

# UltraFusion: Ultra High Dynamic Imaging using Exposure Fusion

## Supplementary Material

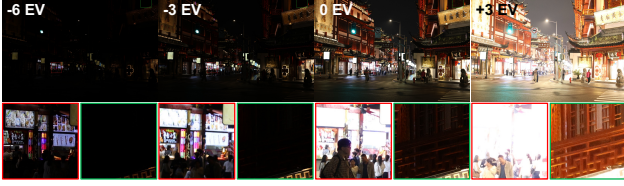


Figure A1. An example of 9-stops scenes.

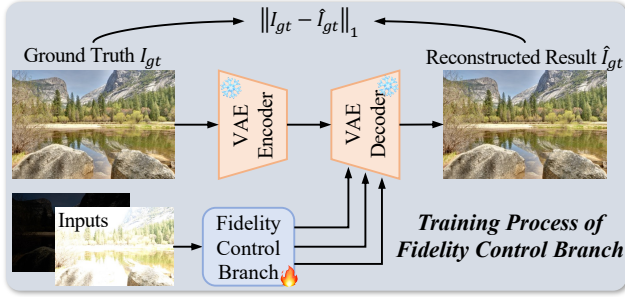


Figure A2. Detailed process of training FCB.

### A. Why We Need Handle 9-Stops?

Some challenging night-time scenes require up to 9 stops of exposure difference to cover the full dynamic range. As shown in Fig. A1, we need -6 EV to capture highlights (red box) and +3 EV (green box) to capture dark details.

### B. Training process of Fidelity Control Branch

To better illustrate how fidelity control branch is implemented, we show its training process in Fig. A2. Unlike the inference stage of our UltraFusion, the input of the VAE during FCB training is the ground truth. Our goal is to enable FCB to learn features that assist VAE decoding through shortcuts.

### C. Evaluation Details

In RealHDRV [39] dataset, the HDR ground truth corresponds to the 0 EV input. However, many 0 EV images in RealHDRV [39] dataset only contain few over-exposed regions need to be recovered. To better demonstrate ultra high dynamic imaging performance of various methods, we extract LDR image of 2 EV or 3 EV (according to the under-exposed input is -2 EV or -3 EV) from HDR groundtruth as over-exposed input, by reversing the process adopted to fuse the HDR groundtruth. Finally, after our augmentation, the RealHDRV [39] dataset contains 50 paired under/over-exposed inputs with 4 or 6 stops.

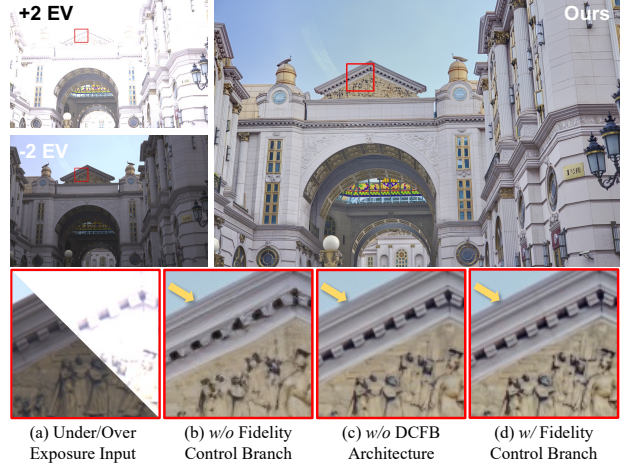


Figure A3. Illustrating the effectiveness of leveraging the similar architecture as decompose-and-fuse control branch in fidelity control branch.

### D. Ablation Study on Fidelity Control Branch

As shown in Fig. A3, the fidelity control branch effectively preserves the faithful structure of inputs. However, simply using two RGB images as inputs leads to some texture loss, as shown in Fig. A3(c). We demonstrate in Fig. A3(d) that by adopting similar architecture as decompose-and-fuse control branch (DCFB), more high-frequency details are retained and the overall visual quality is enhanced.

### E. Cross Attention Architecture

We utilize cross attention in the decompose-and-fuse control branch to fuse features from different modalities. The structure of the cross attention module is illustrated in Fig. A4. The cross attention module accepts three inputs, *i.e.*, overexposed image feature  $X_{oe} \in \mathbb{R}^{H \times W \times C}$ , short-exposed structural features  $X_{ue}^S \in \mathbb{R}^{H \times W \times C}$ , and short-exposed color features  $X_{ue}^C \in \mathbb{R}^{H \times W \times C}$ . First, we concatenate  $X_{ue}^S$  and  $X_{ue}^C$  and apply an  $1 \times 1$  convolution to adjust channel dimension back to  $C$ , obtaining the under exposure feature  $X_{ue}$ . Subsequently, LayerNorm is separately applied to  $X_{oe}$  and  $X_{ue}$ , followed by  $3 \times 3$  depth-wise convolutions to produce the corresponding  $Q$ ,  $K$  and  $V$ . Next, we perform attention operations on obtained  $Q$ ,  $K$  and  $V$ . After reshaping the output of attention operation, we input it to another  $1 \times 1$  convolution layer and add the result to  $X_{oe}$  to produce output condition feature  $X_{out}$ . The whole

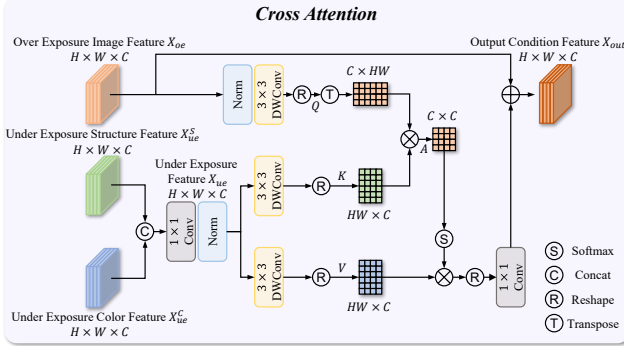


Figure A4. Detailed architecture of cross attention.



Figure A5. Visual comparison with SCTNet [44] on Kalantari's dataset [15]. Our framework can be extended to 3 exposures flexibly.

process can be summarized as follows:

$$X_{out} = X_{oe} + \text{Conv}_{1 \times 1}(V \text{Softmax}(\frac{Q^T K}{\tau})), \quad (\text{A1})$$

where  $\tau$  is a learnable scaling factor.

## F. Extend to 3 Exposures

Our UltraFusion is focus on 2 exposures as it already generates very good results and reduces the user's capture burden. Extending to 3 exposures is straightforward. We use the normal-exposed image as the reference and process it similarly. For the other two exposures, we extract guided features using the guidance extractor, then use normalized summation of them as input to the cross attention module. In the 3-exposure setup, we train UltraFusion on Kalantari's dataset [15] according to conventional settings and test on the corresponding test set. The comparison is performed with officially released state-of-the-art HDR reconstruction methods. The qualitative results are shown in Fig. A5, respectively.

## G. Effectiveness of Pre-Alignment

To conduct a more fair comparison, we also pre-align the test set and summarize the performance of each competing method in Tab. A1. Our UltraFusion still achieves the state-of-the-art performance. The consistent performance improvement of each method also demonstrates that the pre-alignment module is reasonable.

Table A1. Quantitative comparisons on RealHDRV [39] dataset.

Type	Method	RealHDRV			
		TMQI $\uparrow$	MUSIQ $\uparrow$	PAQ2PIQ $\uparrow$	HyperIQA $\uparrow$
HDR Rec.	HDR-Transformer	0.8710	63.30	70.99	0.5197
	SCTNet	0.8758	63.48	71.22	0.5222
	SAFNet	0.8789	62.88	70.91	0.5091
MEF	Defusion	0.8275	57.87	69.73	0.4974
	MEFLUT	0.8505	62.85	70.93	0.5073
	HSDS-MEF	0.8690	63.43	72.53	0.5272
Ours	UltraFusion	<b>0.8925</b>	<b>67.51</b>	<b>73.40</b>	<b>0.5833</b>



Figure A6. Comparing MEF-SSIM map with TC-MoA [69].

## H. Discussion on MEF-SSIM

MEF-SSIM is a widely used metric to evaluate fidelity after exposure fusion. However, sometimes lower MEF-SSIM does not indicate poor fidelity. As shown in Fig. A6, in brighter areas, ours UltraFusion achieves higher MEF-SSIM than TC-MoA, demonstrating high fidelity. In dark areas, it makes some necessary local adjustments, resulting in more natural transitions but lower MEF-SSIM.

## I. Compare with Inpainting Methods

To further illustrate our UltraFusion is the first guided inpainting model that can perform artifact-free HDR imaging, we compare our method with two diffusion-based image editing methods: Anydoor [4] and Stable Diffusion V2 Inpainting [37].

**Anydoor.** We compare our UltraFusion with an image customization method Anydoor [4]. Given a background image, a corresponding mask, and a reference image, AnyDoor can inpaint the reference into the masked region of the background image. Therefore, we utilize the over-exposed image as the background, mask out the over-exposed regions, and provide the contemporary regions from the under-exposed image as the reference. As shown in Fig. A7 (b), while AnyDoor can restore the highlight regions, the restored results fail to maintain consistency with the under-exposed image. Different from Anydoor, our UltraFusion effectively leverages the information from the under-exposed image, achieving a more reliable restoration.

**Stable Diffusion Inpainting.** Since Stable Diffusion V2 Inpainting [37] lacks the ability to fuse differently exposed inputs, we first obtain an initial fused result through a pre-alignment stage and our baseline model (*i.e.*, ControlNet [65]), as shown in Fig. A8 (d). Then, we use the estimated occlusion mask (Fig. A8 (c)) as the inpainting mask



Figure A7. Compare with an image customization method Anydoor [4]. Our method can preserve high-frequency details from the under-exposed image.



Figure A8. Visual comparisons with an inpainting method. We adopt Stable Diffusion V2 Inpainting [37] for comparison. All the inputs are resized to  $512 \times 512$  to meet the size requirement of the inpainting model.

for Stable Diffusion Inpainting to inpaint the occluded regions. It can be observed from Fig. A8 (e) that, although the artifact effect is mitigated, due to the absence of partial under-exposed information as guidance, the result from Stable Diffusion Inpainting fails to maintain consistency with the under-exposed image. Moreover, since Stable Diffusion Inpainting is not trained on our designed synthetic data, it is not robust to align errors, leading to further distortion in well-exposed regions. Finally, without a fidelity control branch, the overall structure of the image undergoes significant deformation. In contrast, our UltraFusion is able to generate a faithful and artifact-free output (Fig. A8 (f)).

## J. Additional Visual Comparisons

We provide additional visual comparisons on three datasets (*i.e.*, our UltraFusion benchmark, RealHDRV dataset [39] and MEFB dataset [66]). Please refer to our [project page](#).

For our benchmark, we present the results of our UltraFusion and competitors on 20 scenes used for the user study.

For the RealHDRV dataset, we selected 10 scenes with significant local motion. For the MEFB dataset, we randomly selected 10 scenes for visual comparison.