

MobileMEF: Fast and Efficient Method for Real-Time Mobile Multi-Exposure Fusion

Lucas Nedel Kirsten, Zhicheng Fu, Nikhil Ambha Madhusudhana

Abstract—Recent advances in camera design and imaging technology have enabled the capture of high-quality images using smartphones. However, due to the limited dynamic range of digital cameras, the quality of photographs captured in environments with highly imbalanced lighting often results in poor-quality images. To address this issue, most devices capture multi-exposure frames and then use some multi-exposure fusion method to merge those frames into a final fused image. Nevertheless, most traditional and current deep learning approaches are unsuitable for real-time applications on mobile devices due to their heavy computational and memory requirements. We propose a new method for multi-exposure fusion based on an encoder-decoder deep learning architecture with efficient building blocks tailored for mobile devices. This efficient design makes our model capable of processing 4K resolution images in less than 2 seconds on mid-range smartphones. Our method outperforms state-of-the-art techniques regarding full-reference quality measures and computational efficiency (runtime and memory usage), making it ideal for real-time applications on hardware-constrained devices. Our code is available at: <https://github.com/LucasKirsten/MobileMEF>.

Index Terms—Multi-exposure fusion, image fusion, efficient methods

I. INTRODUCTION

Photographs captured in environments with highly imbalanced lighting often result in poor-quality images due to underexposed and overexposed regions. This issue arises from the limited dynamic range of digital cameras, which is significantly lower than in real-world scenes [1]. High dynamic range (HDR) imaging techniques have been developed to address this limitation, with multi-exposure image fusion (MEF) being a prominent solution [2], [3]. MEF methods merge multiple images captured at different exposure levels to produce a single image that intends to retain the scene details and color fidelity [4]. Despite the advancements, many existing MEF methods rely on hand-crafted features or transformations, leading to robustness issues under varying conditions [5].

Traditional MEF methods, such as those based on Laplacian pyramids [4], [1], are computationally intensive due to the multiple operations required for generating pyramid sub-images [5]. This computational overhead becomes particularly problematic for hardware-constrained applications like smartphones, especially when processing high-resolution 4K images. Single-scale fusion methods [6] have been proposed to alleviate the computational burden, but they often produce lower-quality images due to noticeable seams and

Lucas N. Kirsten (corresponding author, email: lucask@motorola.com) is with Motorola Mobility Comércio de Produtos Eletrônicos Ltda, Jaguariúna, SP 13918-900 BR. Zhicheng Fu and Nikhil Ambha Madhusudhana are with Lenovo Research.

gray differences [1], [2]. Additionally, while deep learning-based approaches have shown promise in improving MEF [3], they often do not consider real-world deployment constraints, resulting in a trade-off between speed and quality that limits their application on mobile platforms [7], [8], [9], [10].

To address these challenges, we propose a novel MEF method based on an Encoder-Decoder deep learning architecture designed to optimize mobile device performance, named *MobileMEF*. MobileMEF draws inspiration from recent advancements in deep learning [11], [12] but introduces several key modifications to enhance efficiency and effectiveness for MEF tasks. As illustrated in Fig. 1, our method achieves superior image quality results with the lowest or near-lowest required operations compared to existing state-of-the-art (SOTA) methods. This balance of high-quality image output and low computational demand makes our method highly suitable for deployment on mobile devices, enabling efficient processing of 4K resolution images without compromising performance. Our main contributions to the field are:

- We present MobileMEF, an optimized model architecture designed for hardware-constrained MEF applications. MobileMEF outperforms state-of-the-art techniques regarding full-reference quality measures and computational efficiency (runtime and memory usage), being capable of processing 4k resolution images in less than 2 seconds on mid-range smartphones;
- We propose a new bypass module based on Single-Scale Fusion [6] with YUV color space images [1] that forwards an estimate of the fused image channels from the input frames to the model output predictions;
- We propose a new Gradient loss [13] formulation based on cropping, capable of capturing fine details and overall image context of the predicted and ground-truth images.

II. RELATED WORK

Traditional MEF methods can be divided into Multi-Scale and Single-Scale based methods [5], [2]. Multi-Scale methods, like the Mertens [4] algorithm, use a sequence of Laplacian pyramids to decompose input frames into sub-images, alleviating exposure differences and smoothing local transitions by computing three quality measures related to saturation, contrast, and exposure levels. The Fast YUV [1] method improves on this by using the YUV color space and computing pyramids only on the Y channel, reducing computational efforts. In contrast, Single-Scale methods, such as the one proposed by Ancuti et al. [6], approximate the operations in the pyramid step of the Mertens algorithm to a single step, reducing computational resource requirements.

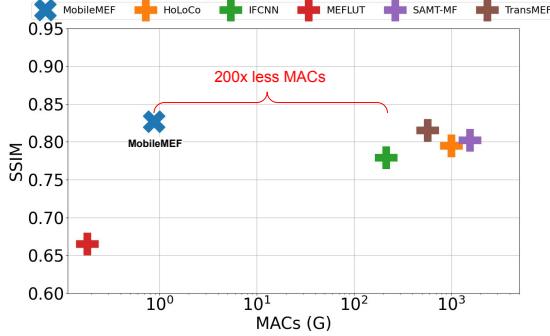


Fig. 1. Comparison of computational cost and performance of SOTA MEF methods: HoLoCo [8], IFCNN [7], MEFLUT [14], SAMT-MEFT [10], TransMEF [9]. MobileMEF can process 4k images in 1.82 seconds on a mid-range mobile device using GPU.

As in many other fields, methods based on deep learning have been proposed to solve the MEF problem [3]. For example, the IFCNN [7] model presented a general MEF framework based on convolutional neural networks. TransMEF [9] is a transformer-based model that uses self-supervised multi-task learning based on an encoder-decoder architecture. HoLoCo [8] proposes a method based on local feature constraints built upon contrastive learning. The MEFLUT [14] model learns to encode the fusion weights as 1D lookup tables to achieve a highly efficient method. More recently, SAMT-MEFT [10] has presented a mean teacher-based semi-supervised learning framework tailored for MEF.

In this work, we propose MobileMEF, an end-to-end encoder-decoder model designed for fast and efficient computation of MEF. Our work is built on top recent advances in deep-learning and MEF efficient designs, such as the LPIENet [11] architecture, the convolutional blocks of ConvNeXt [12], the Fast YUV [1] and the Single-Scale MEF method proposed by Ancuti et al. [6]. Furthermore, we extend the formulation of the widely used Gradient loss [13] to capture fine details and overall image context to improve the fused image quality.

III. THE PROPOSED METHOD

We propose MobileMEF, a new method for MEF based on an Encoder-Decoder deep learning architecture with efficient building blocks designed to work on mobile phones, presented in Fig. 2. The overall model architecture is inspired by the recent success of the LPIENet [11] image enhancer model, but we redesigned some of their main components to work for MEF and to improve performance in mobile devices with 4K resolution images. Specifically, we redesign the model's input pipeline to work with multiple input images in the YUV color space; we replace their base convolutional blocks to use one based on ConvNeXt [12]; added Squeeze-and-Excitation [15] attention block to the Encoder and Decoder model blocks; and added a Single-Scale Fusion (SSF) bypass module based on [6] and [1] that intends to forward fused information of the inputs to the model's output. We proceed to provide details regarding our method.

A. Input pipeline

We first concatenate the inputs' channels of the multiple input frames in the form that our model will receive inputs with shape $(H, W, K \times C)$, where H is the height, W the width, C the image channels, and K the number of input frames. However, note that this simple scheme can cause computational overhead since the model requires to operate in K times more channels in its first layers. To address this issue, we follow the proposal of other works [14], [7] and convert the RGB input frames to the YUV color space. Converting the inputs to YUV allows us to work with smaller inputs in the UV channels since they are only related to color [1], and recent works have demonstrated that similar approaches yield superior performance with reduced computational costs (runtime and memory) [5]. Hence, our input data pipeline is divided into two parts: one related to the Y channels of the input frames with shape (H, W, K) ; and the other related to the UV channels with shape $(\kappa H, \kappa W, K \times 2)$, where $0 < \kappa < 1$ is a down-sample ratio value. Moreover, note that this scheme requires using two Encoder-Decoder models with SSF modules: one for the Y inputs, and the other for the UV, as we proceed to explain.

B. Network architecture

The overall proposed network architecture for MobileMEF is presented in Fig. 2, and is based on the success of recent works for efficient model design for mobile applications [11], [12], [16]. For K input frames with shape (H, W, C) , where H is the height, W the width, and C the image channels, the forward process is designed as:

- 1) The input frames are converted to the YUV color space and split in the channel axis. The inputs related to the UV channels are downsampled by a ratio κ . The Y and resized UV inputs are stacked in the channel axis, providing the two inputs with shape (H, W, K) for the Y channels, and $(\kappa H, \kappa W, K \times 2)$ for the UV channels;
- 2) Both inputs are downsampled by a ratio Γ ;
- 3) Each downsampled input is forwarded by a set of N encoder and decoder blocks;
- 4) The output from the decoder is then fed to a sequence composed of one Inverted residual (InvBlock) [16] and one Spatial Attention (SAT) module;
- 5) The downsampled inputs are processed in parallel by the SSF module that merges the K input frames;
- 6) The predicted fused image from the model's output is added to the merged image from the SSF module;
- 7) The summed image is then upsampled by the ratio Γ to produce the fused channel (Y or UV).

We advocate that using a small downsampling ratio $\Gamma \leq 2$ to the input frames has negligible effects on the result fused image. However, it significantly reduces computational efforts since inputting high-resolution images to the first layers of the model can cause severe computation overheads. We proceed to provide details regarding the blocks of the proposed network architecture following the name convention in Fig. 2.

ConvBlock. We based our Convolutional Block on the ConvNeXt [12], as recent works have demonstrated its superiority

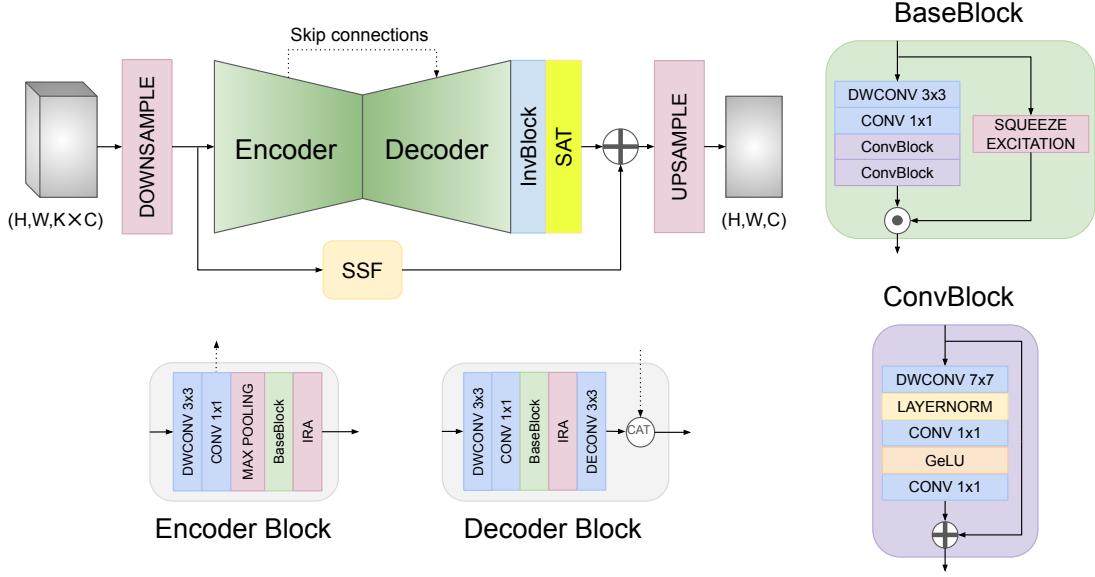


Fig. 2. Overall proposed MobileMEF model architecture. The encoder-decoder model receives an input with C channels and K frames and returns the fused frames. The SSF module merges the K input frames and adds it to the model’s output.

on feature extraction [17] due to its similarity with the recent Visual Transformers [18], but with a more efficient design. Our convolutional blocks are composed of: a Depthwise Convolutional layer with 7×7 kernel size, a Layer Normalization [19], a Pointwise Convolution that doubles the number of input features, a GeLU [20] activation layer, and finally a Pointwise convolution that restores the same number of input features as the inputs. The output of the final Pointwise convolution is then added to the input of the ConvBlock.

BaseBlock. Similarly with [11], our Base Blocks are composed of a first 3×3 Depthwise Convolution layer, a Pointwise Convolution that doubles the number of input features, and a sequence of two convolutional blocks. However, we include a Squeeze and Excitation [15] module that is intended to add a channel attention mechanism related to the inputs. Specifically related to the MEF problem, our experiments demonstrated that the Squeeze and Excitation module can significantly boost the results with neglectable computation overhead. We assign this performance boost to the fact that the input frames are stacked in the channel axis, where the Squeeze and Excitation module operates.

Encoder Block. The Encoder Block compresses the inputs’ spatial dimension while expanding its feature size [21]. First, a 3×3 Depthwise Convolution layer followed by a Pointwise Convolution that doubles the number of input features is used. The output of these convolutions is stored to be used for the skip connections. Next, a Max Pooling 2D layer is employed to reduce the spatial size by a ratio of 2, and is followed by a BaseBlock and an Inverted Residual Attention (IRA) module [11]. The IRA module was proposed in [11] to add both spatial and channel attention to each Encoder/Decoder block, and is illustrated in Fig. 3.

Decoder Block. The Decoder Block expands the spatial dimension of the input features while reducing its feature size

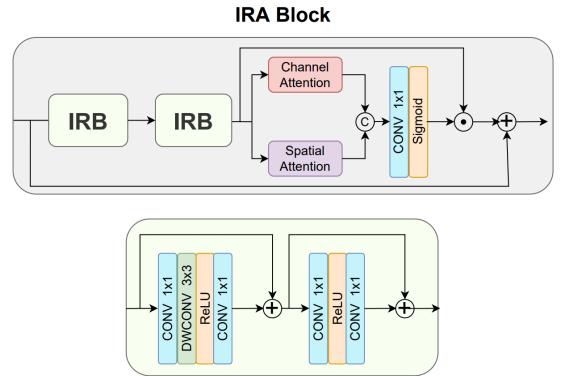


Fig. 3. IRA Block as proposed in LPIENet [11].

to retain only the most important features extracted during the encoder step [21]. First, we employ a 3×3 Depthwise Convolution layer followed by a Pointwise Convolution that halves the number of input features. Next, we use a BaseBlock and an IRA module [11]. Finally, a Transposed Convolution with the same number of input and output features is used to upsample the features to be concatenated with the respective encoder layer output (i.e., the skip connection).

InvBlock. The Inverted residual (InvBlock) [16] module adds residual information through an efficient inverted structure, and is commonly used in many mobile applications. It comprises a branch of Pointwise convolution, Depthwise convolution, another Pointwise convolution (all with the same number of output features); and another branch composed of a Pointwise convolution with a ReLU activation. The output of both branches is then added to compose the final InvBlock output.

SAT. The Spatial Attention module is designed to enhance and suppress certain spatial features of the inputs through a

progressive receptive field augmentation [11]. It is composed of three Convolution layers followed by ReLU (the first two) and Sigmoid activation (the last one). The convolutions use an increasing kernel size scheme of 7×7 , 5×5 , and 3×3 , respectively. All convolutions have the same number of output features, that is the number of expected predicted output channels (1 for the Y channel, and 2 for the UV channels). The last output tensor (the one related to the Sigmoid activation) is then dot multiplied by the input tensor of the SAT module.

C. SSF module

Following the LPIENet [11] network architecture, and as in many other image enhancement works [22], we also add a skip connection that sums the input image with the network output. This simple strategy is usually used to improve model convergence and performance since it allows the network to focus only on correcting the input image, reducing the necessity of also retaining features for reconstructing it during the Encoder-Decoder step [22]. However, for the MEF problem, the inputs and outputs have different numbers of channels, since the main objective is to fuse the input frames into a unique output. One could argue that a simple reduction function, such as the mean or median, could be used to match the network output channels. Nevertheless, in [5] the authors have demonstrated that such stack techniques tend to compromise the final image quality. Moreover, they also show that Single-Scale Fusion (SSF) methods (e.g., [6]) are capable of producing similar results to Multi-Scale ones (e.g., [4], [1]), with the advantage of reducing computational efforts.

We propose to use a learnable version of the SSF method named “SSF-YUV” described in [5] for the skip input path of our model in order to provide a “closer” estimate of the fused input to the model prediction. We refer to it as the *SSF module* and proceed to detail it. This module is based on the works of Ancuti et al. [6], which describes a Single-Scale Fusion method based on approximations on the Mertens [4] algorithm; and the Fast YUV [1], which describes an efficient MEF method designed to fuse images in the YUV color space.

According to [6], a fused image channel (e.g., for RGB image, the red, green, or blue channel) can be obtained with:

$$\mathcal{F} = \sum_{i=1}^K \omega_B \circledast \bar{\mathcal{W}} + \alpha \cdot |\mathcal{P}_1(\mathcal{I}_i) \odot \mathcal{P}_1(\bar{\mathcal{W}})|, \quad (1)$$

where K is the number of input frames, ω_B is a blur kernel, $\bar{\mathcal{W}} \in \mathbb{R}^2$ is the normalized *weight maps*¹ of the fusion method, α is a scalar constant set empirically, $\mathcal{I}_i \in \mathbb{R}^2$ is the i -th input frame channel, and $\mathcal{P}_1(I) = I - \text{UP}(\text{DOWN}(I))$ is a single step Laplacian Pyramid function, where UP and DOWN are up-sampling and down-sampling operations, respectively that halve the spatial dimensions. The definition of these up- and down-sampling operations is usually accompanied by convolving the inputs with Gaussian kernels. However, in our

¹Note that the term “weights” here have a different meaning than the one used for the weights of a deep learning model. In traditional MEF methods (such as [6], [1], [4]), the “weight maps” are the output of hand-coded functions applied to the input frames to extract some visual information of it, such as contrast, saturation, and exposure.

experiments, we noted that a standard resizing with bi-linear interpolation could achieve similar or superior results.

For YUV images, the UV channels relate solely to color, so using some weighted average method such as Eq. 1 for fusing these channels would shift all colors to gray [1]. To address this, we use Eq. 1 to compute the fusion solely in the Y channel, whereas for the UV channels we use [1] proposal:

$$\mathcal{F}_U = \max_K \mathcal{I}_U \quad (2)$$

and

$$\mathcal{F}_V = \max_K \mathcal{I}_V, \quad (3)$$

where \mathcal{I}_U and \mathcal{I}_V are the U and V channels of the input frames, respectively. This formulation reasons that, when imaging, for a reasonably exposed area, the color should be bright and so its color components should be close to its maximum value.

Based on the [1] work, we also use two weight maps related to computing the contrast and exposure of the YUV input frames. We compute the contrast weight as:

$$\mathcal{W}_C = \omega_C \circledast \mathcal{I}_Y, \quad (4)$$

where ω_C is the Laplacian kernel, and \mathcal{I}_Y is the Y channel of the input frames. And for the exposure weight:

$$\mathcal{W}_E = |\mathcal{I}_U| \odot |\mathcal{I}_V|. \quad (5)$$

The final normalized weight maps is computed as:

$$\bar{\mathcal{W}} = \frac{\mathcal{W}_C \odot \mathcal{W}_E}{\sum_{i=1}^K \mathcal{W}_C \odot \mathcal{W}_E}. \quad (6)$$

D. Loss function

We define the loss function to train our complete deep-learning model based on a weighted sum of \mathcal{L}_1 (mean absolute error) and Gradient loss [13]. The final loss function is:

$$\mathcal{L} = \mathcal{L}_1 + \lambda \cdot \mathcal{L}_{\text{Grad}}, \quad (7)$$

where $\mathcal{L}_{\text{Grad}}$ is the Gradient loss, and λ is a scalar constant used to scale the loss.

For the Gradient loss, the features extracted from a large pre-trained classification model (e.g., VGG16 or VGG19 [23]) are used to compute the \mathcal{L}_1 distance between the ground-truth and predicted features, to preserve image structure and generate perceptual-pleasant details. These classification models usually are trained with small inputs (e.g., 224×244). However, our model is intended to work with 4K high-resolution images, and feeding such large images would harm the classifier’s capabilities of extracting features. One solution would be to resize the ground truth and predicted images to match the classifier input size, but it can cause the loss of fine details. Current works [11], [8] usually rely on cropping the high-resolution inputs to fit a “more suitable” image size (e.g., 512×512) to reduce the loss of fine details to the Gradient loss. Nevertheless, this is also not ideal, because it may lose some contextual information of the whole image.

We propose a simple method to improve the Gradient loss performance, based on resizing and cropping the input

and ground-truth images to preserve contextual information and finer details. Given the predicted and ground-truth images $\{\mathcal{I}(x, y), \mathcal{I}_{\text{GT}}(x, y)\} \in \mathbb{R}^{H \times W}$ respectively (omitted the channel axis for simplicity), and a feature extractor classifier Ψ trained with input size $h \times w$, we extract the crops from the images in the form:

$$\delta(\mathcal{I}) = \begin{cases} I(x + \Delta x, y + \Delta y) & \text{if } 0 \leq x \leq w, 0 \leq y \leq h \\ 0 & \text{otherwise} \end{cases}, \quad (8)$$

where Δx and Δy are used to define the starting point for the crop region. The final Gradient loss is then computed using the following composition:

$$\mathcal{L}_{\text{Grad}} = \mathcal{L}_1(\Psi(\mathcal{I}^r), \Psi(\mathcal{I}_{\text{GT}}^r)) + \frac{1}{M} \cdot \sum_{i=1}^M \mathcal{L}_1(\Psi(\delta_i(\mathcal{I})), \Psi(\delta_i(\mathcal{I}_{\text{GT}}))), \quad (9)$$

where $\{\mathcal{I}^r, \mathcal{I}_{\text{GT}}^r\} \in \mathbb{R}^{h \times w}$ are the resized input and ground-truth images respectively, and M is the number of extracted crops. We argue that this definition of the Gradient loss can extract fine details and preserve image context, as we proceed to demonstrate in our results.

IV. EXPERIMENTAL SETUP

A. Dataset

We employ the widely used open-source benchmark MEF dataset SICE [24] to train and evaluate our proposed method. It is composed of both indoor and outdoor scenes, captured with seven types of consumer-grade cameras. We used the most common split of SICE composed of 360 scenes, of which 302 are used for training and 58 for testing. The resolution of most images is between 3000×2000 and 6000×4000 pixels. Each scene comprises at least three images of different Exposure Values (EV) with values -1, 0, and +1, plus one ground-truth image.

B. Implementation details

We employ a reduction factor of $\kappa = 1/4$ on the UV channels (recall Sec. III-A), and of $\Gamma = 2$ for the input frames (both Y and UV, recall Sec. III-B). In all resize operations we use the bi-linear interpolation. We use $N = 5$ encoder/decoder blocks for the Y inputs, and $N = 3$ for the UV inputs, since they are 4 times smaller than the Y ones. The number of features for the first Pointwise convolution of the encoder branch is set to 4 for the Y inputs, and 8 for the UV inputs. This value is doubled for each new encoder block and halved on the respective decoder one (recall Sec. III-B).

For the SSF module (Eq. 1), we use $\alpha = 0.2$ as in [6], and $\omega_B \in \mathbb{R}^{5 \times 5}$ kernel is initialized with all its values equal $1/25$. For the contrast weight (Eq. 4), $\omega_C \in \mathbb{R}^{3 \times 3}$ is initialized with values:

$$\omega_C = \begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \quad (10)$$

which is the Laplacian kernel. Both ω_B and ω_C are treated as learnable parameters during the model training and are implemented using Depthwise convolutions.

For the loss function, we set $\lambda = 1$ since we use \mathcal{L}_1 for computing both $\mathcal{L}_{\text{Grad}}$ (Eq. 9) and the complete \mathcal{L} loss (Eq. 7). For the proposed crop-like $\mathcal{L}_{\text{Grad}}$ (Eq. 9), we extract $M = 5$ crops of the input and ground truth images using Eq. 8, with $(\Delta x, \Delta y) = \{(0, 0), (W-w, 0), (0, H-h), (W-w, H-h), (\frac{W-w}{2}, \frac{H-h}{2})\}$, which correspond to the top-left, top-right, bottom-left, bottom-right and center regions of the images, respectively. The VGG19 [23] is used as the feature extractor Ψ (Eq. 9), with input size $(h, w) = (448, 448)$.

The model was implemented in Python 3 programming language [25] and using the Tensorflow 2 [26] and Keras 2 [27] frameworks. We trained the model using the Adam optimizer [28] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a fixed learning rate of 10^{-4} . We use full-sized (i.e., no crops) images and resize them to 4096×2816 pixels, to allow the network to learn high-resolution features, and use batch size equals 1. We selected two different input pipelines regarding the choice of $K = 2$ frames from each scene to test the robustness of our model: (i) using the EVs -1 and +1 frames, as in [14], [7]; (ii) using the most under and over exposed frames for each scene (the EV value changes depending on the scene), as in [8], [9], [10]. We report results for 300 training epochs, but we also show in Sec. V-B that it can be diminished to 100 without major performance loss. The model was trained in a NVIDIA A100-SXM4 GPU with 40GB of memory, AMD EPYC 7J13 64-Core Processor CPU with 30 processing units, and 200GB of RAM.

C. Evaluation protocol

We evaluated our proposed method with 11 quantitative image quality metrics, and with runtime and memory usage measurements during model inference. Our results were compared with three traditional MEF methods²: Mertens [4], Fast YUV [1], and Ancuti et al. [6]; and five SOTA methods³ based on deep learning, namely: HoLoCo [8], MEFLUT [14], TransMEF [9], IFCNN [7], and SAMT-MEF [10]. We proceed to provide details regarding our evaluation methodology.

Quantitative evaluation. For evaluating the quality of the fused images, we employed nine full-reference metrics: Structural Similarity Index (SSIM) [29], Multi-scale Structural Similarity Index (MS-SSIM) [29], Peak signal-to-noise ratio (PSNR), ΔE CIEDE2000 (ΔE 2000) [30], Visual Information Fidelity (VIFp) [31], Feature Similarity Index Measure (FSIM) [32], Spectral Residual based Similarity (SR-SIM) [33], Visual Saliency-induced Index (VSI) [34], and Mean Deviation Similarity Index (MDSI) [35]; and two no-reference ones: MEF-SSIM [36] and Qc [37].

To ensure a fair comparison, as the SICE dataset includes images of varying resolutions (see Sec. IV-A), the predicted images from the tested MEF methods are resized to match the

²We used the implementations available at: <https://github.com/LucasKirsten/Benchmark-Image-Fusion>.

³We used the implementations provided in their official Github repositories.

resolution of their ground-truth counterparts before computing the metrics.

Computational resources evaluation. In order to establish a fair benchmark among our method and competitors, we converted all models to the ONNX 13 [38] format, which is a flexible and popular format for distributing deep learning models for inference purposes. Then, we used the converted models for inference with dummy inputs and measured runtime and memory usage on both CPU and GPU processors. The reported measurements are the average of 20 runs for each evaluation. We tested the models in a notebook with Intel(R) Xeon(R) W-11955M 2.60Hz CPU with 64GB of RAM and x64-based processor, and NVIDIA RTX A3000 Laptop GPU with 6GB of memory. The models were evaluated with input resolution of 512^2 , 768^2 , 1024^2 , 1280^2 , 1536^2 , 1792^2 , 2048^2 , and 4096^2 . However, some methods could not be evaluated on high resolutions due to hardware limitations (i.e., out of memory error).

For the mobile evaluation, we accessed the runtime and memory usage of MobileMEF using a smartphone with 4 GB of RAM, four 2.4 GHz Kryo 265 Gold and four 1.9 GHz Kryo 265 Silver processors, Snapdragon 680 4G Qualcomm SM6225 chipset, and Adreno 610 GPU. Our trained model was converted to the TensorFlow Lite format⁴ to run on Android devices. The measurements were performed with the TensorFlow Lite benchmark tool⁵ with 10 simulated runs using GPU, with a fixed image resolution of 4096×4096 pixels.

V. RESULTS AND DISCUSSIONS

A. Comparison with other works

We present the results for our quantitative evaluation in Tab. I, supported by visual results in Figs. 4 and 5. **MobileMEF achieves SOTA results for most of the full-reference metrics:** 6/9 using EVs -1 and +1 as input frames (second best in MS-SSIM and VSI), and 7/9 using the most under and over-exposed frames as input (second best in VIFp), suppressing the performance of methods that require 200 to 1700 times more operations (MACs column). These results hint the effectiveness of our method in producing high-quality fused images with minimal computational resources, suggesting that it is well-suited for applications requiring high-quality images without excessive computational costs, such as in smartphone cameras.

Although our method does not achieve the highest MEF-SSIM and Qc values, it still performs competitively. Nevertheless, it's important to note that *some SOTA methods incorporate one (or more) of the tested metrics in their loss function, potentially giving them an advantage in these metrics*. Moreover, no-reference metrics may hinder some intricacies over the predicted images compared to their ground-truth counterparts, such as penalizing regions that require the model to interpolate some content due to a lack of information in the input frames.

Regarding our visual results presented in Figs. 4 and 5, our method demonstrates noticeable improvements in visual quality. It consistently delivers images with better contrast and detail preservation, closely resembling the ground truth, especially when compared to alternatives like HoLoCo [8], MEFLUT [14], and IFCNN [7]. This superior performance is evident in the enhanced clarity and naturalness of the fused images, which better handle the challenges of diverse exposure settings.

The results for evaluating the computational resources are presented in Figs. 6 and 7 regarding runtime and memory, respectively. Our method achieves the lowest or near-lowest execution times across all resolutions and on both devices evaluation (CPU and GPU), even when compared to the MEFLUT [14] method that requires fewer operations (recall Tab. I). Specifically, our method is on average **1.67× faster on CPU** and **1.45× faster on GPU** compared to the fastest one on each resolution. Regarding the memory usage evaluation, our method consistently outperforms all others, with the lowest memory consumption across all tested image resolutions. Specifically, our method is on average **1.41× more memory efficient** compared to the most efficient one on each resolution.

This makes our method effective in terms of the quality of fused images and highly efficient and scalable, ideal for real-time applications and high-resolution image processing, especially for applications with limited computational resources, such as smartphones. Hence, in comparison to other SOTA methods, our method provides a superior balance of computational efficiency and high-quality performance.

B. Ablation studies

In order to assess the effectiveness of our proposals, we conduct ablations studies related to (i) the neural network architecture (Sec. III-B and III-C), (ii) the crop-like Gradient loss proposal (Sec. III-D), and (iii) the number of training epochs. For these studies, we trained our model with half input size resolution and using EV -1 and +1 as inputs, reduced the train epochs to 100 for studies (i) and (ii), and evaluated the model performance using only full-reference metrics, which we understand to be more reliable measures of performance when the ground-truth images are available. Specifically, we used SSIM, MS-SSIM, and PSNR, which are the most broadly used metrics for image quality assessment.

Network Architecture. We present the results for the neural network architecture ablation in Tab. II, where the “Baseline” method is the LPIENet [11] architecture with modified inputs to agree with the MEF pipeline (i.e., allow multiple YUV frame inputs); the “Filter Reduction (F.R.)” column refers to diminishing the number of filters proposed in the original LPIENet model (they use 16 as the initial filter size for all channels, whereas we used 4 and 8 for the Y, and UV inputs, respectively); the “Network Optimization (N.O.)” column refers to the optimizations presented in Section III-B; and the “SSF-Module” column refers to the usage of this new module as explained in Section III-C. The “Memory” and “Runtime” were computed using a mobile device, as described in Section IV-C.

⁴Available at: <https://www.tensorflow.org/lite/>

⁵Available at: <https://www.tensorflow.org/lite/performance/measurement>

TABLE I

QUANTITATIVE RESULTS COMPARING MOBILEMEF TO TRADITIONAL AND DEEP-LEARNING BASED SOTA METHODS USING TWO DIFFERENT EV FRAMES INPUT PIPELINES. THE MACS (G) COLUMN WAS COMPUTED USING AN IMAGE RESOLUTION OF 1280^2 . **BOLD** VALUES MARK THE BEST RESULTS, AND UNDERLINE MARK THE SECOND BEST ONES.

Method	Inputs	MACs (G) ↓	SSIM ↑	MS-SSIM ↑	PSNR ↑	MEF-SSIM ↑	Qc ↑	$\Delta E 2000$ ↓	VIFp ↑	F-SIM ↑	SR-SIM ↑	VSI ↑	MDSI ↓
Mertens	EVs -1 and -1	0.14	0.7327	0.8679	16.7615	0.9358	0.8619	13.4556	0.3955	0.8982	0.9332	0.9592	0.3646
Fast YUV		0.08	0.7299	0.8611	16.4935	0.9060	0.8549	14.8283	0.3839	0.8926	0.9314	0.9552	0.3756
Ancuti et al.		0.25	0.7543	0.8834	17.5773	0.9607	0.8927	12.8665	0.4149	0.9236	0.9460	0.9682	0.3403
HoLoCo		1004.41	<u>0.7973</u>	0.8972	19.9179	0.8890	0.8135	9.8006	0.3801	<u>0.9372</u>	<u>0.9645</u>	0.9784	<u>0.2997</u>
MEFLUT		0.18	0.6651	0.8323	14.5945	0.9490 [†]	0.8456	16.6180	0.3256	0.8859	0.9137	0.9546	0.3842
TransMEF		569.47	0.7461 [†]	0.8807	17.3411	0.9757	<u>0.8872</u>	12.8441	0.4118	0.9224	0.9447	0.9692	0.3370
IFCNN		213.18	0.7794	0.8697	19.0874 [†]	0.9440	<u>0.7885</u>	<u>9.6578</u>	0.4672	0.9163	0.9468	0.9719	0.3112
SAMT-MEF		1552.58	0.7489 [†]	0.8769	18.4657 [†]	0.9103	0.8458	12.0584	0.3339	0.9249	0.9511	0.9712	0.3215
MobileMEF		0.88	0.8270	0.8937	20.6532	0.8965	0.8010	9.3952	0.3676	0.9376	0.9647	0.9778	0.2990
Mertens		0.14	0.7838	0.8423	18.6549	0.9030	0.6602	10.5726	0.3064	0.8781	0.9158	0.9561	0.3738
Fast YUV	Most under and over exposed frames	0.08	0.7732	0.8213	17.9362	0.7865	0.6408	14.1170	0.2930	0.8601	0.9108	0.9458	0.3919
Ancuti et al.		0.25	0.7795	0.8197	18.0990	0.9048	<u>0.6696</u>	10.9890	0.2838	0.8489	0.8858	0.9486	0.4004
HoLoCo		1004.41	0.7951	0.8594	19.3602 [†]	0.8875 [†]	0.6551	9.4392	0.2880	<u>0.9180</u>	0.9516	0.9744	<u>0.3190</u>
MEFLUT		0.18	0.5454	0.7264	10.3333	0.9272 [†]	0.4878	25.9636	0.1703	0.7833	0.8312	0.9244	0.4365
TransMEF		569.47	0.8156 [†]	0.8357	18.6669	0.9668	0.6497	9.8191	0.2493	0.8778	0.9134	0.9584	0.3731
IFCNN		213.18	0.7646	<u>0.8640</u>	18.9990 [†]	<u>0.9652</u>	0.6897	10.3337	0.4014	0.9075	0.9296	0.9683	0.3263
SAMT-MEF		1552.58	0.8024 [†]	0.8596	19.7875 [†]	0.9076	0.6389	8.9166	0.2872	0.9026	0.9343	0.9698	0.3412
MobileMEF		0.88	0.7977	0.8799	20.7976	0.8957	0.6325	8.8717	0.3306	0.9320	0.9622	0.9771	0.3036

[†]The metric is also used in the method loss function formulation.

TABLE II

ABLATION STUDIES REGARDING THE PROPOSED NETWORK ARCHITECTURE: FILTER REDUCTION (F.R. COLUMN), NETWORK OPTIMIZATIONS (N.O. COLUMN), AND SSF MODULE (SSF COLUMN). THE MACS (G) COLUMN WAS COMPUTED USING AN IMAGE RESOLUTION OF 4096^2 . THE MEMORY AND RUNTIME COLUMNS WERE COMPUTED USING A MOBILE DEVICE. **BOLD** VALUES MARK THE BEST RESULTS, AND UNDERLINE MARK THE SECOND BEST ONES.

	F.R.	N.O.	SSF	MACs (G)	Memory (MB)	Runtime (μ s)	MS-SSIM	SSIM	PSNR
Baseline				88.93	1710.02	5.63E+06	<u>0.9087</u>	0.8532	18.6070
	✓			10.74	<u>1125.29</u>	1.49E+06	<u>0.9083</u>	<u>0.8612</u>	20.8147
	✓	✓		8.94	1100.66	1.72E+06	0.9069	0.8591	21.2006
MobileMEF	✓	✓	✓	<u>9.05</u>	1156.23	1.82E+06	0.9088	0.8681	21.3015

TABLE III

ABLATION RESULTS FOR THE PROPOSED GRADIENT LOSS, \mathcal{L}_{Grad} .

\mathcal{L}_{Grad}	MS-SSIM	SSIM	PSNR
Default	0.8867	0.8367	18.9027
Proposed	0.9088	0.8681	21.3015

Our complete network yields the best quality measures for all the full-reference metrics. Moreover, the proposed network optimizations reduced the number of required operations and the amount of memory usage, with a slight increase in runtime compared to the Baseline model with reduced filters. Specifically related to the SSF module, note that it slightly impacts the computational efforts, with a small increase of 1.2% of MACs, 5% on memory usage, and 3.4% on runtime. Nevertheless, such a small increase is justified due to the consistent improvement in all the quantitative metrics, and also supported by manual inspection of the results (i.e., qualitative analysis).

Crop-like Gradient Loss. The crop-like Gradient loss ablation results are presented in Tab. III. We observe an increase in all three full-reference metrics. Specifically, the MS-SSIM improves by 2.5%, the SSIM 3.8%, and the PSNR 12.7%.

Training Epochs. The results for the trained epochs ablation are presented in Tab. IV. We observe small differences among the model's performance on different epoch checkpoints, supported by the small standard deviation values of the metrics. This finding serves as an indicative of the model's

TABLE IV

EPOCHS ABLATION RESULTS. **BOLD** VALUES MARK THE BEST RESULTS, AND UNDERLINE MARK THE SECOND BEST ONES. THE FINAL ROWS CORRESPOND TO THE MEAN (MEAN) AND STANDARD DEVIATION (STD) VALUES, RESPECTIVELY.

Epochs	SSIM	MS-SSIM	PSNR
100	0.8242	0.8931	20.5245
200	0.8246	0.8926	20.5374
300	0.8270	0.8937	<u>20.6532</u>
400	0.8251	<u>0.8952</u>	20.8056
500	<u>0.8266</u>	0.8962	20.6528
MEAN	0.8255	0.8942	20.6347
STD	0.0013	0.0015	0.1135

robustness for overfitting.

VI. CONCLUSIONS

In this work, we introduced MobileMEF, a novel MEF method based on an Encoder-Decoder deep learning architecture optimized for fast and efficient image processing on mobile devices. Our extensive evaluations demonstrate that our method outperforms SOTA approaches in most tested full-reference quality measures and computational efficiency (runtime and memory usage), making it highly suitable for real-time applications on devices with limited resources, such as smartphones. Furthermore, our ablation studies confirmed the effectiveness of our architectural choices (i.e., the efficient building blocks and the SSF module), the crop-like Gradient loss proposal, and the robustness of our model across different training epochs. Finally, our work offers a robust solution for

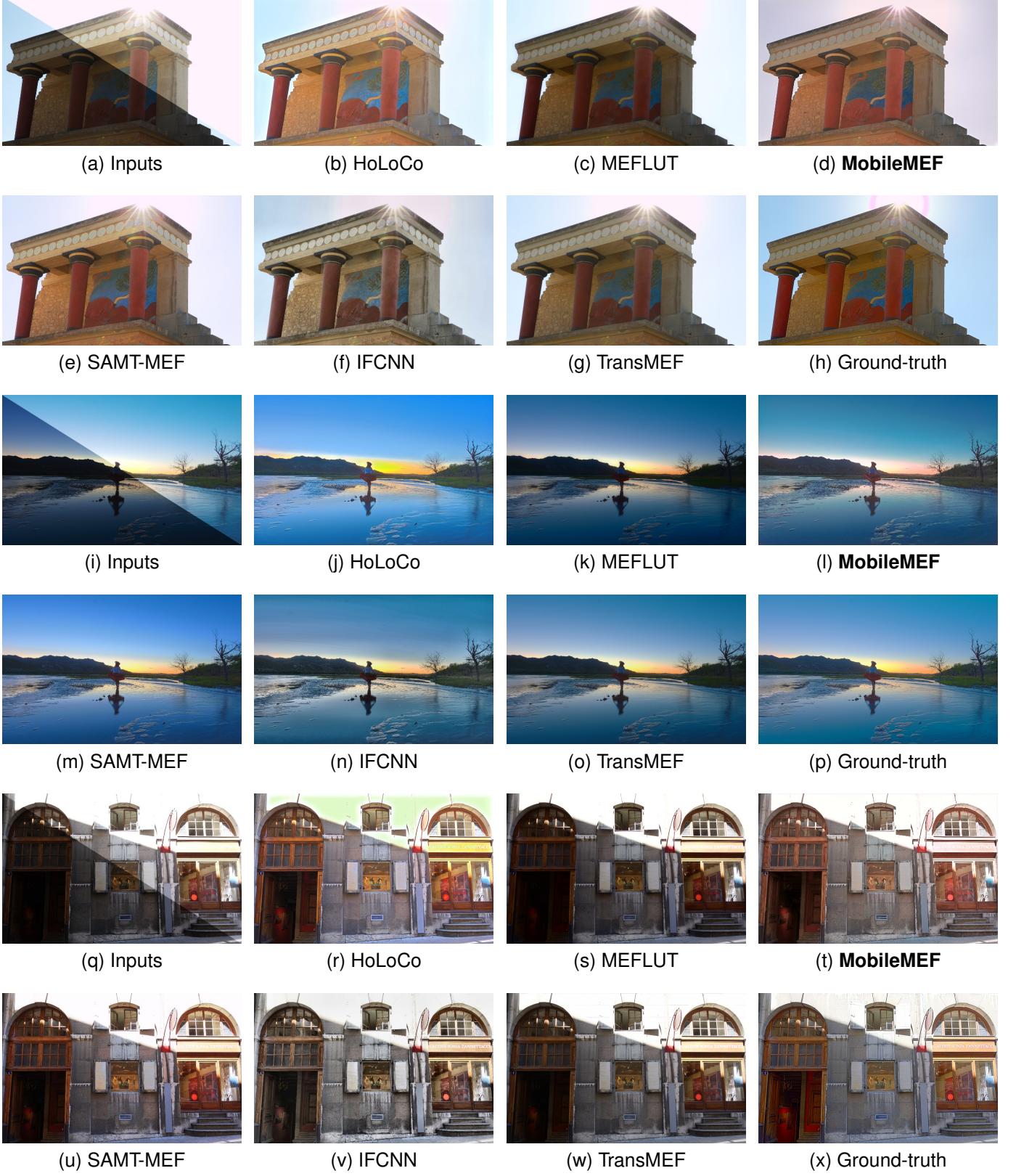


Fig. 4. Visual comparison of MobileMEF with SOTA methods using EV +1 and -1 as input frames.

MEF, balancing high image quality and low computational demands, thus paving the way for more efficient and effective HDR imaging on hardware-constrained applications. In future

works, we intend to enhance the model's adaptability to a wider range of input conditions to mitigate its current limitations.

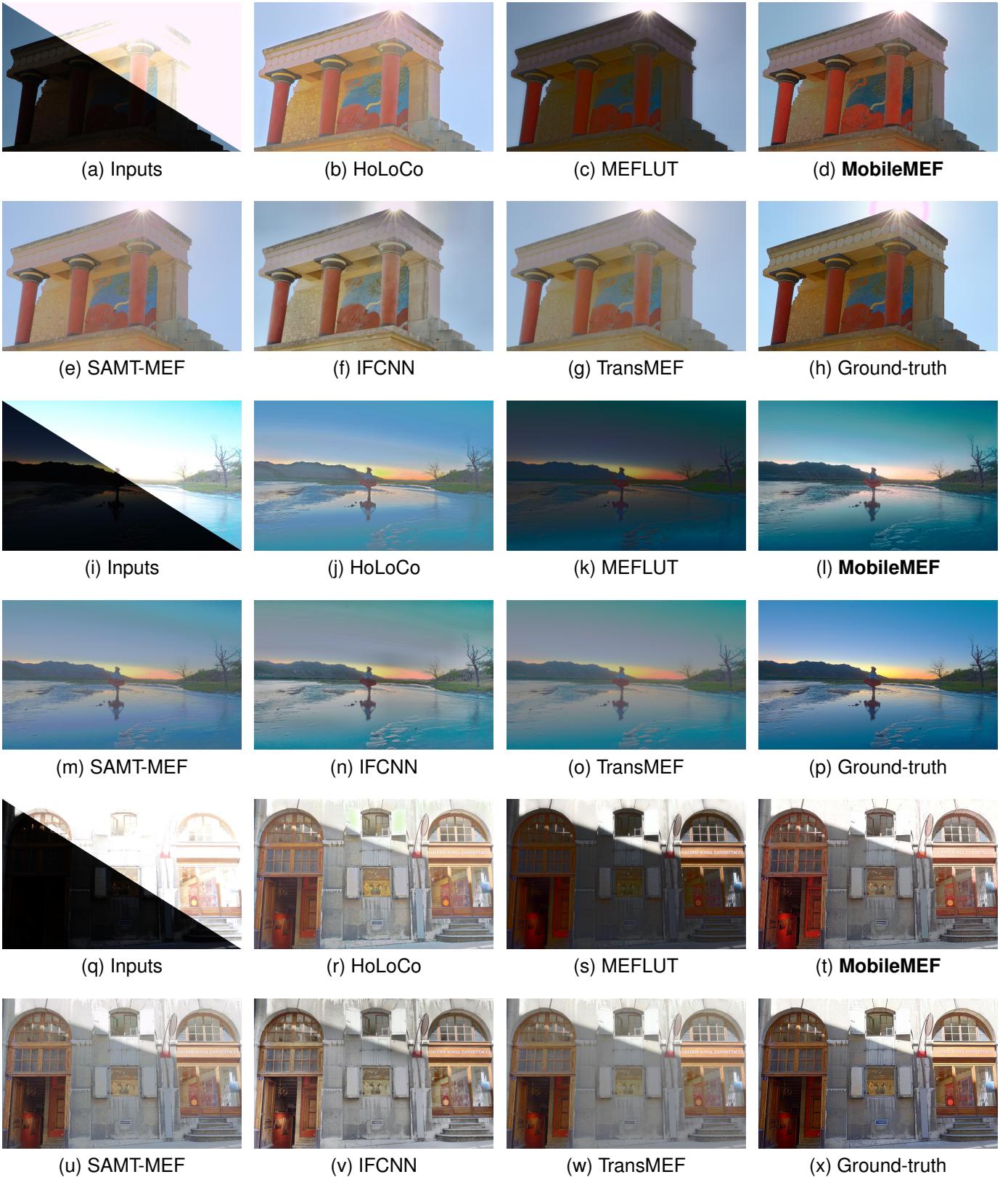


Fig. 5. Visual comparison of MobileMEF with SOTA methods using the most under and over exposed input frames.

REFERENCES

- [1] Y. Liu, D. Wang, J. Zhang, and X. Chen, "A fast fusion method for multi-exposure image in yuv color space," in *2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, 2018, pp. 1685–1689.
- [2] H. Kaur, D. Koundal, and V. Kadian, "Image fusion techniques: a survey," *Archives of computational methods in Engineering*, vol. 28,

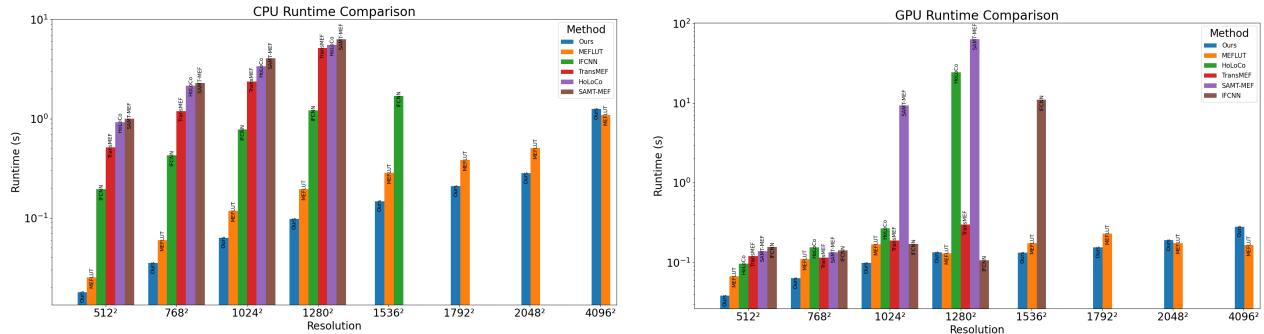


Fig. 6. Results for the runtime evaluations in CPU (left) and GPU (right).

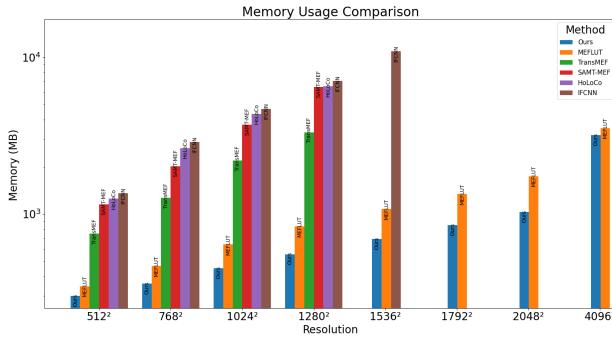


Fig. 7. Results for the memory evaluations.

pp. 4425–4447, 2021.

- [3] Y. Liu, X. Chen, Z. Wang, Z. J. Wang, R. K. Ward, and X. Wang, “Deep learning for pixel-level image fusion: Recent advances and future prospects,” *Information Fusion*, vol. 42, pp. 158–173, 2018.
- [4] T. Mertens, J. Kautz, and F. Van Reeth, “Exposure fusion,” in *15th Pacific Conference on Computer Graphics and Applications (PG’07)*. IEEE, 2007, pp. 382–390.
- [5] L. Kirsten, “Benchmark evaluation of image fusion algorithms for smartphone camera capture,” *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 11, no. 9, pp. 5363—5369, 2024.
- [6] C. O. Ancuti, C. Ancuti, C. De Vleeschouwer, and A. C. Bovik, “Single-scale fusion: An effective approach to merging images,” *IEEE Transactions on Image Processing*, vol. 26, no. 1, pp. 65–78, 2017.
- [7] 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.
- [8] J. Liu, G. Wu, J. Luan, Z. Jiang, R. Liu, and X. Fan, “Holoco: Holistic and local contrastive learning network for multi-exposure image fusion,” *Information Fusion*, vol. 95, pp. 237–249, 2023.
- [9] L. Qu, S. Liu, M. Wang, and Z. Song, “Transmef: A transformer-based multi-exposure image fusion framework using self-supervised multi-task learning,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, no. 2, 2022, pp. 2126–2134.
- [10] Q. Huang, G. Wu, Z. Jiang, W. Fan, B. Xu, and J. Liu, “Leveraging a self-adaptive mean teacher model for semi-supervised multi-exposure image fusion,” *Information Fusion*, p. 102534, 2024.
- [11] M. V. Conde, F. Vasluiianu, J. Vazquez-Corral, and R. Timofte, “Perceptual image enhancement for smartphone real-time applications,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 1848–1858.
- [12] A. Ghorbel, W. Hamidouche, and L. Morin, “Convnext-charm: Convnext-based transform for efficient neural image compression,” in *2023 11th European Workshop on Visual Information Processing (EUVIP)*. IEEE, 2023, pp. 1–6.
- [13] C. Ma, Y. Rao, Y. Cheng, C. Chen, J. Lu, and J. Zhou, “Structure-preserving super resolution with gradient guidance,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 7769–7778.

- [14] T. Jiang, C. Wang, X. Li, R. Li, H. Fan, and S. Liu, “Meflut: Unsupervised 1d lookup tables for multi-exposure image fusion,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 10542–10551.
- [15] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.
- [16] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “Mobilenetv2: Inverted residuals and linear bottlenecks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.
- [17] M. Goldblum, H. Souri, R. Ni, M. Shu, V. Prabhu, G. Somepalli, P. Chattopadhyay, M. Ibrahim, A. Bardes, J. Hoffman *et al.*, “Battle of the backbones: A large-scale comparison of pretrained models across computer vision tasks,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.
- [19] J. L. Ba, J. R. Kiros, and G. E. Hinton, “Layer normalization,” *arXiv preprint arXiv:1607.06450*, 2016.
- [20] D. Hendrycks and K. Gimpel, “Gaussian error linear units (gelus),” *arXiv preprint arXiv:1606.08415*, 2016.
- [21] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.
- [22] D. Lepcha, B. Goyal, A. Dogra, K. P. Sharma, and D. N. Gupta, “A deep journey into image enhancement: A survey of current and emerging trends,” *Information Fusion*, vol. 93, pp. 36–76, 2023.
- [23] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
- [24] J. Cai, S. Gu, and L. Zhang, “Learning a deep single image contrast enhancer from multi-exposure images,” *IEEE Transactions on Image Processing*, vol. 27, no. 4, pp. 2049–2062, 2018.
- [25] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009.
- [26] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, “Tensorflow: A system for large-scale machine learning,” in *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, 2016, pp. 265–283.
- [27] F. Chollet *et al.*, “Keras,” <https://keras.io>, 2015.
- [28] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [29] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.
- [30] M. R. Luo, G. Cui, and B. Rigg, “The development of the cie 2000 colour-difference formula: Ciede2000,” *Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur*, vol. 26, no. 5, pp. 340–350, 2001.
- [31] H. R. Sheikh and A. C. Bovik, “Image information and visual quality,”

- IEEE Transactions on image processing*, vol. 15, no. 2, pp. 430–444, 2006.
- [32] L. Zhang, L. Zhang, X. Mou, and D. Zhang, “Fsim: A feature similarity index for image quality assessment,” *IEEE transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.
 - [33] L. Zhang and H. Li, “Sr-sim: A fast and high performance iqa index based on spectral residual,” in *2012 19th IEEE international conference on image processing*. IEEE, 2012, pp. 1473–1476.
 - [34] L. Zhang, Y. Shen, and H. Li, “Vsi: A visual saliency-induced index for perceptual image quality assessment,” *IEEE Transactions on Image processing*, vol. 23, no. 10, pp. 4270–4281, 2014.
 - [35] H. Z. Nafchi, A. Shahkolaei, R. Hedjam, and M. Cheriet, “Mean deviation similarity index: Efficient and reliable full-reference image quality evaluator,” *Ieee Access*, vol. 4, pp. 5579–5590, 2016.
 - [36] K. Ma, K. Zeng, and Z. Wang, “Perceptual quality assessment for multi-exposure image fusion,” *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3345–3356, 2015.
 - [37] 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, pp. 178–182, 2005.
 - [38] J. Bai, F. Lu, K. Zhang *et al.*, “Onnx: Open neural network exchange,” <https://github.com/onnx/onnx>, 2019.