothnessNet' is trained. The 'SmoothnessNet' is trained to assign high probabilities $P(\ell_p = \ell_q)$ to adjacent pixels p, q that are likely to belong to the same input image and high probabilities $P(\ell_p \neq \ell_q)$ to pixels that are likely to belong to different input images and thus to the graph-cut solution.

SmoothnessNet is an efficient siamese architecture that allows all M input image combinations of arbitrary N images, to be processed in parallel, in order to predict the label compatibility between adjacent pixels p, q for every label combination $\ell_p, \ell_q \epsilon [1, N]$.

Fig. 3 demonstrates the SmoothnessNet architecture used to predict the probabilities for label compatibility for two input images and all four label combinations. Each branch of SmoothnessNet takes as input one of the M concatenated image pair combinations of $i, j \epsilon [1, N]$, where N is the total number of input images. The branches of SmoothnessNet share same architecture and weights. Each branch consists of five convolutional layers. More precisely, the first convolutional layer has filters $[3 \times 3 \times 2 \times 16]$ followed by tanh, next three convolutional layers have filters $[3 \times 3 \times 16 \times 16]$ and each convolutional layer is followed by tanh, last convolutional

layer has filters $[3 \times 3 \times 16 \times 4]$ and provides the pairwise outputs for the 4 directions in the N8-neighborhood. The decision layer of 'SmoothnessNet' consists of 4 softmax activation functions. The pairwise outputs of all branches that correspond to same direction of the N8-neighborhood go through same softmax function. The probabilities $P\left(\ell_p, \ell_q\right)$ are predicted for all 4 directions, where p, q adjacent pixels in the N8 neighborhood. Lastly, the Smoothness term V which is a function of the predicted label probabilities is estimated as

$$V_{pq} = -\log\left[P\left(\ell_p \neq \ell_q\right)\right] \mathrm{dis}(p,q)^{-1} \mathbb{I}_{\ell_p \neq \ell_q} \tag{5}$$

where $\mathrm{dis}(p,q)$ is the euclidean distance between pixels p and q.

'SmoothnessNet' is trained on pairs of input images, thus four branches with inputs all image combinations are used to train the network weights. All branches of 'SmoothnessNet' share same architecture and weights.

The loss function used to train the SmoothnessNet is:

$$\begin{aligned}\mathcal{L}_v = &-\log\left(P\left(\ell_p \neq \ell_q\right)\right) \frac{\left|\nabla_{pq}^{00} x - \nabla_{pq} y\right| + \left|\nabla_{pq}^{11} x - \nabla_{pq} y\right|}{\left|\nabla_{pq}^{01} x - \nabla_{pq} y\right| + \left|\nabla_{pq}^{10} x - \nabla_{pq} y\right|} \\ &-\log\left(P\left(\ell_p = \ell_q\right)\right) \frac{\left|\nabla_{pq}^{01} x - \nabla_{pq} y\right| + \left|\nabla_{pq}^{10} x - \nabla_{pq} y\right|}{\left|\nabla_{pq}^{00} x - \nabla_{pq} y\right| + \left|\nabla_{pq}^{11} x - \nabla_{pq} y\right|}\end{aligned} \tag{6}$$

SmoothnessNet predicts all pairwise weights between adjacent pixels in the N8-neighborhood. The pairwise weights are estimated by processing pixels in a large neighborhood, equal to the receptive field of the Smoothness Network, however, the use of pairwise weights allows efficient graph solutions with pairwise solver to be used instead of graph cut solvers of high-order neighborhoods. Another major advantage of the proposed Smoothness network architecture is the use of efficient siamese architecture, which allows many input images to be processed simultaneously and thus the final smoothness probabilities.

III. Dataset Generation

The optimal way to train deep learning networks for MFIF would be a real world dataset of multi-focus image pairs with ground truth. Since such a dataset is not available, synthetic datasets have been developed in order to train the networks of deep learning-based MFIF methods. The creation of synthetic dataset for MFIF is an open research issue and thus different synthetic datasets have been introduced in order to train deep learning networks for MFIF.

In order to train the classification-based networks Cnn-Fusion [30], p-CNN [31], Ensemble-CNN [32], the developed datasets included whole patches that were either clean or blurred with Gaussian, in order to simulate the out-of-focus effect. However, the size of the patches in the aforementioned methods is limited to 16x16, and 32x32. In [35], the synthetic dataset used to train the DRPL network included input image pairs with both well-focused and out-of-focus pixels. In [44] the dataset was created using RGB-D image sets. A major issue of the aforementioned datasets is that they do not account for input image pairs corrupted with noise, which is present in real-world multi-focus images due to various sources. Thus the networks are likely to be sensitive to image noise.

In this paper, we introduce a novel synthetic dataset for MFIF that contains both clean image pairs and image pairs contaminated with different types of noise. More precisely the proposed dataset contains synthetic multi-focus image pairs without noise, with Gaussian noise, with Salt & Pepper noise and with Poisson noise.

The Pascal VOC 2012 dataset [51], which is used for the classification challenge, has been used in order to create the proposed synthetic dataset for our multi-focus image fusion problem. Thus, 2500 random images with their respective annotation maps were selected from the dataset. All images were resized to 224×224 using bicubic interpolation, while their respective annotation maps were resized to 224×224 using the nearest neighbor method. Nearest neighbor was preferred to resize the annotation maps, in order to preserve the same labels of the original annotation maps to the resized ones, and avoid the introduction of artificial categories due to interpolation. Afterwards, the labels of the annotation maps were reduced to 2 labels, foreground and background.

Using the new annotation maps with the two labels, two synthetic multi-focus images were created for each of the 2500 images. The first image preserved the background information (according to the annotation map) of the initial image, while the rest of the image was blurred with a Gaussian $\mathcal{N}\left(0, \sigma_1^2\right)$. The second image preserved the foreground information (according to the annotation map) of the initial image, while the rest of the image was blurred with a Gaussian $\mathcal{N}\left(0, \sigma_2^2\right)$. For each of the 2500 multi-focus image sets, the standard deviation of the Gaussians were randomly selected, $\sigma_1, \sigma_2 \epsilon\, [0.5, 5]$.

The clean synthetic dataset includes 2000 synthetic multi-focus image pairs which are used for training purposes, 250 pairs used for validation and 250 image pairs used for testing.

In order to better simulate the noise existing in real multi-focus images, due to various sources, such as sensor noise, image compression and image transmission, the images of training, validation and testing datasets were augmented with images contaminated with different types of noise. More precisely, three noise cases are used, additive Gaussian noise with $\sigma_n = [10, 20, 30]$, Salt & Pepper noise with density $d = [0.01, 0.03, 0.05]$ and Poisson noise.

All input image pairs of the synthetic datasets, are contaminated with noise, while the annotation map is used to guide the creation of the ground truth noisy image, by selecting the respective pixels from the noisy input images.

The final training dataset contains 16000 input image pairs with their respective ground truth fused images. The final validation dataset contains 2000 input image pairs with their respective ground truth fused results. The final testing dataset contains 2000 input image pairs with their respective ground truth fused results.

IV. Experimental results

First, we describe the training details and the data sets used for evaluation. Qualitative and quantitative results are then included in order to evaluate the performance of the proposed

and the compared state-of-the-art MFIF methods in both real world datasets and on synthetic datasets with Gaussian noise, Salt & Pepper noise and Poisson noise.

A. Training Details

The experiments included in this paper, are based on Mathworks Matlab R2019b, implemented on NVidia RTX 2080-8G with Max-Q Design, GPU graphics processing unit. The optimizer used is Adam with $\beta_1 = 0.9, \beta_2 = 0.99$ and learning rate 0.001. The batch size is set to 16 for both UnaryNet and SmoothnessNet. The training is done for 45 epochs. The developed code is available at https://github.com/obouzos/.

B. Datasets for experimental results

In order to evaluate the performance of the proposed mf-CNNCRF and the state-of-the-art MFIF methods, experiments on 4 real world datasets and on 3 synthetic datasets have been contacted. The included datasets are: The RGB dataset (Lytro dataset) which consists of 20 RGB multi-focus image pairs and can be found in [52], the grayscale dataset, which includes 17 grayscale multi-focus image pairs and can be found in [22], the MFFW dataset, that consists of 13 multi-focus image fusion pairs and can be found in [53], the low-light multi-focus dataset which is introduced in this paper and includes 10 image pairs that can be found in https://github.com/obouzos/ and the three synthetic datasets that contain Gaussian noise, Salt & Pepper noise and Poisson noise respectively.

In order to create the three synthetic datasets that contain noise for the experimental results, 50 random images along with their respective segmentation maps of the Pascal VOC 2012 dataset [51], which were not included in the training, validation or test set have been selected. The labels of the segmentation maps are reduced to two, background and foreground labels. Based on the segmentation map with the two labels, two synthetic multi-focus images are developed. The first image has foreground pixels well in focus while background pixels out-of-focus. The second image has background pixels well-focused and foreground pixels out-of-focus. The out-of-focus pixels have been developed by applying Gaussian blur with red($\mu = 0, \sigma \epsilon [0.5, 5]$). The synthetic image pairs are afterwards contaminated with different types of noise in order to create three datasets. For the creation of the 'Gaussian noise dataset', the image pairs are contaminated with Gaussian noise ($\mu = 0, \sigma = 10, 30, 50$) at same locations. For the development of 'Salt & Pepper noise' dataset, image pairs are contaminated with Salt & Pepper noise with density values $d = [0.01, 0.03, 0.05]$. Lastly, the 'Poisson noise' dataset is constructed by contaminating the input image pairs with Poisson noise. For every case, the ground truth image is created by applying the ground truth decision map (segmentation map with two labels) to the noisy input image pairs and selecting the respective well-focused input pixels from each input image.

C. Algorithm comparison

We compare mf-CNNCRF with 16 state-of-the-art MFIF methods including: the spatial domain-based methods: DSIFT [18] and GFDF [20], the transform domain-based methods: DCHWT [54], SIGPRO [16], the combined-based methods: mf-CRF [24], Joint [27], MGIM [28] and the deep learning-based methods: CNNFusion [30], SESF [55], IFCNN [44], ECNN [32], DRPL [35], MSFIN [40], UniFusion [45], SwinFusion [47] and MUFusion [48].

D. Experiments on Real-World datasets

The evaluation of fused images for the real world datasets, that do not have reference image, is not easy. In order to assess fused image quality, image metrics as described in [56] are employed for the RGB and grayscale datasets. According to Liu et al. [56] the image quality metrics to assess fused image quality are classified in four categories: (1) information theory-based metrics, (2) image feature-based metrics, (3) image structural similarity-based metrics and (4) human perception-based metrics. Four metrics, one from each category, are used in order to assess the fused image quality of the proposed and the state-of-the-art MFIF methods in the Lytro and grayscale datasets. The included metrics are: Mutual Information - MI, gradient-based fusion performance - Q_G, Piella's Metric - Q_P, Chen-Blum Metric - Q_{CB}. Higher values in the four metrics indicate better fused image quality.

1) RGB dataset: Fig. 4 shows the source images of the volleyball scene of the RGB dataset along with the fusion results of the compared methods and magnifications of two selected regions. Spatial domain methods DSIFT and GFDF cannot preserve accurately the well-focused pixels of the region in the red box, and the background is out-of-focus. Transform domain DCHWT and IFCNN have lower contrast than the original images. CNNFusion and SESF cannot preserve accurately the background in the region of the red box. SESF has visible artifacts near the shoe of the region in the green area. ECNN cannot preserve in focus the foreground and background in the region of the red box and has artifacts near the shoes in the region of the green box. DRPL has visible artifacts in the region of the green box as shown in the magnification. Joint has artifacts in the blue region. SIGPRO and UNIFusion can not preserve accurately the well focused areas in the red regions. The red region is not well focused in MSFIN. MGIM has artifacts on the shoe in the green region. SwinFusion and MUFusion have lower contrast than the original images and the back of the shoe in the green region is not well focused. mf-CRF and the proposed mf-CNNCRF preserve best the well focused pixels found in the source images for the Volleyball scene of the RGB dataset.

Table I includes the objective evaluation of the proposed method and state-of-the-art methods for the RGB dataset. The proposed method has the a) highest MI, indicating that during fusion, preserves best the information of the input images, b) the highest Q_G, i.e. preserves better the edges that are present in the input images, c) the second best Q_P, i.e. preserves better the structures of the input images and d) highest Q_{CB}, indicating better visual quality according to the human perception metric. Thus, mf-CNNCRF outperforms the state-of-the-art methods in the 'Lytro' dataset in most metrics.

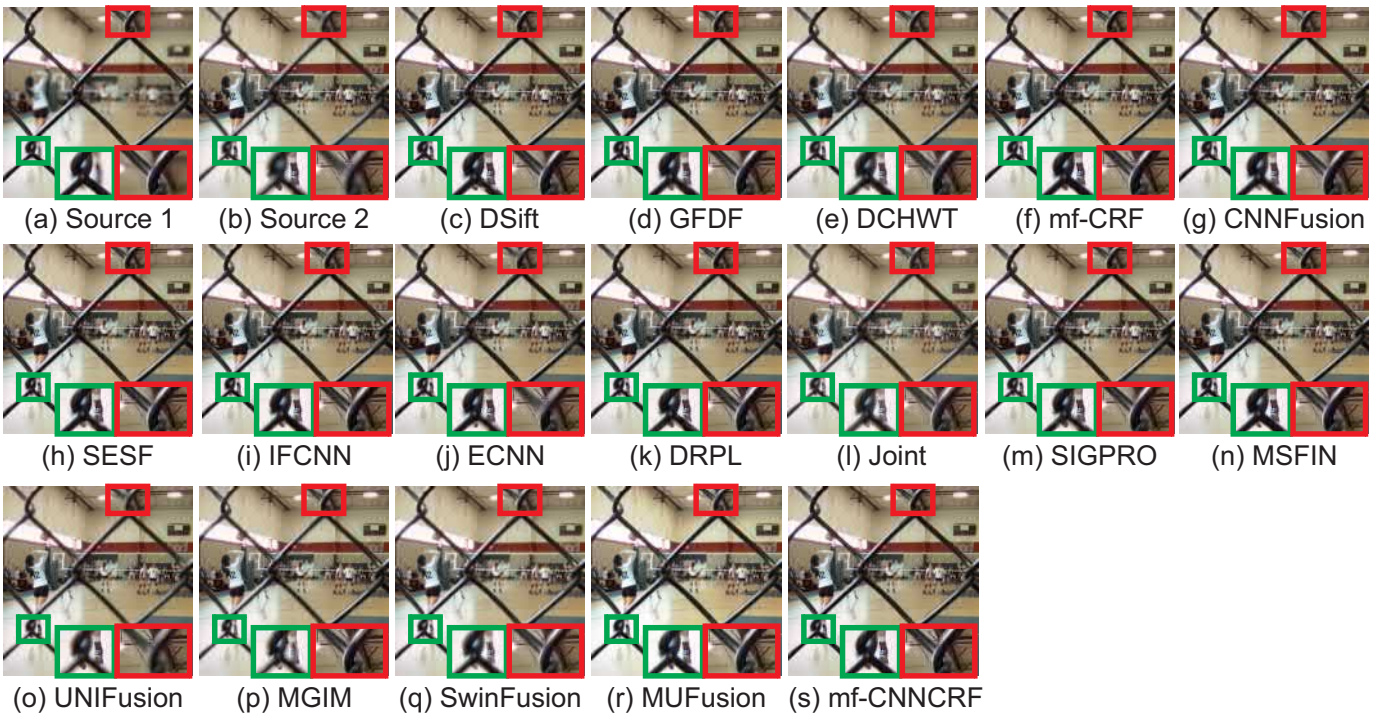

Fig. 4. Source and fused images for the 'Volleyball' scene of the RGB dataset

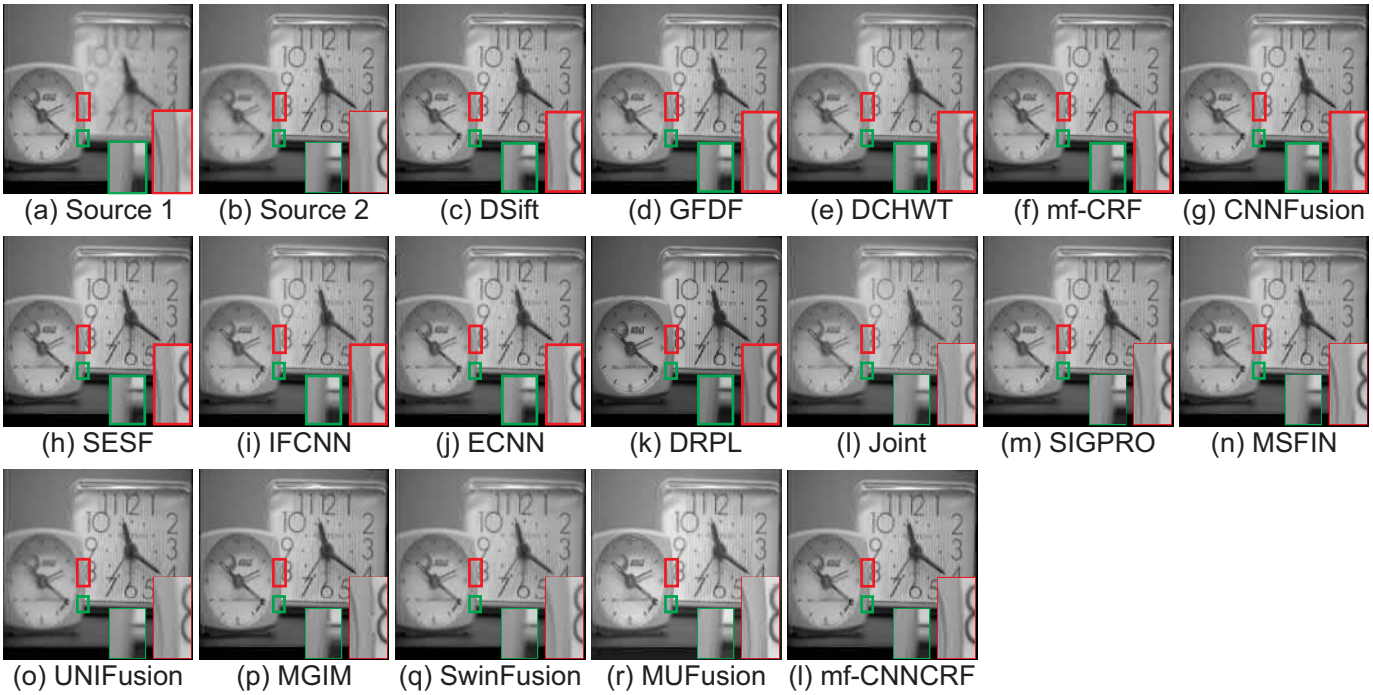

Fig. 5. Source and fused images 'Clocks' scene of the grayscale dataset.

2) Grayscale dataset: Fig. 5 demonstrates the two source images and the fused images for the 'Clocks' scene of the grayscale dataset. The spatial domain methods DSift and GFDF can not preserve accurately the clock boundary in both magnified areas. The transform domain method DCHWT can not preserve the boundary of the left clock in both magnified regions. mf-CRF preserves accurately the boundary in both regions. In CNNFusion and SESF, the boundary of the left clock is out of focus in both regions. In IFCNN there is artifact in the boundary in the region that corresponds to the red box. In ECNN the left clock boundary is out of focus in both magnified regions. In DRPL in the region that corresponds to the red region, there is artifact on the clock boundary. In Joint, SIGPRO, MSFIN, MGIMG the boundary of the clock in both

TABLE I
MEAN VALUES OF METRICS MI, QG, QP, QCB ON THE RGB DATASET.

Methods	MI	Q_G	Q_P	Q_{CB}
DSIFT [18]	8.9315	0.7633	0.8951	0.8092
GFDF [20]	8.7410	0.7633	**0.8975**	0.8114
DCHWT [54]	6.7298	0.7184	0.8893	0.6924
mf-CRF [24]	8.9506	0.7641	0.8967	0.8098
CNNFusion [30]	8.6420	0.7629	0.8963	0.8084
SESF [55]	8.6729	0.7611	0.8954	0.8049
IFCNN [44]	7.0400	0.7337	0.8934	0.7292
ECNN [32]	8.8784	0.7599	0.8861	0.8075
DRPL [35]	8.9324	0.7628	0.8935	0.8066
Joint [27]	6.9991	0.7435	0.8528	0.7176
Sigpro [16]	8.8518	0.7632	0.8964	0.8104
MSFIN [40]	8.9507	0.7639	0.8969	0.8099
UNIFusion [45]	7.3254	0.7403	0.8764	0.7410
MGIM [28]	8.8876	0.7553	0.8356	0.8000
SwinFusion [47]	6.5127	0.7118	0.8508	0.6725
MUFusion [48]	6.0381	0.6739	0.8385	0.6497
mf-CNNCRF	**8.9533**	**0.7641**	0.8972	**0.8121**

TABLE II
MEAN OF FOUR METRICS ON THE GRAYSCALE DATASET.

Methods	MI	Qg	Qp	Qcb
DSIFT [18]	8.6267	0.7405	0.8212	0.7851
GFDF [20]	8.3826	0.7402	0.8232	0.7840
DCHWT [54]	5.9965	0.6781	0.8096	0.6752
mf-CRF [24]	8.7034	0.7422	0.8210	0.7891
CNNFusion [30]	8.3190	0.7390	0.8233	0.7832
SESF [55]	8.4289	0.7402	0.8232	0.7828
IFCNN [44]	5.9641	0.6743	0.8063	0.6725
ECNN [32]	8.6543	0.7381	0.8094	0.7813
DRPL [35]	8.5751	0.7229	0.8067	0.7628
Joint [27]	6.7541	0.7212	0.8076	0.7234
Sigpro [16]	8.5721	0.7395	0.8173	0.7799
MSFIN [40]	8.6855	0.7365	0.8107	0.7863
UNIFusion [45]	7.0361	0.7112	0.7865	0.7214
MGIM [28]	8.5324	0.7231	0.7635	0.7658
SwinFusion [47]	5.8265	0.6806	0.7815	0.6602
MUFusion [48]	5.1646	0.5959	0.5130	0.7335
mf-CNNCRF	**8.7985**	**0.7431**	**0.8321**	**0.7904**

regions is not well preserved. In UniFusion and SwinFusion the boundary of the clock in green region is not well preserved. The boundary of the clock in the red region is not accurately preserved in MUFusion method. The boundary of both regions is accurately preserved in both regions by mf-CNNCRF.

Table II includes the objective evaluation of the proposed method and state of the art methods for the grayscale dataset [22]. In essence, mf-CNNCRF exhibits the highest MI for the grayscale dataset, and thus can preserve better the original information of the input images. The proposed method has the highest Q_G value and thus can preserve better the gradient information of the input images. Moreover, mf-CNNCRF preserves better the structures of the input images, since it has highest average Q_P value. Lastly, mf-CNNCRF has better fused image quality according to the human-perception metric Q_{CB}.

3) MFFW dataset: Table III includes the objective evaluation of the proposed and the compared methods. The proposed method has higher mean Mutual information value compared to the other methods, thus preserves the original better. The gradient information of the original images is preserved best in the proposed fused image, since mf-CNNCRF has higher mean Q_G value for the MFFW dataset. mf-CNNCRF ranks

TABLE III
MEAN VALUES OF METRICS MI, Q_G, Qp, Qcb ON THE MFFW DATASET.

Methods	MI	Q_G	Qp	Qcb
DSIFT [18]	8.2660	0.7434	0.8311	0.7335
GFDF [20]	7.8174	0.7444	**0.8458**	0.7526
DCHWT [54]	5.6786	0.6881	0.8389	0.6324
mf-CRF [24]	8.4046	0.7453	0.8373	0.7568
CNNFusion [30]	7.7264	0.7383	0.8400	0.7438
SESF [55]	7.7109	0.7419	0.8376	0.7397
IFCNN [44]	5.8065	0.6867	0.8258	0.6420
ECNN [32]	8.2431	0.7397	0.8286	0.7351
DRPL [35]	8.1611	0.7363	0.8253	0.7176
Joint [27]	6.3374	0.7107	0.7798	0.6616
Sigpro [16]	8.1812	0.7449	0.8398	0.7470
MSFIN [40]	8.3676	0.7370	0.8254	0.7420
UniFusion [45]	5.0203	0.6455	0.7893	0.6432
MGIM [28]	8.1964	0.7366	0.7838	0.7378
SwinFusion [47]	5.4657	0.6402	0.7845	0.6078
MUFusion [48]	5.1433	0.5850	0.7301	0.5717
mf-CNNCRF	**8.4133**	**0.7455**	0.8431	**0.7594**

second in the metric Q_P with very close mean value to GFDF method. Lastly, according the the human inspired metric Q_{CB} mf-CNNCRF has higher fused image quality than the state-of-the-art compared methods.

Fig. 6 demonstrates the source images and the fused images for the scene 'Flowers' of the MFFW dataset. Spatial domain methods DSift and GFDF have artifacts in the green region. DCHWT has artifacts around the flower in the green region and the boundaries are not well preserved in the blue region. SIGPRO has artifacts around the boundaries of the flower in the blue region. In mf-CRF the boundaries of the flower in blue region are not well preserved. Joint and MGIM methods have artifacts in both regions. In CNNFusion there are artifacts around the flower in the green region. SESF and IFCNN methods have artifacts in both regions. ECNN can not accurately preserve the boundaries of the flower in the green region. DRPL and MSFIN have artifacts in the blue region. UNIFusion, SwinFusion and MUFusion have artifacts in both regions.

4) Low light dataset: In low light environments, camera sensors produce dense noise in order to compensate for the low light. A dataset of 10 multi-focus image fusion pairs captured in low-light environment is introduced. These images have visible camera sensor noise, which was produced due to the low light available in each schene. The low light multi-focus image fusion dataset is available online at: https://github.com/obouzos/.

Fig. 7 demonstrates the source images and the fused images for the set from the low-light multi-focus dataset. Since the environment has low-light, the camera sensor noise is dense. Two regions were selected and magnified. Spatial domain methods DSIFT and GFDF fail to accurately preserve the boundaries of well focused objects in both regions. DCHWT, SIGPRO and mf-CRF can not preserve accurately the boundaries of the region in green. Joint and MGIM can not preserve accurately the cup boundaries in the red region, which remain out-of-focus. In CNNFusion and SESF, the cup in the red region is not well focused. SESF also cannot preserve well the object boundaries in the green region. IFCNN can not preserve well the well-focused object boundaries in both green and red

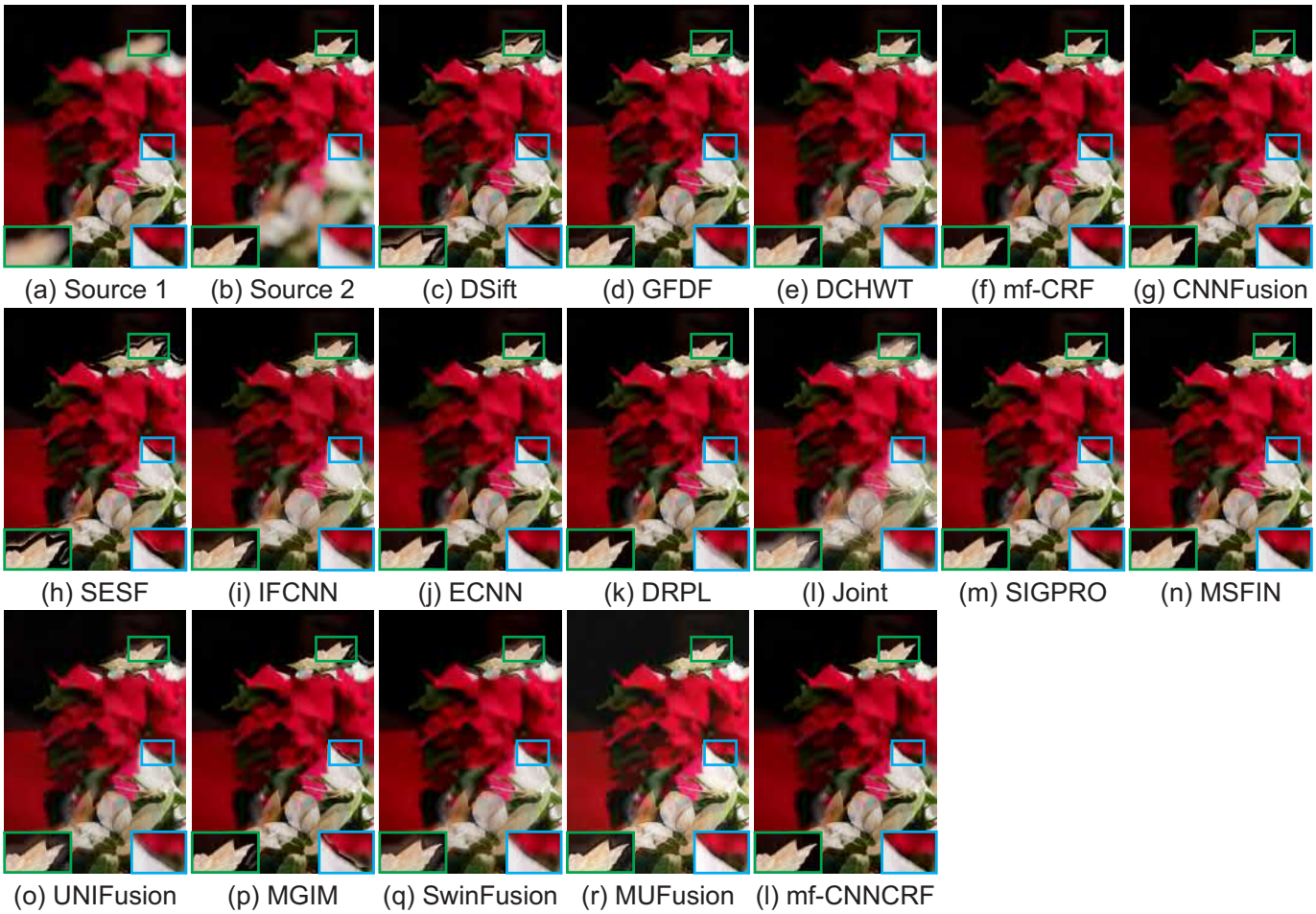

Fig. 6. Source and fused images for scene flowers of the MFFW dataset.

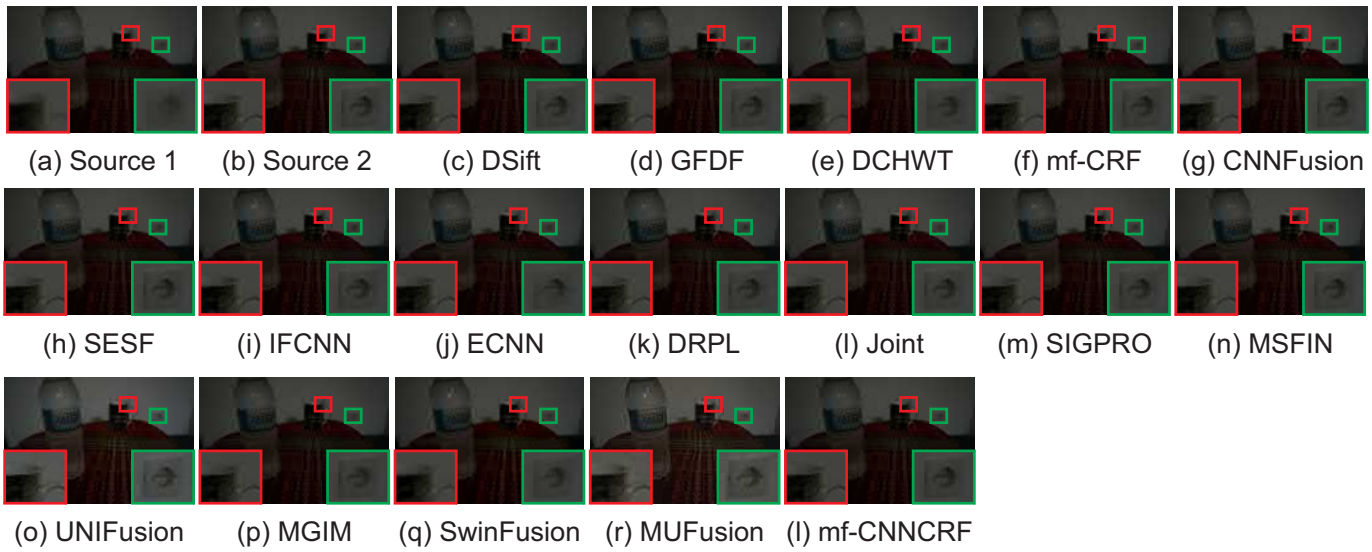

Fig. 7. Source and fused images from the developed low-light dataset.

regions. In ECNN and DRPL the socket in green region is out-of-focus. MSFIN can not preserve well the boundaries in both regions. UNIFusion introduces artifacts in the fused image and the boundaries of objects in both regions are not well focused. In SwinFusion, the red region is out-of-focus and part of the object in the green region is also out-of-focus. In MUFusion, the fused image has artifacts and the object boundaries in both regions are not well preserved. In the proposed method, the

TABLE IV
MEAN VALUES OF METRICS MI, QG, QCB ON THE LOW LIGHT DATASET.

Methods	MI	Q_g	Qcb
DSIFT [18]	7.5831	0.6997	0.7464
GFDF [20]	7.3641	0.7063	0.7456
DCHWT [54]	6.6218	0.6305	0.6697
mf-CRF [24]	7.5523	0.7004	0.7531
CNNFusion [30]	7.3637	0.7043	0.7417
SESF [55]	7.2753	0.6996	0.7306
IFCNN [44]	6.1427	0.6131	0.6782
ECNN [32]	7.4429	0.6949	0.7179
DRPL [35]	7.3200	0.6896	0.7033
Joint [27]	6.8412	0.6699	0.6981
SIGPRO [16]	7.3737	0.6979	0.7365
MSFIN [40]	7.5622	0.6975	0.7372
UNIFusion [45]	5.6853	0.3573	0.6432
MGIM [28]	7.6539	0.6918	0.7465
SwinFusion [47]	5.7312	0.5111	0.6234
MUFusion [48]	5.1575	0.3079	0.5833
mf-CNNCRF	**7.5872**	**0.7101**	**0.7556**

object boundaries in both regions are well-preserved and the fused image does not have artifacts.

Table IV includes the objective evaluation of the proposed and the compared methods. The proposed method has higher Mutual Information than the compared methods for the Low Light Dataset, thus preserves best the original information. Moreover, mf-CNNCRF exhibits highest value of Q_G, which indicates that preserves best the gradients of the original low light images. Lastly, mf-CNNCRF has highest Qcb value which is indicates that the proposed fused image has higher visual quality than the compared methods. The proposed method is more robust to the real-world noise of unidentified type compared to the other methods.

E. Experiments on Noisy Synthetic datasets

In order to evaluate mf-CNNCRF and state-of-the art methods in the presence of noise, two quality metrics are used, Root Mean Square Error (RMSE) and Peak Signal to Noise Ratio (PSNR).

Since ground truth image is available, $RMSE$ and $PSNR$ metrics are used to evaluate the quality of the fused images of the proposed and the state-of-the-art methods. Lower RMSE values and higher PSNR values indicate better fused image quality.

The UNIFusion method was retrained on the proposed dataset and the results on the noisy datasets are included for comparison.

1) Gaussian Noise evaluation: Fig. 8 depicts the source images and the fused images for input images corrupted with Gaussian noise with $\sigma = 30$. DSift cannot preserve the well focused region of the red box. GFDF can preserve both regions. In DCHWT, the image has lower contrast than the input images and can not preserve the well focused regions in the red area. The mf-CRF preserves both regions, however, requires prior knowledge of σ. CNNFusion and SESF can not preserve the well focused pixels of the red area. IFCNN has lower contrast compared to the input images. ECNN can not preserve the well focused pixels in red box area. In DRPL some of the well focused pixels in the area of the red box are out of focus. In Joint, SIGPRO, UNIFusion and retrained Unifusion the red region is not well focused. In MSFIN and MGIM both regions are out-of-focus. In SwinFusion, the red region is out-of-focus. In MUFusion red region is out-of-focus and fused image has more noise than the original images. The proposed mf-CNNCRF preserves accurately the well-focused pixels in both region, without requiring any prior knowledge about the noise and the noise characteristics.

TABLE V
RMSE/PSNR FOR COMPARED METHODS FOR GAUSSIAN NOISE $N\left(0, \sigma^2\right)$.

Methods	RMSE			PSNR		
	$\sigma = 10$	$\sigma = 30$	$\sigma = 50$	$\sigma = 10$	$\sigma = 30$	$\sigma = 50$
DSIFT [18]	1.5694	1.5591	1.5604	35.9909	35.4506	35.2891
GFDF [20]	1.5060	1.5443	1.5666	37.0070	36.7745	36.7274
DCHWT [54]	3.4690	4.1629	4.2734	32.4730	30.4008	29.8461
mf-CRF [24]	1.2860	1.3507	1.3937	38.0907	37.4330	36.7118
CNNFusion [30]	1.6829	1.6774	1.6654	36.1205	35.9824	35.9422
SESF [55]	1.5289	1.6520	1.6543	36.7794	36.2192	36.2759
IFCNN [44]	3.3746	5.9648	6.8673	32.1912	25.9303	23.4013
ECNN [32]	2.0064	3.1150	4.2832	32.9758	28.9209	25.8115
DRPL [35]	2.1795	3.6554	3.9212	35.4851	29.1489	28.0977
Joint [27]	1.4907	2.4706	2.8470	33.6382	31.7076	31.0473
Sigpro [16]	1.3252	1.5748	1.6880	37.4120	35.6770	34.7140
MSFIN [40]	2.1577	3.5311	3.6073	33.9450	28.5070	28.1089
UNIFusion [45]	4.5071	5.2038	5.0883	26.8822	25.5543	26.1245
UNIFusion (retrained)	4.4546	4.7466	4.6918	27.3591	27.2498	27.7961
MGIM [28]	1.9627	4.1260	3.9405	32.9954	26.0799	26.5090
SwinFusion [47]	1.9674	3.3456	5.2186	30.8878	28.2156	26.0017
MUFusion [48]	1.5928	2.2021	3.1983	21.3574	16.7232	14.3383
mf-CNNCRF	**0.5396**	**0.6717**	**0.7653**	**41.5071**	**40.4088**	**39.5173**

In order to evaluate the performance of fusion methods in the presence of Gaussian noise, additive Gaussian Noise with $\mu = 0$ and $\sigma = [10, 30, 50]$ is applied to the synthetic dataset of 50 image pairs. Gaussian Noise is added at the same positions in both input images of every set. Table V includes the RMSE and PSNR metrics comparison, in order to compare the proposed method with state-of-the-art MFIF methods. It is important to note that the proposed method does not have any prior knowledge about the σ^2, while the mf-CRF approach requires prior knowledge for the transform domain coefficient shrinkage. The proposed mf-CNNCRF has the lowest mean RMSE value and highest mean PSNR value for all $\sigma \epsilon \{10, 30, 50\}$, demonstrating the robustness of the proposed method in the presence of Gaussian noise.

Table V includes the RMSE and PSNR metrics in order to evaluate the performance of the proposed and state-of-the-art methods in the presence of Gaussian noise. Lower RMSE values indicate better image quality and fused image that is closer to the target fused image. The proposed method has the lowest RMSE value for all $\sigma = (10, 30, 50)$ and the fused image is closer to the target fused image than the compared methods. Higher PSNR values indicate better fused image quality. The proposed method has the highest PSNR value for all $\sigma = (10, 30, 50)$. Object evaluation demonstrates the robustness of mf-CNNCRF in the presence of Gaussian noise,

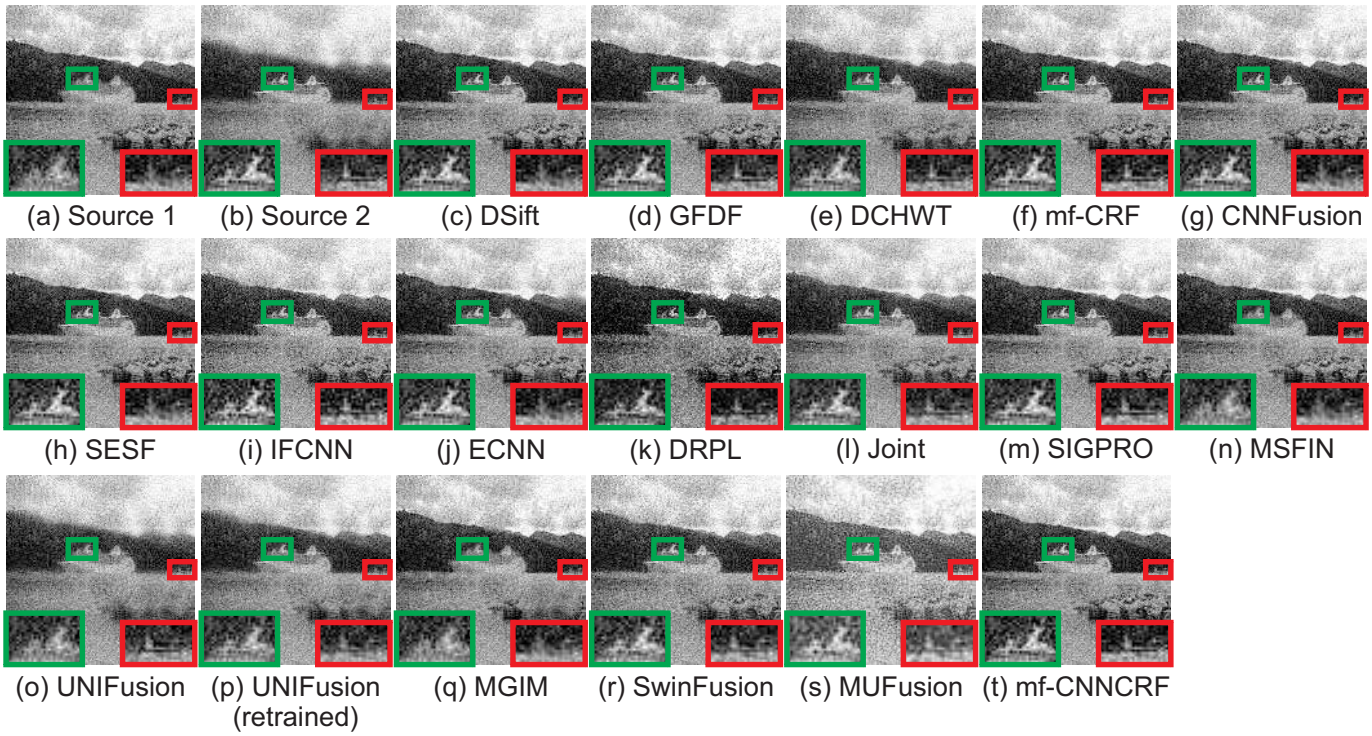

Fig. 8. Source and fused images for the Gaussian dataset for $\sigma = 30$.

without prior knowledge of the noise type and statistics.

2) Salt & Pepper evaluation: Fig. 9 includes the qualitative evaluation of the compared methods for Salt & Pepper noise with density $d = 0.05$. DSIFT cannot preserve accurately the well-focused pixels in the area with the green box. In GFDF and DCHWT, the fused image has lower contrast than the input images. In DCHWT, the area in red box is out of focus and are visible artifacts in the area of the green box. The mf-CRF can not preserve the focused pixels in both areas of the green and red boxes. In CNNFusion and SESF, both selected regions have out of focus pixels. In IFCNN, the fused image has more noise than each of the input images. In ECNN, both regions of the green and red boxes remain out of focus. In DRPL, there are artifacts in the area of the green box. In Joint, SwinFusion and MUFusion, the red region is not well focused. In SIGPRO, UNIFusion, retrained UNIFusion and MGIM the green region is not well focused. The image produced by MUFusion has lower contrast than the original images and the green region is out-of-focus. The proposed mf-CNNCRF preserves the well-focused pixels in both selected areas of the green and red boxes.

In order to evaluate the performance of fusion methods in the presence of Salt & Pepper noise, Salt & Pepper noise with density $d = [0.01, 0.03, 0.05]$ is applied to the clean image pairs and the ground-truth is extracted by selecting the respective pixels of input images based on the ground-truth binary map. Table VI includes the RMSE and PSNR values for densities $d = [0.01, 0.03, 0.05]$. The proposed mf-CNNCRF has the lowest RMSE value and highest PSNR value than the state-of-the-art MFIF methods. Thus, mf-CNNCRF is more robust to Salt & Pepper noise for densities $d = [0.01, 0.03, 0.05]$, compared to the state-of-the-art MFIF methods.

TABLE VI
RMSE/PSNR FOR COMPARED METHODS FOR SALT & PEPPER NOISE WITH DENSITY d.

Methods	RMSE			PSNR		
	$d = 0.01$	$d = 0.03$	$d = 0.05$	$d = 0.01$	$d = 0.03$	$d = 0.05$
DSIFT [18]	1.8015	2.2517	2.5080	27.0784	22.1538	20.2426
GFDF [20]	1.7128	2.2293	2.7115	28.9642	24.0510	21.6412
DCHWT [54]	4.3809	5.8072	6.4859	24.2867	20.5738	18.8345
mf-CRF [24]	1.8513	2.7776	3.0777	24.8823	20.1078	18.0188
CNNFusion [30]	2.1320	2.9551	3.5687	27.8879	22.7490	20.4593
SESF [55]	2.6945	3.5997	4.1030	24.4180	19.4385	17.3319
IFCNN [44]	3.6754	5.6483	6.7252	23.7389	19.4124	17.4541
ECNN [32]	2.5459	4.0012	4.8820	24.9360	18.9538	16.1941
DRPL [35]	1.3734	2.4688	3.1398	23.4983	18.9296	16.8639
Joint [27]	2.8903	4.0363	4.6225	24.1998	20.6993	19.0134
Sigpro [16]	1.3926	1.5614	1.8350	30.9404	27.2993	24.3097
MSFIN [40]	1.1142	1.5945	2.0632	30.8606	25.1948	23.3883
UNIFusion [45]	2.4086	2.8693	3.3102	30.2885	26.4748	23.9677
UNIFusion (retrained)	3.0456	3.5387	3.9595	28.1812	25.0043	22.9938
MGIM [28]	2.0248	2.4053	2.6948	26.1130	22.0677	20.3552
SwinFusion [47]	2.5273	3.5622	4.3311	24.9633	21.1266	19.2653
MUFusion [48]	1.7963	2.5323	3.1271	17.4746	14.7837	13.6316
mf-CNNCRF	**0.6086**	**0.7336**	**0.9662**	**36.9693**	**32.8821**	**28.5414**

3) Poisson noise evaluation: Fig. 10 includes the qualitative evaluation for the dataset containing Poisson noise. Two regions are selected and magnified. In DSIFT, both regions are out of focus. In GFDF, the chair in the green box remains out-of-focus. In DCHWT, the image has lower contrast than the input images. In mf-CRF, the chair in the area with green

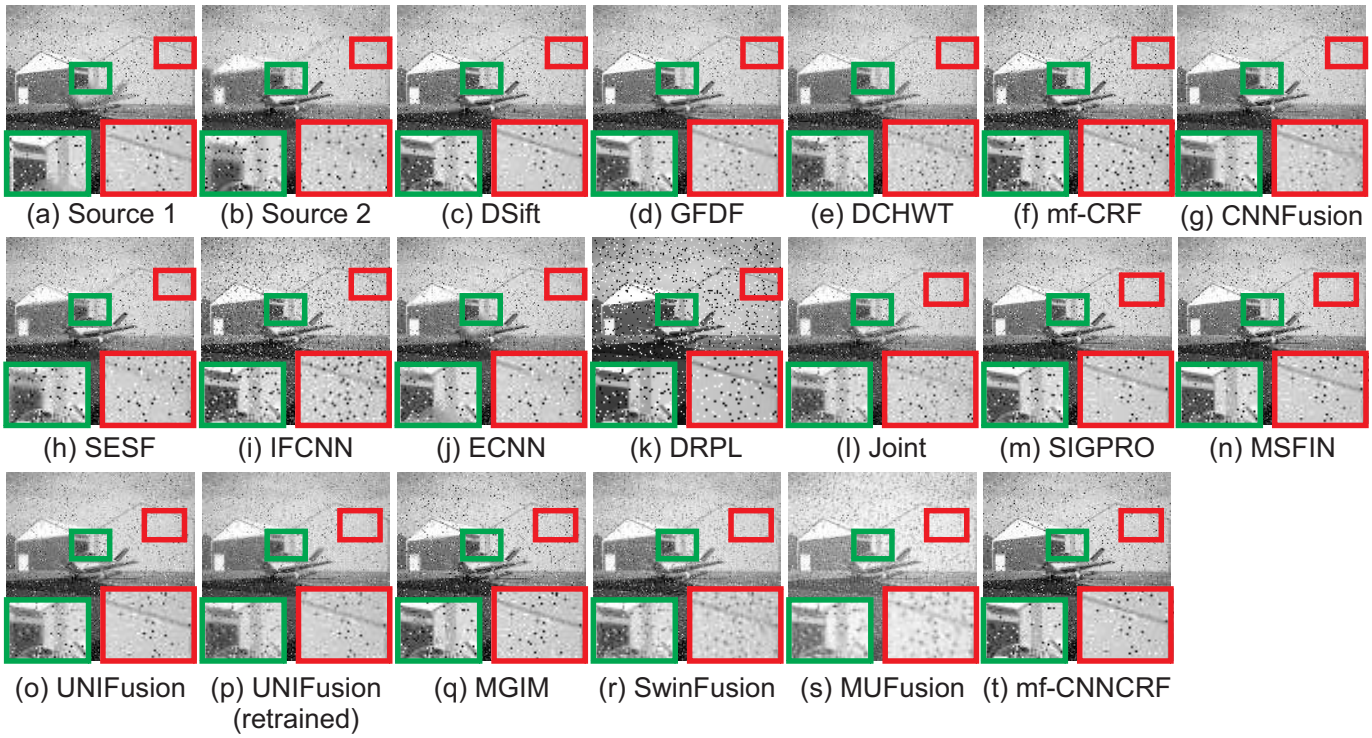

Fig. 9. Salt & Pepper Comparison for density $d = 0.05$.

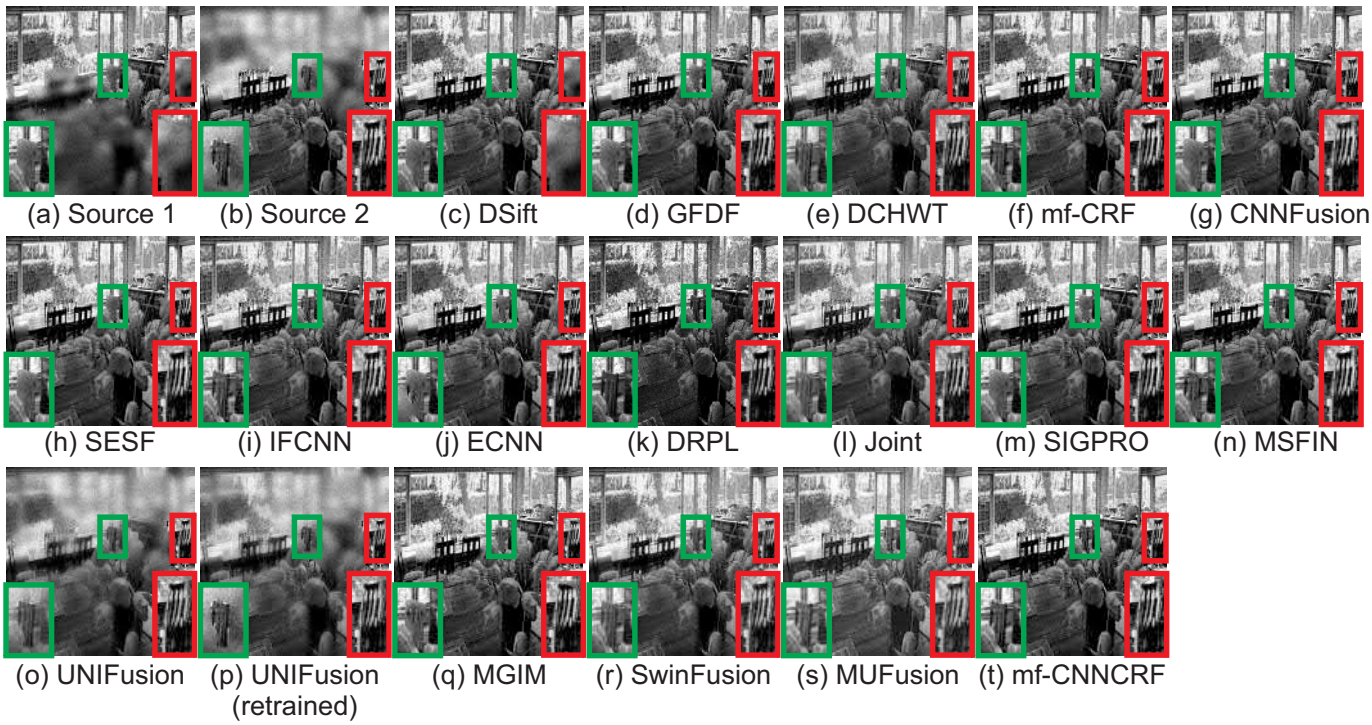

Fig. 10. Source and fused images for the Poisson dataset.

box is half out-of-focus. In CNNFusion and SESF, the chair in the selected region with green box remains out of focus. In IFCNN, the fused result has lower contrast than each of the input images. In ECNN, the selected area of the green box is out-of-focus. In DRPL, there are visible artifacts on the window in the area of the green box. In Joint, SIGPRO and MSFIN the green region is not well focused. In UNIFUsion, retrained UNIFusion and MGIM the green region is out-of-focus. In SwinFusion and MUFusion the fused image has lower contrast than the original images and parts of objects in green region are out-of-focus. The proposed mf-CNNCRF preserves accurately the well focused pixels in both selected

TABLE VII
RMSE/PSNR FOR COMPARED METHODS FOR POISSON NOISE.

Methods	RMSE	PSNR
DSIFT [18]	2.7556	31.4108
GFDF [20]	2.7779	32.0567
DCHWT [54]	5.1700	28.4661
mf-CRF [24]	2.2856	34.9521
CNNFusion [30]	2.9810	31.5903
SESF [55]	3.5593	29.9626
IFCNN [44]	4.9599	28.6489
ECNN [32]	3.1361	29.5647
DRPL [35]	3.9808	28.5953
Joint [27]	3.7800	29.0042
Sigpro [16]	3.1323	30.3853
MSFIN [40]	2.7784	32.5744
UNIFusion [45]	6.8286	23.7262
UNIFusion (retrained)	6.4224	24.0025
MGIM [28]	3.5217	28.1717
SwinFusion [47]	3.5539	27.9271
MUFusion [45]	2.4491	22.2399
mf-CNNCRF	**2.0350**	**36.4569**

regions in the presence of Poisson noise. Lastly, in order to evaluate the performance of compared methods to the presence of Poisson noise, Poisson noise was used to corrupt the input image pairs and ground-truth image was computed based on the ground-truth binary map. Table VII includes the objective evaluation of compared methods in the presence of Poisson noise with the metrics RMSE and PSNR. The proposed mf-CNNCRF exhibits the lowest RMSE value and the highest PSNR values verifying the robustness of mf-CNNCRF in the presence of Poisson noise.

V. COMPUTATIONAL COST COMPARISON

In this section, a comparison between the computational cost of the benchmarked approaches is presented. Since the presented methods are implemented in different frameworks, we focused our study only on the deep learning based approaches. In addition, the comparison is focused on two factors: a) the number of trainable parameters, b) the prediction time. In the proposed method, both CNN networks of the CRF model are efficient siamese architectures. Thus, the branches of each CNN network, share the same weights and the same architecture. Table VIII includes the total number of the trainable parameters for the compared deep learning-based methods. It is evident that CNNFusion has the greatest number of trainable parameters, followed by MSFIN and ECNN. In contrast, the proposed method has the smallest number of trainable parameters, followed by UniFusion. This shows that the proposed approach has the smallest complexity from all examined deep learning based approaches. Table VIII allows depicts the average prediction time for grayscale input image pairs of size 224×224 for all trained deep learning based approaches. The measurements were all performed on the PC, described in Section IV-A. It is clear that ECNN and CNNFusion are the methods with the highest inference time, while SESF is the fastest deep learning-based method

TABLE VIII
COMPLEXITY COMPARISON BETWEEN DEEP LEARNING METHODS IN TERMS OF NUMBER OF TRAINABLE PARAMETERS AND AVERAGE PREDICTION TIME (SEC).

Methods	Parameters	Time (s)
CNNFusion [30]	8759×10^3	33.476
SESF [55]	74.8×10^3	0.347
IFCNN [44]	74.2×10^3	0.749
ECNN [32]	1587.2×10^3	86.123
DRPL [35]	1070×10^3	0.424
MSFIN [40]	4588.9×10^3	0.437
UNIFusion [45]	38.7×10^3	6.278
SwinFusion [47]	973.7×10^3	0.874
MUFusion [45]	554.7×10^3	0.53
mf-CNNCRF	22.5×10^3	0.637

with 0.347 sec. The proposed method features a favourable processing time of 0.637 sec, which is in the same par with most methods.

VI. CONCLUSION

In this paper, a CNN-based CRF model, named mf-CNNCRF, for multi-focus image fusion is proposed. The proposed method uses the rich prior potentials learned through CNN training and the long range interactions of the CRF model in order to reach structured inference with global or close to global solution for multi-focus image fusion. The proposed framework uses efficient siamese architectures, in order to support arbitrary number of the input images that can be processed in parallel, making mf-CNNCRF computationally efficient. The developed dataset includes both clean training images and training images with Gaussian noise, Salt & Pepper noise and Poisson noise. Experimental results demonstrate that the proposed mf-CNNCRF outperforms state-of-the-art MFIF methods on both qualitative and quantitative evaluations for both clean images and images with Gaussian noise, Salt & Pepper noise and Poisson noise and exhibits high generalization capabilities. It is important to note that the proposed algorithm does not require knowledge of the noise type or statistics. This is the first work, to the best of our knowledge, that uses a CNN architecture to learn rich CRF priors. Our future work will focus on refining the network architectures for UnaryNet and SmoothnessNet and replace the a-expansion algorithm with a deep learning based network.

REFERENCES

[1] Y. Liu, L. Wang, J. Cheng, C. Li, and X. Chen, "Multi-focus image fusion: A survey of the state of the art," *Information Fusion*, vol. 64, pp. 71–91, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1566253520303109

[2] B. Aiazzi, L. Alparone, A. Barducci, S. Baronti, and I. Pippi, "Multispectral fusion of multisensor image data by the generalized laplacian pyramid," in *IEEE International Geoscience and Remote Sensing Symposium*, vol. 2, 1999, pp. 1183–1185.

[3] L. Bogoni and M. Hansen, "Pattern-selective color image fusion," *Pattern Recognition*, vol. 34, no. 8, pp. 1515–1526, 2001. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S003132030000087X

[4] H. Li, B. S. Manjunath, and S. K. Mitra, "Multi-sensor image fusion using the wavelet transform," in *Image Processing, 1994. Proceedings. ICIP-94., IEEE International Conference*, vol. 1. IEEE, 1994, pp. 51–55.

[5] Q. Zhang and B.-l. Guo, "Multifocus image fusion using the nonsubsampled contourlet transform," *Signal Processing*, vol. 89, no. 7, pp. 1334–1346, 2009.

[6] F. Nencini, A. Garzelli, S. Baronti, and L. Alparone, "Remote sensing image fusion using the curvelet transform," *Information Fusion*, vol. 8, no. 2, pp. 143–156, 2007.

[7] B. Yang and S. Li, "Multifocus Image Fusion and Restoration With Sparse Representation," *IEEE Transactions on Instrumentation and Measurement*, vol. 59, no. 4, pp. 884–892, 2010.

[8] Y. Liu and Z. Wang, "Simultaneous image fusion and denoising with adaptive sparse representation," *IET Image Processing*, vol. 9, no. 5, pp. 347–357, 2015.

[9] Q. Zhang and M. D. Levine, "Robust Multi-Focus Image Fusion Using Multi-Task Sparse Representation and Spatial Context," *IEEE Transactions on Image Processing*, vol. 25, no. 5, pp. 2045–2058, 2016.

[10] Z. Zhou, S. Li, and B. Wang, "Multi-scale weighted gradient-based fusion for multi-focus images," *Information Fusion*, vol. 20, pp. 60–72, 2014.

[11] J. Sun, H. Zhu, Z. Xu, and C. Han, "Poisson image fusion based on Markov random field fusion model," *Information fusion*, vol. 14, no. 3, pp. 241–2535, 2013.

[12] N. Mitianoudis and T. Stathaki, "Pixel-based and region-based image fusion schemes using ICA bases," *Information Fusion*, vol. 8, no. 2, pp. 131–142, 2007.

[13] X. Luo, Z. Zhang, C. Zhang, and X. Wu, "Multi-focus image fusion using hosvd and edge intensity," *Journal of Visual Communication and Image Representation*, vol. 45, pp. 46–61, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1047320317300433

[14] 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, pp. 147–164, 2015.

[15] S. Li and B. Yang, "Hybrid multiresolution method for multisensor multimodal image fusion," *IEEE Sensors Journal*, vol. 10, no. 9, pp. 1519–1526, 2010.

[16] X. Li, F. Zhou, H. Tan, Y. Chen, and W. Zuo, "Multi-focus image fusion based on nonsubsampled contourlet transform and residual removal," *Signal Processing*, vol. 184, p. 108062, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165168421001006

[17] A. Hyvärinen, J. Hurri, and P. O. Hoyer, "Independent Component Analysis BT - Natural Image Statistics: A Probabilistic Approach to Early Computational Vision," A. Hyvärinen, J. Hurri, and P. O. Hoyer, Eds. London: Springer London, 2009, pp. 151–175.

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

[19] S. Li, X. Kang, J. Hu, and B. Yang, "Image matting for fusion of multi-focus images in dynamic scenes," *Information Fusion*, vol. 14, no. 2, pp. 147–162, 2013.

[20] X. Qiu, M. Li, L. Zhang, and X. Yuan, "Guided filter-based multi-focus image fusion through focus region detection," *Signal Processing: Image Communication*, vol. 72, pp. 35–46, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0923596518302832

[21] X. Bai, Y. Zhang, F. Zhou, and B. Xue, "Quadtree-based multi-focus image fusion using a weighted focus-measure," *Information Fusion*, vol. 22, pp. 105–118, 2015.

[22] Y. Zhang, X. Bai, and T. Wang, "Boundary finding based multi-focus image fusion through multi-scale morphological focus-measure," *Information Fusion*, vol. 35, pp. 81–2535, 2017.

[23] B. K. Shreyamsha Kumar, "Image fusion based on pixel significance using cross bilateral filter," *Signal, Image and Video Processing*, vol. 9, no. 5, pp. 1193–1204, 2015. [Online]. Available: https://doi.org/10.1007/s11760-013-0556-9

[24] O. Bouzos, I. Andreadis, and N. Mitianoudis, "Conditional random field model for robust multi-focus image fusion," *IEEE Transactions on Image Processing*, vol. 28, no. 11, pp. 5636–5648, Nov 2019.

[25] Y. Yang, S. Tong, S. Huang, and P. Lin, "Multifocus image fusion based on nsct and focused area detection," *IEEE Sensors Journal*, vol. 15, no. 5, pp. 2824–2838, 2015.

[26] Z. Wang, S. Wang, and Y. Zhu, "Multi-focus image fusion based on the improved pcnn and guided filter," *Neural Processing Letters*, vol. 45, no. 1, pp. 75–94, 2017. [Online]. Available: https://doi.org/10.1007/s11063-016-9513-2

[27] X. Li, F. Zhou, and H. Tan, "Joint image fusion and denoising via three-layer decomposition and sparse representation," *Knowledge-Based Systems*, vol. 224, p. 107087, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705121003506

[28] J. Chen, X. Li, L. Luo, and J. Ma, "Multi-focus image fusion based on multi-scale gradients and image matting," *IEEE Transactions on Multimedia*, vol. 24, pp. 655–667, 2022.

[29] X. Zhang, "Deep learning-based multi-focus image fusion: A survey and a comparative study," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021.

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

[31] H. Tang, B. Xiao, W. Li, and G. Wang, "Pixel convolutional neural network for multi-focus image fusion," *Information Sciences*, vol. 433-434, pp. 125–141, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0020025517311647

[32] M. Amin-Naji, A. Aghagolzadeh, and M. Ezoji, "Ensemble of cnn for multi-focus image fusion," *Information Fusion*, vol. 51, pp. 201–214, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1566253518306043

[33] C. Du and S. Gao, "Image segmentation-based multi-focus image fusion through multi-scale convolutional neural network," *IEEE Access*, vol. 5, pp. 15 750–15 761, 2017.

[34] X. Guo, R. Nie, J. Cao, D. Zhou, and W. Qian, "Fully convolutional network-based multifocus image fusion," *Neural Computation*, vol. 30, no. 7, pp. 1775–1800, 2020/11/27 2018.

[35] J. Li, X. Guo, G. Lu, B. Zhang, Y. Xu, F. Wu, and D. Zhang, "Drpl: Deep regression pair learning for multi-focus image fusion," *IEEE Transactions on Image Processing*, vol. 29, pp. 4816–4831, 2020.

[36] H. Ma, Q. Liao, J. Zhang, S. Liu, and J. H. Xue, "An a-matte boundary defocus model-based cascaded network for multi-focus image fusion," *IEEE Transactions on Image Processing*, vol. 29, pp. 8668–8679, 2020.

[37] B. Xiao, B. Xu, X. Bi, and W. Li, "Global-feature encoding u-net (geu-net) for multi-focus image fusion," *IEEE Transactions on Image Processing*, vol. 30, pp. 163–175, 2021.

[38] K. He, J. Sun, and X. Tang, "Guided image filtering," in *Computer Vision –ECCV 2010*, K. Daniilidis, P. Maragos, and N. Paragios, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 1–14.

[39] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in *Advances in Neural Information Processing Systems*, J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger, Eds., vol. 24. Curran Associates, Inc., 2011, pp. 109–117.

[40] Y. Liu, L. Wang, J. Cheng, and X. Chen, "Multiscale feature interactive network for multifocus image fusion," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–16, 2021.

[41] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2fusion: A unified unsupervised image fusion network," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2020.

[42] H. Li, R. Nie, J. Cao, X. Guo, D. Zhou, and K. He, "Multi-focus image fusion using u-shaped networks with a hybrid objective," *IEEE Sensors Journal*, pp. 1–1, 2019.

[43] H. Li and X. Wu, "Densefuse: A fusion approach to infrared and visible images," *IEEE Transactions on Image Processing*, pp. 1–1, 2019.

[44] Y. Zhang, Y. Liu, P. Sun, H. Yan, X. Zhao, and L. Zhang, "Ifcnn: A general image fusion framework based on convolutional neural network," *Information Fusion*, vol. 54, pp. 99 – 118, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1566253518305505

[45] C. Cheng, X.-J. Wu, T. Xu, and G. Chen, "Unifusion: A lightweight unified image fusion network," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–14, 2021.

[46] W. Zhao, D. Wang, and H. Lu, "Multi-focus image fusion with a natural enhancement via a joint multi-level deeply supervised convolutional neural network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 4, pp. 1102–1115, 2019.

[47] J. Ma, L. Tang, F. Fan, J. Huang, X. Mei, and Y. Ma, "Swinfusion: Cross-domain long-range learning for general image fusion via swin transformer," *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 7, pp. 1200–1217, 2022.

[48] C. Cheng, T. Xu, and X.-J. Wu, "Mufusion: A general unsupervised image fusion network based on memory unit," *Information Fusion*, vol. 92, pp. 80–92, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253522002202

[49] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4353–4361.

[50] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 11, pp. 1222–1239, 2001.
[51] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.
[52] M. Nejati, S. Samavi, and S. Shirani, "Multi-focus image fusion using dictionary-based sparse representation," *Information Fusion*, vol. 25, pp. 72–84, 2015.
[53] S. Xu, X. Wei, C. Zhang, J. Liu, and J. Zhang, "Mffw: A new dataset for multi-focus image fusion," 2020. [Online]. Available: https://arxiv.org/abs/2002.04780
[54] B. K. Shreyamsha Kumar, "Multifocus and multispectral image fusion based on pixel significance using discrete cosine harmonic wavelet transform," *Signal, Image and Video Processing*, vol. 7, no. 6, pp. 1125–1143, 2013. [Online]. Available: https://doi.org/10.1007/s11760-012-0361-x
[55] B. Ma, Y. Zhu, X. Yin, X. Ban, H. Huang, and M. Mukeshimana, "Sesf-fuse: an unsupervised deep model for multi-focus image fusion," *Neural Computing and Applications*, vol. 33, no. 11, pp. 5793–5804, 2021. [Online]. Available: https://doi.org/10.1007/s00521-020-05358-9
[56] Z. Liu, E. Blasch, Z. Xue, J. Zhao, R. Laganiere, and W. Wu, "Objective Assessment of Multiresolution Image Fusion Algorithms for Context Enhancement in Night Vision: A Comparative Study," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 1, pp. 94–109, 2012.

Odysseas Bouzos received the Diploma Degree from the Department of Electrical and Computer Engineering, Democritus University of Thrace (DUTH), Greece, in 2013. He also received an MSc in High Dynamic Range image fusion from the same department in 2014. He received his PhD in 2022 on Optimization techniques for Image Fusion at the same department. His research interests include Probabilistic Graphical Models, Deep Learning, Machine Learning, Computer Vision and Image Fusion. From 2015 until 2016 he was an intern at Centre for Robotics, MINES ParisTech, PSL Research University, Paris, France. From 2009 until 2010 he worked as a user of the ATLAS experiment at CERN - the European Organisation for Nuclear Research.

Ioannis Andreadis received the Diploma Degree from the Department of Electrical and Computer Engineering, Democritus University of Thrace (DUTH), Greece, in 1983 [1978-1983] IKY Scholarship] and the M.Sc. [Electrical Power Systems Engineering] and Ph.D. [Machine Vision] Degrees from the University of Manchester Institute of Science and Technology, UK, in 1985 and 1989, respectively. He was awarded the IET Image Processing Premium in 2009. He received the best paper award (Computer Vision and Applications) in PSIVT 2007 as well the best paper award in EUREKA 2009. He is author of an Institute of Physics (IOP) "Select Paper" in 2010. He has supervised 13 PhD Theses, 20 MSc Theses and 70 Diploma dissertations. Professor Andreadis was a member of the Board of Governors of the European Commission Joint Research Center (JRC) from 2008-2010. He is a Fellow of the Institution of Engineering and Technology (2006), (IET - London, UK).

Nikolaos Mitianoudis (S'98 - M'04 - SM'11) received the diploma in Electronic and Computer Engineering from the Aristotle University of Thessaloniki, Greece in 1998. He received the MSc in Communications and Signal Processing from Imperial College London, UK in 2000 and the PhD in Audio Source Separation using Independent Component Analysis from Queen Mary, University of London, UK in 2004. Between 2003 and 2009, he was a Research Associate at Imperial College London, UK working on the Data Information Fusion-Defense Technology Centre project "Applied Multi-Dimensional Fusion", sponsored by General Dynamics UK and QinetiQ. From 2009 until 2010, he was an Academic Assistant at the International Hellenic University. Currently, he is an Assistant Professor in Audio and Image Processing at the Democritus University of Thrace, Greece. He is an Associate Editor of the IEEE Trans. on Image Processing (2018-2024). His research interests include Image Fusion, Computer Vision, Deep Learning, Semantic Segmentation, Music Information Retrieval and Blind Source Separation/Extraction.