Baltic J. Modern Computing, Vol. 13 (2025), No. 4, pp. 933–957
https://doi.org/10.22364/bjmc.2025.13.4.10

ColorMEF: A Novel Transformer Based Multi-Exposure Fusion Model

Matiss LOCANS, Evalds URTANS

Riga Technical University, Riga, Latvia

matiss.locans@edu.rtu.lv, evalds.urtans@rtu.lv

ORCID 0000-0001-9813-054, ORCID 0000-0001-9813-0548

Abstract. This research presents a novel multi-exposure fusion model termed ColorMEF. In contrast to conventional approaches that rely primarily on image luminance data, ColorMEF integrates chromatic color information to augment the image fusion process. The incorporation of chromatic information enables ColorMEF to outperform existing models, as substantiated by evaluation metrics such as SSIM, DISTS, and VSI. Furthermore, the model is systematically trained using an end-to-end framework that optimizes the MEF-SSIM function based on the input data. ColorMEF achieves state-of-the-art performance on 4 of 5 conventional full-reference metrics: SSIM 0.9181 (+0.0092 versus the best prior), GMSD 0.0645 (-0.0043, $\downarrow 6.3\%$), DISTS 0.9266 (+0.0144), and VSI 0.9857 (+0.0025); on VIF ColorMEF is slightly lower (0.5224 vs 0.5394 for IFCNN). On the paper-introduced FSMEF-SSIM metric, ColorMEF ranks *second* (0.9435 vs 0.9465 for MEF-Net). These results indicate that coupling the chroma and luma during fusion improves structural fidelity and perceived quality across diverse scenes.

Keywords: Machine learning, Multiple exposure fusion, Vision transformers, Guided filtering, Chromatic colors

1 Introduction

The technique of Multiple Exposure Fusion (MEF) involving disparate exposure images represents a relatively nascent challenge within the domain of computer vision. By applying this technique with low dynamic range (LDR) images, it is possible to acquire high dynamic range (HDR) images, which contain more information than individual images and make the processing of said information much simpler because it is within a singular image space.

To fuse multiple exposure images, it is necessary to acquire them. It can be done by taking multiple LDR images with different exposure times. By changing the exposure time of the camera sensor, the amount of light to which the sensor is exposed changes.

The increased exposure time translates into longer light exposure, which translates into much brighter images. This can cause certain objects in captured images to be underexposed, causing them to appear much darker and indistinguishable from their surroundings, or overexposed, causing them to appear too bright, creating the same effect of underexposed areas but in the opposite direction.

Since LDR images can fail to capture all of the scene information in a singular image, by taking multiple LDR images with different exposure times and fusing them together, it is possible to attain a singular fused HDR image, which contains most visible features from individual exposures in a single image.

There are a multitude of solutions for the MEF task, spanning classical and deep learning methods alike, with deep learning methods getting better and outperforming their classical counterparts. However, most MEF methods still struggle with similar limitations. Most methods utilize color transformations of images to acquire luminance information from the said images. While luminance information is fused together by using the proposed deep learning architectures, color information is often relegated to being fused by a weighted sum operation. Although this fusion makes sense, as luminance contains much more structural and contrast information than colors do, there remains an unexplored avenue of color fusion with deep learning methods as well.

We propose ColorMEF, a transformer-based multi-exposure fusion model, trained in a self-supervised manner. The overall structure of the model is akin to an encoder-decoder network, and it is trained by using previously used multi-exposure fusion specific metrics for self-supervision, to learn input image feature fusion. The model works in an end-to-end manner, using the segmentation approach to generate weight maps for input exposure images. The weight maps are then applied on top of the input exposures, and they are fused by summing values.

Unlike other solutions, color information, which can be described as chromatic colors, is also fused inside the model. We adapt a transformer-based encoder architecture for global feature extraction from the image, as well as a CNN skip connection for small local feature extraction. We used parallel transformer units to achieve this goal, as experimental results show that color information can affect how luminance is weighed, affecting the final resulting image.

2 Related work

Classical MEF typically relies on multi-scale pyramid fusion (saliency weighted Laplacian or ratio of Gaussians pyramids),exposure-weighted blending (well-exposedness, local contrast, and saturation weights), gradient- or edge-preserving schemes (bilateral/guided filtering) to avoid halos, and optimization formulations that penalize seams and artifacts. These methods are fast and data-free, but struggle with misalignment/ghosting and often treat color as a passenger variable, blending it with the same scalar weights chosen for luminance; chroma-specific artifacts (bleeding, false color in saturated regions) are common.

In recent years, numerous deep learning-based multi-exposure fusion (MEF) solutions have emerged. The first deep learning-based MEF approach was proposed by (Prabhakar et al., 2017) named DeepFuse utilizing CNN features for image fusion. Un-

like classical algorithms, DeepFuse fuses images directly; however, it only fuses luminance information, while color information is fused using the weighted sum (Prabhakar and Babu, 2016). Various other solutions adapt this general approach to MEF by fusing luminance information directly and returning it from the model, while changing the underlying architecture to further try to improve fusion results. PGMI (Zhang, Xu, Xiao, Guo and Ma, 2020) utilizes individual branches for different exposures compared to DeepFuse (Prabhakar et al., 2017), while also utilizing pathwise transfer blocks to exchange information between them. However, this proposes a limitation, locking PGMI to be able to fuse only two images at a time. Other solutions simply use established CNN architectures, e.g. DenseNet (Huang et al., 2017) and fuse multiple exposure images by concatenating luminance information together in channel dimension, returning already fused image outputs directly from the model (Xu, Ma, Jiang, Guo and Ling, 2020), (Xu, Ma, Le, Jiang and Guo, 2020)

The previously mentioned approach has some drawbacks. Trusting the model to provide a complete MEF output is suboptimal. This means that models have to be trained on a large amount of data to be able to produce such output. Some of existing deep learning-based approaches are more similar to classical algorithms, generating weight maps in order to perform MEF (Ma et al., 2020), (Xu, Zhao, Wang, Zhang, Liu and Zhang, 2020). When weight maps are generated, the input images are used to generate the fused image by weighing them and then summing the resulting weighted images. This also allows the solutions to be manipulated with weights, for example, MEF-Net (Ma et al., 2020) uses two differing resolutions in the MEF task. Fusion weights are acquired from low-resolution images and, in conjunction with guided filtering (He et al., 2010), weights can be upscaled to higher resolutions, allowing for the fuse of higher resolution images.

Other existing MEF solutions also include the use of GAN (Goodfellow et al., 2014), such as MEF-GAN (Xu, Ma and Zhang, 2020). MEF-GAN in particular is set apart from other solutions by the fact that it fuses RGB images, instead of luminance information from YCbCr images. It is also one of the first models to use an attention mechanism; more specifically, it uses a self-attention mechanism similar to (Wang et al., 2017).

Transformer (Vaswani et al., 2017) architectures have proved to be powerful in a multitude of tasks. Naturally, their strength has also been preserved for vision tasks, e.g., vision transformers (ViT) (Dosovitskiy et al., 2020). One of the more recent solutions, TransMEF (Qu et al., 2021), utilizes the transformer architecture to extract global feature information. This lets the model overcome the receptive field issues that many CNN models have. However, due to the nature of ViT image patching, while they are very proficient at processing global feature patches, they lack the ability to discern the small features within the patches themselves. Because of this, a CNN-based local feature extractor can be used to compensate for these deficiencies. TransMEF (Qu et al., 2021), in particular, is not trained for MEF in a direct way, rather it is trained as a feature extractor. The MEF component is operationalized through a fusion rule, which, although it facilitates the simultaneous amalgamation of multiple exposure images, concurrently detracts from the overall performance and the resultant outcomes.

As mentioned previously, MEF datasets are hard to come by. One of the most prominent data sets used is SICE (Cai et al., 2018), which contains multiple scenes with generated ground truths. For those who wish to use different data for image fusion, it is necessary to obtain MEF data via other means. Another aspect of training data acquisition is the usage of other data sets in conjunction with data augmentation to distort the original image for input and use the original as ground truth (Qu et al., 2021). This can allow the use of large-scale natural image datasets, but the performance might be inhibited because distortion effects are not sufficient to model different exposure levels like they are in originally created images. In the absence of ground-truth data, a self-supervised training approach can be utilized for model training. One of the most popular MEF optimization functions, MEF-SSIM (Ma et al., 2015), has been used as an optimization parameter to train models self-supervised. MEF-SSIM is utilized to process input data and detect image regions of highest contrast, which then are used to train models for the MEF task, avoiding the need for ground truth images.

3 Methodology

The contribution of this work is color-guided luminance fusion that instead of fusing Y and then blending chroma, jointly processes Y and (Cb, Cr) in a full-parallel transformer, including a cross-attention branch that lets chroma steer which luminance structures to trust. Compared to brightness-only MEF, this reduces false structure selections near saturated colors and under white-balance shifts. Compared to post-hoc colorization/blending, it avoids chroma-luma inconsistencies by learning them together. Architecturally, we pair a DPT-style global transformer with CNN skip connections and per-channel deep guided filters for high-res weight upsampling. In this section, we take a look at the proposed model architecture and explain the choices made in the architecture and the philosophy behind them. We also look at the data used in the training and the training process itself.

3.1 Dataset

For training and validation, we use a closed source dataset, which contains about 7600 multi-exposure scenes. Each scene consists of three distinct exposure images, namely dark exposure b_0, middle exposure b_1, and bright exposure b_2. All images are of size 1667×1250 and correspond to a 4:3 aspect ratio. It is important to mention that this large-scale dataset does not contain ground truth images.

For testing and comparison with other methods, we use the SICE (Cai et al., 2018) dataset, which contains 589 multi-exposure scenes. We used the data split provided by the author of the dataset to determine the use of images for model testing. We take the validation split for the model testing. Three exposure images from each scene were sampled, so the model had limited available information during testing. The availability of ground truths in SICE comes as a boon, allowing us to utilize different full reference metrics in order to quantify model performance. However, since (Cai et al., 2018) state that reference images are generated using existing MEF and HDR solutions, this data set can only be used to compare the structural information of images primarily, leaving color information comparisons as a secondary objective.

3.2 Model

The proposed model ColorMEF, as mentioned before, takes inspiration from other previously mentioned methods, as well as different solutions from other computer vision subfields. It is publicly available in open source code [1]. The general approach for the MEF task is borrowed from (Ma et al., 2020). We first downsample the input sequence and make low-resolution weight-map predictions. Using guided filtering (He et al., 2010), we acquire high-resolution weight maps, which are then used for the generation of output exposure by multiplying them by the high-resolution input exposure sequence and adding the results. However, this general process is applied to all three image channels and not exclusively to luminance. The proposed model processes luminance and chromatic color information jointly and separately at different times. Also, chromatic colors are processed together since they represent a 2D point upon a color space. Our reasoning is that processing them jointly is easier for the model since it can fuse both values in conjunction, rather than fusing each of them separately. ColorMEF is trained end to end, optimizing the MEF-SSIM (Ma et al., 2015) criterion on top of high-resolution images.

ColorMEF adheres to the weight-map fusion paradigm as outlined in (Ma et al., 2020). Initially, low-resolution weights are estimated and subsequently upsampled to high resolution through the employment of deep guided filters (DGFs). The fusion of each channel is accomplished via weighted summation. Distinctly from preceding studies, we perform the fusion *luminance and chroma* within the network itself. A DPT-like U-shaped transformer (Ranftl et al., 2021) is utilized to extract global features, while shallow CNN skip pathways are employed to capture local details. A fully-parallel transformer block (Touvron et al., 2022) is implemented, comprising self-attention on Y, combined self-attention on (Cb, Cr), and cross-attention between the two, in order to achieve chroma-guided luminance. Prior to weight prediction, feature refinement is carried out through dictionary convolution units (DCUs) (Xu, Zhao, Wang, Zhang, Liu and Zhang, 2020).

3.2.1 Feature extraction block Weight map prediction can be defined as a segmentation task, since weight maps can be interpreted as segments of original images, where information evaluation is performed. Therefore, it makes sense to utilize segmentation networks for this approach. We use a dense prediction transformer (DPT) (Ranftl et al., 2021) to extract global image characteristics. This transformer is U-shaped and uses ViT (Dosovitskiy et al., 2020) for its encoder part, while CNN blocks are used in the decoder to rebuild images. DPT uses reassembly to reconstruct low-resolution image representations in reassembled blocks. These reconstructions are then used to build segmentation maps for each exposure back to the same low-resolution inputs.

In addition to using DPT for global feature extraction, we use parallel transformer blocks (Touvron et al., 2022) in our architecture. The complete parallel transfomer block consists of three MHA units and three feedforward networks. The visual representation of this block can be viewed in Fig 1 on page 938. This is where color information comes in. The image information is divided into two categories: luminance

[1] https://github.com/scrayish/ColorMEF

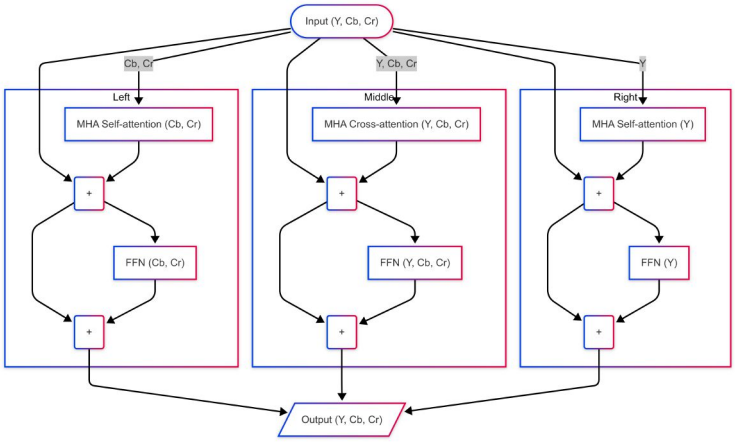


Fig. 1: Visual representation of full parallel transformer block used within DPT for extracting global features from luminance and chrominance information

or Y and chrominance or Cb or Cr. In the full parallel transformer block, we apply self-attention to each of these two categories separately. The third MHA is used for cross-attention between luminance and chrominance. Because luminance information between multiple exposures is more distinguishable than chrominance information, by utilizing cross-attention, it ought to be possible to guide chrominance information in some way to achieve better image fusion results. However, in experiments, regardless of the direction of cross-attention, chrominance information has a profound impact on luminance fusion results, since all of the information is deeply intertwined, therefore changing the overall fused image appearance.

As mentioned above, ViTs are good for global information acquisition, since they do not suffer from receptive field problems as CNNs. However, they are limited when it comes to local features. In a similar fashion to TransMEF (Qu et al., 2021), we use a skip connection to extract local feature information. Our CNN feature module is much more classical in its approach, sporting a normalization and activation function for each convolution layer, of which there are three. While chromatic and luminance information is deeply intertwined inside the DPT, we keep them separate for low-level feature extraction, each information type being processed by its own skip connection.

Also similar to TransMEF (Qu et al., 2021), an enhance block is used to merge the global transformer features with the local features of the CNN module. However, the enhance block uses dictionary convolutional units (DCU) (Xu, Zhao, Wang, Zhang, Liu and Zhang, 2020) instead of regular convolutions. This allows us to refine segmentation map results by using a closer link to input information. Also, since chromatic information has a profound influence on luminance information fusion results, it is possible to wrestle back some of the luminance independence by using DCUs. We postulate that the superior clarity of luminance information, in terms of structural and contrastive attributes, compared to chromatic channel information, underpins this phenomenon. The enhancement block head is used to compress the feature information and gain a singu-

lar weight map for each exposure image. The same as skip connections, these enhance blocks are separate for luminance and chrominance information.

3.2.2 Guided image filtering After feature extraction, we applied guided image filtering (He et al., 2010) to enhance the weight maps obtained and upsample them to high resolution to create a high-resolution fusion image. Unlike the MEF-Net approach (Ma et al., 2020), we use a deep guided filter (DGF) (Wu et al., 2018). The deep guided filter utilizes convolutions in order to enhance the weight maps. We also utilize separate DGFs for each information channel, since the information between luminance and chrominance is very distinct. We also separate chrominance back in the Cb and Cr channels and apply a separate DGF to them to reduce the potential production of color artifacts, where a singular DGF is used for both color channels. The complete visual structure scheme of ColorMEF can be viewed in Fig 2 on page 939

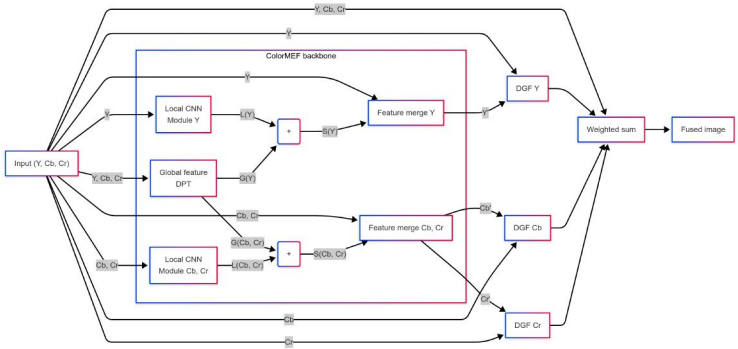


Fig. 2: ColorMEF model full schematic, including guided filtering modules and all necessary post-processing steps. All of these steps happen during model forward pass

3.2.3 Model training The model is trained on a proprietary dataset. A subset of 250 images is randomly selected for validation, while the remaining images comprise the training set. The training process utilizes the MEF-SSIM (Ma et al., 2015) criterion in its entirety. To prevent potential artifacts during training, each image channel, namely Y, Cb, and Cr, is processed separately using the MEF-SSIM criterion. The three loss components are assigned equal weightage ($static\_scale\_coeff = 1.0$, $chroma\_to\_luma\_coeff = 1.0$), and their summation is employed for backpropagation. The model is trained employing a single NVIDIA RTX A6000 GPU with 48 GB of video memory over approximately 200 epochs, with each epoch lasting approximately 2.5–3 hours. Under these conditions, the cumulative wall-clock training time is approximately 21 to 25 days. The optimization process is conducted using the Adam optimizer, starting with an initial learning rate of 1e-4, following a StepLR schedule that reduces the learning rate by 0.9 every 10 epochs (i.e., from 1.0e-4 to approximately

1.22e-5 by epoch 200). A smoothed running loss is calculated using an Exponential Moving Average (EMA) with a beta parameter of 0.9.

Input data are prepared at high and low resolution, similar to MEF-Net (Ma et al., 2020). For high resolution, we use the regular image resolution of 1667×1250 and for low resolution, we use 1280×960. This is a much higher resolution than is normally used for training. Also, since ViT is used for global image information extraction, the low resolution of the model is fixed. In order to train at different resolutions, model configuration should be updated to support said resolution. We also employ learnable positional embeddings, which lock the low resolution further in. It may be possible to train more dynamic-size images if the original transformer positional embedding proposed by (Vaswani et al., 2017) were used.

The training script employs randomized resizing with a high dimension of 1250 and a low dimension of 960. All images are transformed into the YCbCr color space, and the channels are split to facilitate fusion processes. Three exposure levels—dark, middle, and bright—are fused, with explicit supervision applied to both the luma (Y) and chroma (Cb/Cr) channels. During the inference and evaluation phases, the predicted RGB images and exposure weight maps for each sample are saved. The model utilizes a transformer-based architecture with an embedding dimension of 512, employs 8 attention heads, and has an enhancement depth of 3, adopting an identity readout mechanism. Padding is handled through PyTorch reflection padding, with an alternative option for circular panorama padding available. Training and evaluation are conducted sequentially over each epoch, with gradient computations enabled exclusively during training.

Higher resolution also means that more data are required. Although it is possible to train our model using the SICE (Cai et al., 2018) dataset, it is simply far too small for this kind of resolution. This is because a smaller resolution allows for efficient image cropping. By increasing the low-resolution model, image crops become larger, thus individual crops contain less and less unique information per crop, leading to degrading training. Another important aspect of transformer models is the general need for larger datasets than other network architectures. Insufficient data amounts can cause aggressive tiling artifacts in fused images, which the model learns to smooth out when larger data amounts are provided.

3.3 Full Structure MEF-SSIM

In addition to a new model architecture, we also introduce a variation of the MEF-SSIM (Ma et al., 2015) quality measurement metric. Although MEF-SSIM proves itself to be a formidable choice for self-supervised training, it mostly focuses only on the best contrastive information of the input images. Although this works in most cases, there may be cases where additional information from other exposures might be necessary to gain a more balanced fusion output. For this purpose, we propose the Full-Structure (FS) MEF-SSIM.

Algorithm 1 FSMEF-SSIM

1: Variance calculation $\sigma_{y_i}^2$ for input images $y_i \in Y$
2: Variance calculation σ_x^2 for fused image x
3: Covariance calculation σ_{xY} between mean values of fused image μx and input images μY
4: Quality map calculation between input images Y and fused image x with formula

$$\frac{2\sigma_{xy} + C2}{\sigma_x^2 + \sigma_Y^2 + C2}$$

5: Based on varinace calculation σ_Y^2 image segments are grouped into multiple sampling maps $k_i \in K$
6: Sampling maps are used to sample MEF-SSIM results $s_i \in S$ and sampled maps are multiplied with scaling coefficients and summed for final FSMEF-SSIM score

$$\sum_{i=1}^{n} s_i k_i, s \in S; k \in K$$

Fundamentally, this measurement works the same as MEF-SSIM. The contrast regions are calculated and ranked from the best to the worst, based on the returned values. However, with FSMEF-SSIM, all structural information is taken into account by weighting it. In this work, it is adjusted to work with three exposure images ($n = 3$), so weights for each structural segment ranging from best to worst are 0.9, 0.09 and 0.01, respectively. We still highly weigh the best regions with a much smaller contribution from the smallest structural regions. This is done because smaller contrast regions might contain large amounts of similar information with little to no detail, resulting in inferior fusion performance. But it can allow one to smooth out transitions between aggressive structural regions. If lower structural regions are weighed much higher, the final fusion image can become homogeneous, which is an undesirable result, since it can make the image look foggy and lose detail.

This kind of approach can help with the fusing of images with larger exposure time differences. Applying FSMEF-SSIM to an exposure stack of images with a smaller exposure time gap will not give large improvements in results, as the images already are spaced close enough to yield good enough detail coverage for successful fusion.

4 Results

As mentioned above, we used the SICE (Cai et al., 2018) dataset validation split as images to test ColorMEF. We avoid using MEF-SSIM for comparison, as it was used as an optimization criterion during model training. Instead, we utilize ground-truth images and we employ full-reference metrics in order to quantify our solution.

We also use other solutions, MEF-Net (Ma et al., 2020), TransMEF (Qu et al., 2021), CSC-MEFN (Xu, Zhao, Wang, Zhang, Liu and Zhang, 2020) and IFCNN (Zhang, Liu, Sun, Yan, Zhao and Zhang, 2020) for comparison. We used the trained weights provided by the author and source code to replicate selected fusion images using their

solutions. We only make small adjustments to make the code work. No changes have been made to the parameters or the model architecture that could affect the results.

4.1 Quantitative results

In this subsection, we take a look at quantitative results. We employ five different metrics for the determination of image quality: SSIM (Wang et al., 2004), VIF (Sheikh and Bovik, 2006), GMSD (Xue et al., 2013), VSI (Zhang et al., 2014), and DISTS (Ding et al., 2020). Each metric looks at image structure quality, although each does it in a different way. We find it valuable to evaluate the quality of the image based on different calculation approaches.

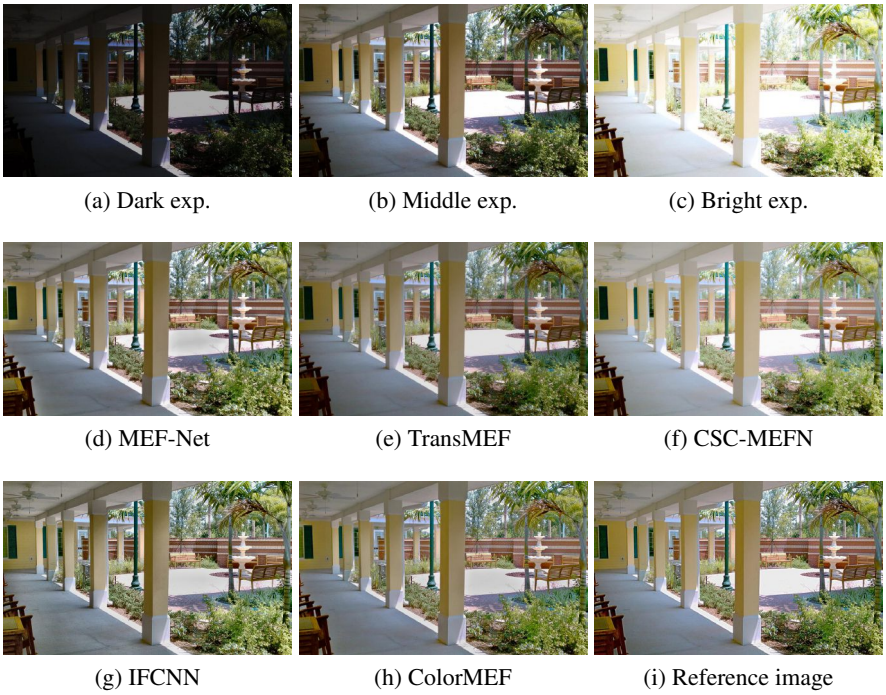
Table 1: Average metric result of all SICE validation test set for all calculated metrics of all tested models. The best results for each metric are marked in bold, while the second-best result is underlined. The star notation indicates the model and metric introduced in this paper

MEF method	MEFNET	TransMEF	CSC-MEFN	IFCNN	ColorMEF*
FSMEF-SSIM*	**0.94646**	0.91000	0.93124	0.88927	0.94348
SSIM	0.90890	0.84143	0.88284	0.87787	**0.91807**
VIF	0.52473	0.44590	0.47646	**0.53935**	0.52240
GMSD	0.06884	0.09076	0.08044	0.08716	**0.06451**
DISTS	0.91214	0.90273	0.90772	0.89346	**0.92656**
VSI	0.98314	0.97694	0.98067	0.97809	**0.98568**

1 contains all average metric values for each quality evaluation criterion tested. We can see that the colorMEF average metric values are best for four of five quality metrics. VIF is the only metric where IFCNN outperforms ColorMEF. Although ColorMEF achieves best results in a multitude of metrics, it is worth mentioning that all models, apart from TransMEF achieve very competitive metric evaluation. This means that all of the models are comparable and can fuse images well to an extent. We also reiterate that (Cai et al., 2018) used existing MEF solutions as well as HDR solutions for ground truth generation. In this way, it is possible to maximize structural information from images, as well as contrast. However, certain details remain subjective, such as coloring, saturation, hue, etc. For more precise results, a qualitative evaluation is also necessary.

4.2 Qualitative results

As seen in Fig 3 on page 943, ColorMEF achieves a balanced fusion between all three exposures. It is not as aggressive as IFCNN (Zhang, Liu, Sun, Yan, Zhao and Zhang, 2020), which mixes shadows and small details on the ground very aggressively. The chairs in the bottom left corner are also well fused, while MEF-Net (Ma et al., 2020) produces a light halo around them. While the reference image sports a high color saturation, ColorMEF tones colors a bit down, while remaining fairly vibrant. It also does not



(a) Dark exp. (b) Middle exp. (c) Bright exp.

(d) MEF-Net (e) TransMEF (f) CSC-MEFN

(g) IFCNN (h) ColorMEF (i) Reference image

Fig. 3: ColorMEF result comparison with other deep learning based MEF solutions. (a) - (c) is the input sequence, (d) - (h) are the model fused images, and (i) is the reference image for this fusion sequence

.

suffer from much darkening as with TransMEF (Qu et al., 2021), or too much brightness as CSC-MEFN (Xu, Zhao, Wang, Zhang, Liu and Zhang, 2020).

Fig 4 on page 944 is another example sequence of fusion. In this specific exposure sequence. Although this example uses much different exposure times, it is simpler because you are exposed indoors. It contains some details, which are conveniently clustered.

Inspecting the resulting images, it is apparent that all the models are good enough to fuse this image. By comparing with the reference image, we also deduce that certain models can fuse features better than the reference image contains. In this picture, MEF-Net (Ma et al., 2020) and ColorMEF have fairly similar results. TransMEF (Qu et al., 2021) is darker than others, while CSC-MEFN (Xu, Zhao, Wang, Zhang, Liu and Zhang, 2020) is brighter. IFCNN (Zhang, Liu, Sun, Yan, Zhao and Zhang, 2020) contains very aggressive details, which yields a high structure score, but can degrade the quality of the image, making it look messy.

Apart from influencing the fusion of luminance information, ColorMEF's approach of fusing colors also allows it to correct colors depending on input exposures and input scene. Although not as saturated as reference image colors, ColorMEF can reduce the

(a) Dark exp. (b) Middle exp. (c) Bright exp.

(d) MEF-Net (e) TransMEF (f) CSC-MEFN

(g) IFCNN (h) ColorMEF (i) Reference image

Fig. 4: ColorMEF result comparison with other deep learning based MEF solutions. (a) - (c) is the input sequence, (d) - (h) are the merged images of the model, and (i) is the reference image for this fusion sequence

.

blandness of images, which makes them look flat, and can impact the ease of postprocessing. The fused image also looks warmer than other solution images.

One caveat to including colors inside the network is their influence on luminance information fusion. While we employ cross-attention in a way that colors should follow luminance, the opposite effect could be observed at times. By increasing color importance and influence, it could be possible to gain more saturation and vibrance at the cost of detailing. This avenue has yet to be explored properly.

To gain additional insight on the qualitative performance of the models, we collect the mean opinion score through a survey. We surveyed 47 respondents and asked them to evaluate 9 image sets. Each set contains the fused image of each mdodel. We do not provide a reference image in this comparison, since our aim is to get an opinion on each model performance. The mean quality score (MOS) for each individual sequence and the average score of all 9 sequences can be viewed in 2.

We asked each image to be evaluated on a scale of 1 to 5, where 1 is a very low quality image and 5 is very high quality image. Based on observations from MOS, it is

Table 2: Mean opinion score for each compared model generated image inside provided image sequence. The top score for each image is marked in bold, while the second score is underlined

	MEF-Net	TransMEF	CSC-MEFN	IFCNN	ColorMEF
Sequence 1	3.59574	2.14894	2.44681	**3.91489**	3.63830
Sequence 2	2.89362	1.68085	2.70213	**3.93617**	3.44681
Sequence 3	3.65957	2.14894	2.89362	**4.08511**	3.55319
Sequence 4	3.46809	2.31915	3.12766	3.51064	**3.70213**
Sequence 5	3.48936	2.55319	3.02128	**4.34043**	3.46809
Sequence 6	2.97872	2.38298	3.00000	**3.97872**	3.19149
Sequence 7	3.57447	2.78723	2.36170	**4.38298**	3.14894
Sequence 8	3.70213	2.85106	2.53191	**4.21277**	3.06383
Sequence 9	3.46809	3.23404	2.27660	**4.12766**	3.38298
Average	3.42553	2.45626	2.70686	**4.05437**	3.39953

apparent that IFCNN (Zhang, Liu, Sun, Yan, Zhao and Zhang, 2020) has received the highest evaluation score of 4.05. This can be explained by its aggressive fusion, which makes the resulting images highly detailed. However, it can also cause harsh artifacts in coloring, as was observed in our own qualitative analysis.

Both ColorMEF and MEF-Net (Ma et al., 2020) are very close in second and third place, differing by approximately 0.03 MOS, which is a very small margin. The similar evaluation scores between these two models make sense, as was observed in qualitative result analysis. However, both methods trail IFCNN by more than 0.6 MOS, which is a substantial margin. Especially since both ColorMEF and MEF-Net scored very highly in quantitative results when comparing fused images to reference images.

Both TransMEF (Qu et al., 2021) and CSC-MEFN (Xu, Zhao, Wang, Zhang, Liu and Zhang, 2020) scored lower in the survey. In qualitative analysis, it was observed that these two models produced the smoothest images out of all five models tested. Their fusion images contained the least amount of small detailing, which could be the determining factor for lower MOS values in comparison to other models, which had a higher amount of small detailing.

In summary, small details and image sharpness can be compiled as the two most important aspects for image quality estimation in people testing. While coloring can impact the quality of an image, it seemingly does not play as large a role in image quality evaluation. In qualitative result analysis, we concluded that IFCNN and MEF-Net images are less saturated than those generated by other models, but score higher MOS because of the higher detail. Although ColorMEF scored the highest in most quantitative evaluations, it is trailing in the middle behind the previously mentioned models.

5 Further research

MEF is a method for obtaining HDR images by combining multiple LDR images. However, there may be some pitfalls to this approach. After MEF, the resulting image is usu-

ally still an LDR image, despite containing more details. Sometimes, users may prefer to perform additional post-processing, such as tone mapping.

ColorMEF is different from other models in that it fuses color information within itself along with luminance information. As mentioned in Section 3, we used MEF-SSIM (Ma et al., 2015) to calculate the loss of the color channel. Although it seems to work in this case, it may be suboptimal to use a structural metric to combine the color information. In an optical inspection of the YCbCr channels of an image, it is apparent that the Y channel contains the most information out of the three channels. Because it is equivalent to a grayscale image, this channel has the most pronounced structural and contrast information. While color channels also contain some contrast and structure, it is much weaker, as well as the mean value of all color information in channel is much more towards the middle of value range than in luminance. A wider theoretical knowledge of color spaces and color information can prove useful in further improving color fusion within neural networks.

When it comes to MEF, each additional image taken adds to the total computational cost of fusing images together. Although most of deep learning-based MEF solutions focus on two-image fusion, those images have to be fairly close to each other in terms of exposure times to fuse images successfully without excessive artifacts. Deep-learning models tend to have a strong bias for structural details given their training regimen. It is this bias that can lead to large artifacts in fused images, as models try to utilize much of the information from one image or the other, causing distinct artifact regions in fused images to degrade their quality. Different methods should be explored, which could let models truly fuse information between images, instead of doing hard-headed segmentation and fusion of most prominent feature areas, if there is a necessity to fuse images with larger difference of exposure times in order to keep image count lower.

6 Conclusions

In this paper, we proposed a novel deep learning method for MEF named ColorMEF. By utilizing and fusing image information within the model itself, we can acquire more balanced images, with reduced artifacts, and are better colors than their fused counterparts using other solutions, despite being trained with much larger resolution, which can provide a larger space for errors. Our outperform other model fused images based on objective metrics that use a reference photo. It showed an improvement of 25% using the GMSD metric, 4% using the SSIM and DISTS metrics, and outperformed other methods in 4 of 5 metrics. A novel model provides higher-quality images without additional hardware requirements at higher resolutions. We also give directions for further research, which could bring even more improvements to color fusion methods when applied together with deep learning-based solutions.

References

Cai, J., Gu, S., Zhang, L. (2018). Learning a deep single image contrast enhancer from multi-exposure images, *IEEE Transactions on Image Processing* **27**(4), 2049–2062.

Ding, K., Ma, K., Wang, S., Simoncelli, E. P. (2020). Image quality assessment: Unifying structure and texture similarity, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44**, 2567–2581.
https://api.semanticscholar.org/CorpusID:215785896

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale, *ArXiv* **abs/2010.11929**.
https://api.semanticscholar.org/CorpusID:225039882

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., Bengio, Y. (2014). Generative adversarial nets, *NIPS*.
https://api.semanticscholar.org/CorpusID:1033682

He, K., Sun, J., Tang, X. (2010). Guided image filtering, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **35**, 1397–1409.
https://api.semanticscholar.org/CorpusID:1264129

Huang, G., Liu, Z., van der Maaten, L., Weinberger, K. Q. (2017). Densely connected convolutional networks, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Ma, K., Duanmu, Z., Zhu, H., Fang, Y., Wang, Z. (2020). Deep guided learning for fast multi-exposure image fusion, *IEEE Transactions on Image Processing* **29**, 2808–2819.

Ma, K., Zeng, K., Wang, Z. (2015). Perceptual quality assessment for multi-exposure image fusion, *IEEE Transactions on Image Processing* **24**, 3345–3356.
https://api.semanticscholar.org/CorpusID:4828378

Prabhakar, K., Babu, R. V. (2016). Ghosting-free multi-exposure image fusion in gradient domain, *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* pp. 1766–1770.
https://api.semanticscholar.org/CorpusID:8764582

Prabhakar, K., Srikar, V. S., Babu, R. V. (2017). Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs, *2017 IEEE International Conference on Computer Vision (ICCV)* pp. 4724–4732.
https://api.semanticscholar.org/CorpusID:216738

Qu, L., Liu, S., Wang, M., Song, Z. (2021). Transmef: A transformer-based multi-exposure image fusion framework using self-supervised multi-task learning, *ArXiv* **abs/2112.01030**.
https://api.semanticscholar.org/CorpusID:244799167

Ranftl, R., Bochkovskiy, A., Koltun, V. (2021). Vision transformers for dense prediction, *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* pp. 12159–12168.
https://api.semanticscholar.org/CorpusID:232352612

Sheikh, H. R., Bovik, A. C. (2006). Image information and visual quality, *IEEE Transactions on Image Processing* **15**, 430–444.
https://api.semanticscholar.org/CorpusID:3716103

Touvron, H., Cord, M., El-Nouby, A., Verbeek, J., J'egou, H. (2022). Three things everyone should know about vision transformers, *ArXiv* **abs/2203.09795**.
https://api.semanticscholar.org/CorpusID:247594673

Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I. (2017). Attention is all you need, *NIPS*.
https://api.semanticscholar.org/CorpusID:13756489

Wang, X., Girshick, R. B., Gupta, A. K., He, K. (2017). Non-local neural networks, *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* pp. 7794–7803.
https://api.semanticscholar.org/CorpusID:4852647

Wang, Z., Bovik, A. C., Sheikh, H. R., Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity, *IEEE Transactions on Image Processing* **13**, 600–612.
https://api.semanticscholar.org/CorpusID:207761262

Wu, H., Zheng, S., Zhang, J., Huang, K. (2018). Fast end-to-end trainable guided filter, *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* pp. 1838–1847.
https://api.semanticscholar.org/CorpusID:3936783

Xu, H., Ma, J., Jiang, J., Guo, X., Ling, H. (2020). U2fusion: A unified unsupervised image fusion network, *IEEE Transactions on Pattern Analysis and Machine Intelligence* .

Xu, H., Ma, J., Le, Z., Jiang, J., Guo, X. (2020). Fusiondn: A unified densely connected network for image fusion, *AAAI Conference on Artificial Intelligence*.
https://api.semanticscholar.org/CorpusID:213637621

Xu, H., Ma, J., Zhang, X.-P. (2020). Mef-gan: Multi-exposure image fusion via generative adversarial networks, *IEEE Transactions on Image Processing* **29**, 7203–7216.
https://api.semanticscholar.org/CorpusID:220470749

Xu, S., Zhao, Z., Wang, Y., Zhang, C., Liu, J., Zhang, J. (2020). Deep convolutional sparse coding networks for image fusion, *ArXiv* **abs/2005.08448**.
https://api.semanticscholar.org/CorpusID:218673456

Xue, W., Zhang, L., Mou, X., Bovik, A. C. (2013). Gradient magnitude similarity deviation: A highly efficient perceptual image quality index, *IEEE Transactions on Image Processing* **23**, 684–695.
https://api.semanticscholar.org/CorpusID:478859

Zhang, H., Xu, H., Xiao, Y., Guo, X., Ma, J. (2020). Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, pp. 12797–12804.

Zhang, L., Shen, Y., Li, H. (2014). Vsi: A visual saliency-induced index for perceptual image quality assessment, *IEEE Transactions on Image Processing* **23**, 4270–4281.
https://api.semanticscholar.org/CorpusID:2995883

Zhang, Y., Liu, Y., Sun, P., Yan, H., Zhao, X., Zhang, L. (2020). Ifcnn: A general image fusion framework based on convolutional neural network, *Inf. Fusion* **54**, 99–118.
https://api.semanticscholar.org/CorpusID:199677411

A Appendix

In this appendix, we show additional images of ColorMEF testing such as Fig 5 on page 949, Fig 6 on page 950, Fig 7 on page 951, Fig 8 on page 952, Fig 9 on page 953, Fig 10 on page 954, Fig 11 on page 955, Fig 12 on page 956. Fig 13 on page 957 and comparison with the models mentioned previously. Specifically, we show image collages presented in the survey to determine MOS (mean opinion score) for each model-fused output. The image collages are the same as those given to the respondents, and each collage image is numbered, where numbering means that the model used to make the output image. As discussed in section Section 3.1, we used the SICE dataset (Cai et al., 2018) for model evaluation, specifically, we used validation images from the proposed data split.

Fig. 5: Collage of fused images for SICE (Cai et al., 2018) sequence 46 (sequence 1 in 2). Images acquired using models by numbers: 1 - MEF-Net (Ma et al., 2020), 2 - TransMEF (Qu et al., 2021), 3 - CSC-MEFN (Xu, Zhao, Wang, Zhang, Liu and Zhang, 2020), 4 - IFCNN (Zhang, Liu, Sun, Yan, Zhao and Zhang, 2020), 5 - Ours (ColorMEF)

Fig. 6: Collage of fused images for SICE (Cai et al., 2018) sequence 62 (sequence 2 in 2). Images acquired using models by numbers: 1 - MEF-Net (Ma et al., 2020), 2 - TransMEF (Qu et al., 2021), 3 - CSC-MEFN (Xu, Zhao, Wang, Zhang, Liu and Zhang, 2020), 4 - IFCNN (Zhang, Liu, Sun, Yan, Zhao and Zhang, 2020), 5 - Ours (ColorMEF)

Fig. 7: Collage of fused images for SICE (Cai et al., 2018) sequence 56 (sequence 3 in 2). Images acquired using models by numbers: 1 - MEF-Net (Ma et al., 2020), 2 - TransMEF (Qu et al., 2021), 3 - CSC-MEFN (Xu, Zhao, Wang, Zhang, Liu and Zhang, 2020), 4 - IFCNN (Zhang, Liu, Sun, Yan, Zhao and Zhang, 2020), 5 - Ours (ColorMEF)

Fig. 8: Collage of fused images for SICE (Cai et al., 2018) sequence 102 (sequence 4 in 2). Images acquired using models by numbers: 1 - MEF-Net (Ma et al., 2020), 2 - TransMEF (Qu et al., 2021), 3 - CSC-MEFN (Xu, Zhao, Wang, Zhang, Liu and Zhang, 2020), 4 - IFCNN (Zhang, Liu, Sun, Yan, Zhao and Zhang, 2020), 5 - Ours (ColorMEF)

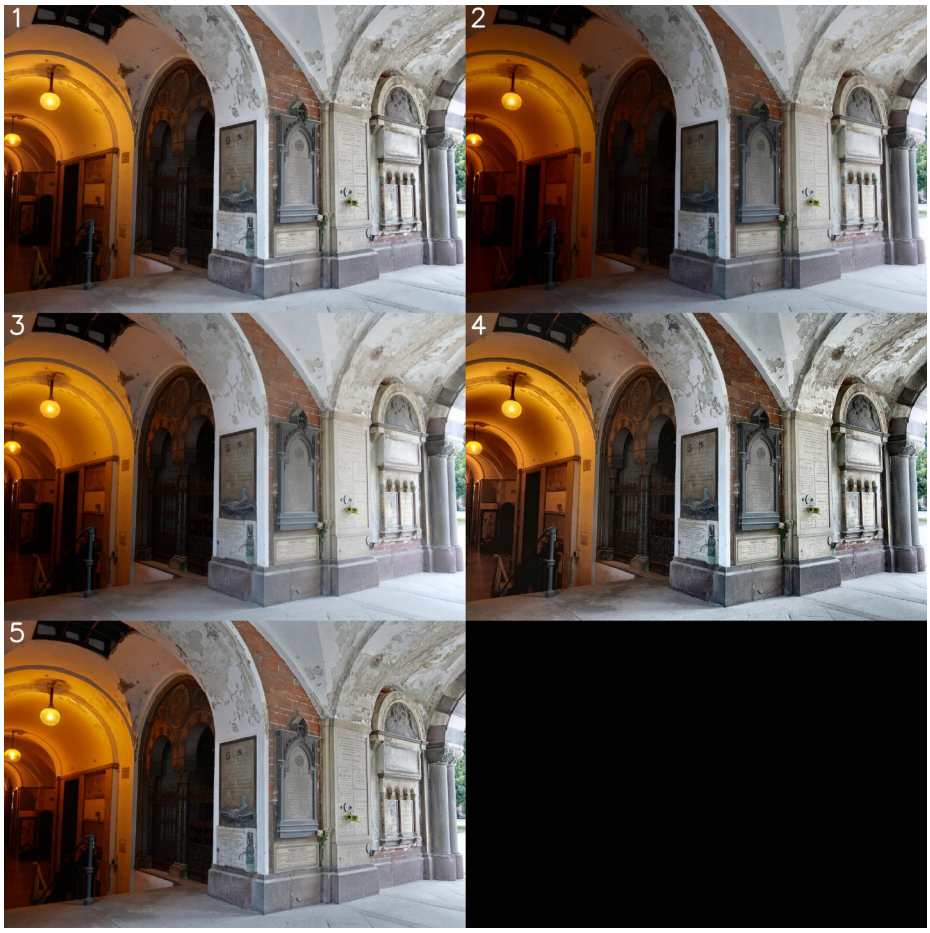


Fig. 9: Collage of fused images for SICE (Cai et al., 2018) sequence 28 (sequence 5 in 2). Images acquired using models by numbers: 1 - MEF-Net (Ma et al., 2020), 2 - TransMEF (Qu et al., 2021), 3 - CSC-MEFN (Xu, Zhao, Wang, Zhang, Liu and Zhang, 2020), 4 - IFCNN (Zhang, Liu, Sun, Yan, Zhao and Zhang, 2020), 5 - Ours (ColorMEF)

Fig. 10: Collage of fused images for SICE (Cai et al., 2018) sequence 78 (sequence 6 in 2). Images acquired using models by numbers: 1 - MEF-Net (Ma et al., 2020), 2 - TransMEF (Qu et al., 2021), 3 - CSC-MEFN (Xu, Zhao, Wang, Zhang, Liu and Zhang, 2020), 4 - IFCNN (Zhang, Liu, Sun, Yan, Zhao and Zhang, 2020), 5 - Ours (ColorMEF)

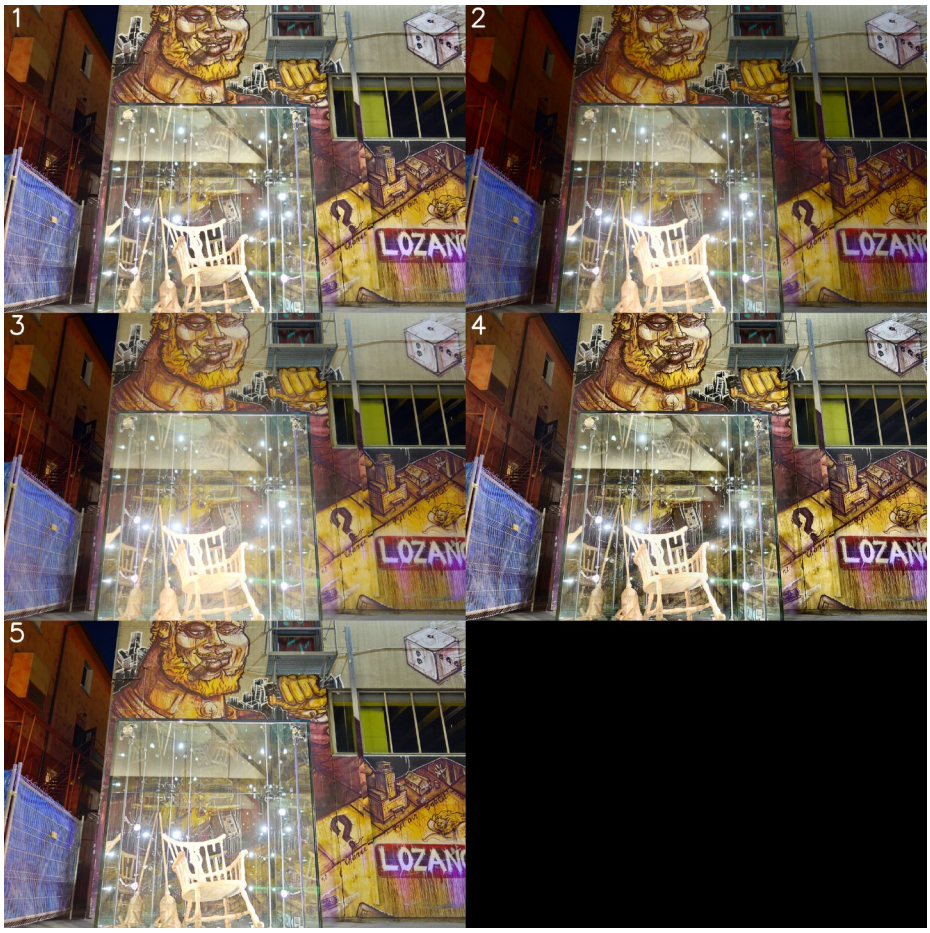


Fig. 11: Collage of fused images for SICE (Cai et al., 2018) sequence 58 (sequence 7 in 2). Images acquired using models by numbers: 1 - MEF-Net (Ma et al., 2020), 2 - TransMEF (Qu et al., 2021), 3 - CSC-MEFN (Xu, Zhao, Wang, Zhang, Liu and Zhang, 2020), 4 - IFCNN (Zhang, Liu, Sun, Yan, Zhao and Zhang, 2020), 5 - Ours (ColorMEF)

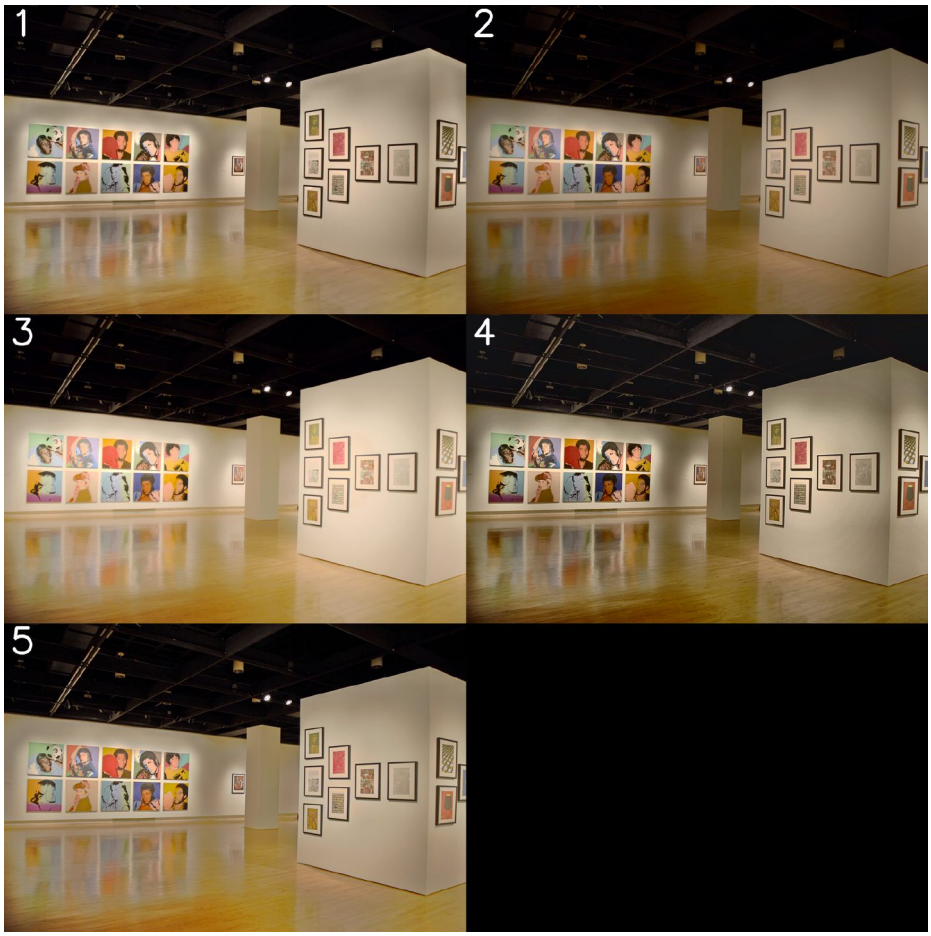


Fig. 12: Collage of fused images for SICE (Cai et al., 2018) sequence 57 (sequence 8 in 2). Images acquired using models by numbers: 1 - MEF-Net (Ma et al., 2020), 2 - TransMEF (Qu et al., 2021), 3 - CSC-MEFN (Xu, Zhao, Wang, Zhang, Liu and Zhang, 2020), 4 - IFCNN (Zhang, Liu, Sun, Yan, Zhao and Zhang, 2020), 5 - Ours (ColorMEF)

Fig. 13: Collage of fused images for SICE (Cai et al., 2018) sequence 52 (sequence 9 in 2). Images acquired using models by numbers: 1 - MEF-Net (Ma et al., 2020), 2 - TransMEF (Qu et al., 2021), 3 - CSC-MEFN (Xu, Zhao, Wang, Zhang, Liu and Zhang, 2020), 4 - IFCNN (Zhang, Liu, Sun, Yan, Zhao and Zhang, 2020), 5 - Ours (ColorMEF)

Received February 26, 2025 , revised September 15, 2025, accepted December 4, 2025