AdaptiveAE: An Adaptive Exposure Strategy for HDR Capturing in Dynamic Scenes

Supplementary Material

Tianyi Xu[1,3,4*] Fan Zhang[1] Boxin Shi[3,4†] Tianfan Xue[2,1†] Yujin Wang[1†]

[1]Shanghai AI Laboratory [2]The Chinese University of Hong Kong

[3]State Key Laboratory of Multimedia Information Processing, School of Computer Science, Peking University

[4]National Engineering Research Center of Visual Technology, School of Computer Science, Peking University

photon@stu.pku.edu.cn, zhangfan@pjlab.org.cn, shiboxin@pku.edu.cn

tfxue@ie.cuhk.edu.hk, wangyujin@pjlab.org.cn

A. More experimental results

A.1. More visual comparison results

As shown in Fig. A1, we present additional qualitative results on the HDRV [26] dataset. Our method strikes a balance between the impact of motion-related artifacts and the overall noise level, resulting in the best quality among all auto-exposure techniques.

Fig. A2 shows qualitative results for the ablation experiments on P_{ghost} on a case in the DeepHDRVideo [3] dataset; this penalty item effectively helps with mitigating ghosting and motion blur.

A.2. Cross post-processing methods test

To validate the importance of addressing motion blur and ghosting during auto-exposure, we compared our results with Wang *et al.* [32], applying deblur methods at various stages: before, during, and after exposure fusion. For fairness, we trained the deblur models [2, 7] on our HDRV-blur dataset, created by adding random motion blur to HDR images from the HDRV dataset using our synthesis pipeline. For pre-fusion deblurring, BANet [29], trained on HDRV-blur, was used to process the predicted LDRs before fusion with DeepHDR [7]. For fusion deblurring, we utilized DeepHDR's intrinsic deblurring ability, trained on HDRV-blur, without employing BANet. For post-fusion deblurring, BANet was applied after DeepHDR fusion. As shown in Tab. A1 and Fig. A3, post-capture deblur minimally reduces blur in the fused HDR image but degrades static regions, highlighting the efficacy of addressing blur during LDR capture.

A.3. Analyzing the role of ISO

In our experiments (Sec. 4), we set the ISO for fixed-ISO baselines to 200, as it serves as a standard choice in most scenarios. This raises the question of whether better results can be achieved by modifying the fixed ISO to an alternative value in the method proposed by Wang *et al*. [32]. To investigate this, we use Wang *et al*. [32] to predict the exposure values (EVs) for three low dynamic range images and systematically test all possible fixed-ISO settings to identify the value that maximizes the PSNR-μ on the test set. We denote this approach as W-optimal, where W refers to Wang *et al*. [32]. As illustrated in Tab. A2 and supported by the qualitative results in Fig. A4, utilizing the optimal fixed ISO results in slight performance improvement. However, this optimal ISO is highly dataset-specific and demonstrates very limited generalization capability, further validating the robustness and superiority of our proposed method over fixed-ISO approaches.

A.4. More discussions on inference time

Our RL agent executes in $<$5ms/scene on an NVIDIA RTX3080. The primary contributor to the latency, six LDR captures, can mostly be eliminated if we use the frames cached in the preview buffer, also known as the ZSL (Zero Slag Latency) buffer, which is the de facto standard for mobile phones. Utilizing an asynchronous camera driver, it has the potential to achieve real-time performance. In contrast, existing methods that use previously captured histograms for exposure prediction incur 30-70ms latency and are not robust to movement. Even without a viewfinder buffer, our inference speed can also be optimized with digital-overlap (DOL) sensors (to $<$100 ms/frame) and AE stats grid (around 32x24, downsampled from ISP 3A).

A.5. More frames

Our design of the reward and the step penalty (Eq. (7)) may result in a predicted exposure bracketing set containing more than three frames. Fig. A5 illustrates a case where our model makes a four-frame decision. Given our design of the step penalty, this occurs in only a small percentage of scenes with significant dynamic ranges and movements.

*This work was done during Tianyi Xu's internship at Shanghai AI Laboratory.

†Corresponding authors.

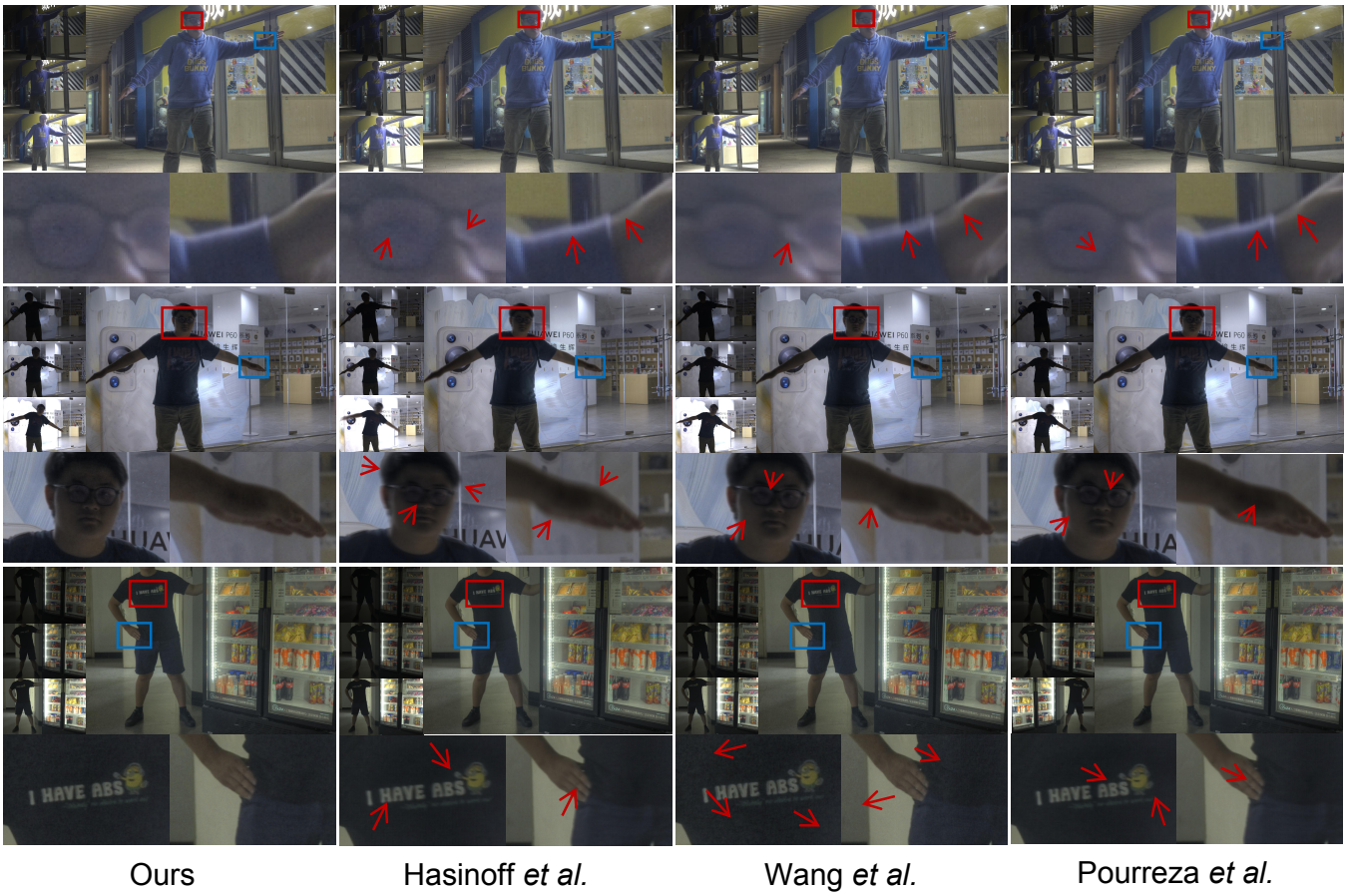

Figure A1. Additional qualitative comparisons with other auto-exposure methods on HDRV dataset [26]. Upper left: Predicted LDRs with varying ISO and shutter speed settings and synthesized using our image synthesis pipeline. Upper right: Fused HDR image using DeepHDR [7] and tone-mapped using Photomatix Enhancer. Below: Zoom-in results for tested methods.

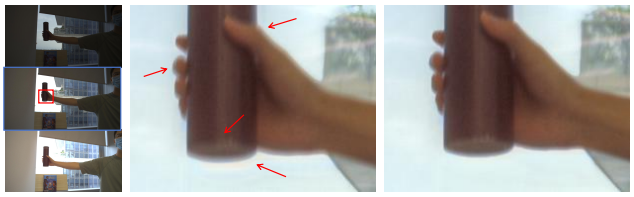

Figure A2. Effectiveness of ghost penalty.

B. Details of the noise synthesis model

We synthesize noise to the blurred HDR image b_j^L according to our noise model, which is based on [6]. The quantity each pixel measures is the radiance level Φ, in units of electrons per second. Therefore, the pixel value I of a raw image can be expressed as:

$$I = \min\left\{\frac{\Phi T \times \text{ISO}}{U} + I_0 + n, I_{\text{max}}\right\}, \qquad \text{(A1)}$$

Table A1. Results for ablation studies for different deblur post-processing techniques. We use Wang *et al.* [32] as the base model (denoted as W), and Pre-BA denotes using BANet [29] to process the LDRs before exposure fusion. BD denotes using blur-aware DeepHDR [7] for exposure fusion, which is trained on the HDRV-blur dataset we synthesized from the HDRV [26] dataset. Post-BA denotes using BANet to deblur the final tone-mapped HDR. **Bold**: The best.

Model	PSNR-μ	SSIM-μ	PU-PSNR	PU-SSIM
W	36.46	0.8902	32.68	0.8933
W+Pre-BA	37.33	0.9095	33.24	0.9100
W+BD	37.01	0.9016	32.83	0.8972
W+Post-BA	37.25	0.9124	32.88	0.9023
Ours	**39.70**	**0.9408**	**34.67**	**0.9465**

where T denotes the exposure time in seconds, U is a camera-dependent constant, I_0 represents the electrons created by dark current, n is the signal- and gain- dependent sensor noise and I_{max} indicates the full well capacity.

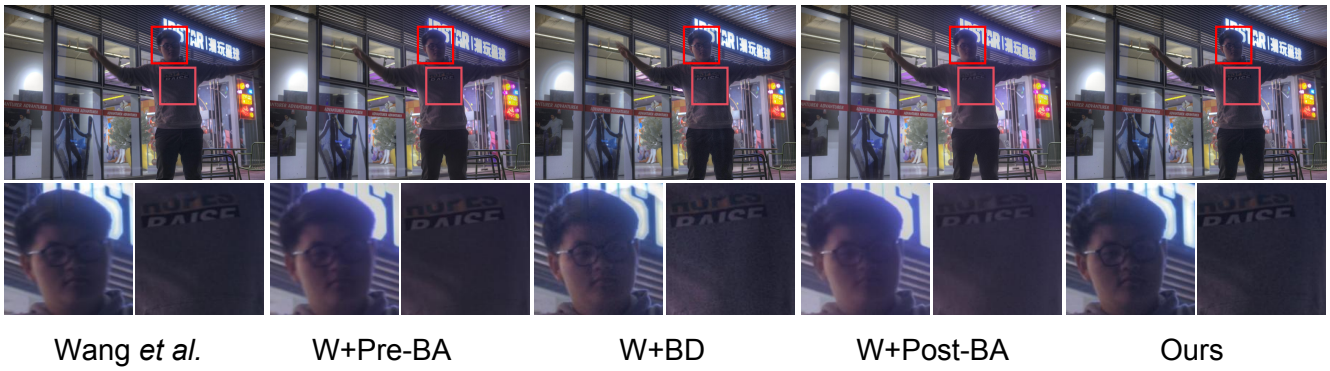

Figure A3. Necessity of considering motion blur during LDR capturing. We compared our method on the HDRV dataset [26] with Wang *et al*. [32] combined with different post-processing deblurring methods. W denotes Wang *et al*., Pre-BA denotes applying BANet [29] to LDRs for fusion. BD denotes DeepHDR [7] trained on the HDRV-blur dataset. Post-BA denotes applying BANet to the tone-mapped fused HDR result.

Figure A4. Ablation on the role of ISO in fixed-ISO methods. We choose the PSNR-optimal fixed-ISO for Wang *et al*. [32] (represented by W) that optimizes the PSNR of generated HDR on HDRV-Test dataset [26], denoted as W-optimal. Upper left: Predicted LDRs with varying ISO and shutter speed settings and synthesized using our image synthesis pipeline. Upper right: Fused HDR image using DeepHDR [7] and tone-mapped using Photomatix Enhancer. Below: Zoom-in results for tested methods.

Figure A5. An example scene for which our models give a 4-frame decision. 3-frame denotes a truncated version of the prediction, which has obvious ghosting patterns.

Conforming to the paradigm of [6], we model noise as a zero-mean variable, coming from three independent sources, including photon noise, which represents the Poisson distribution of photon arrivals and depends linearly upon the number of recorded electrons, ΦT, readout noise, which comes from sensor readout, and analog-to-digital conversion(ADC) noise, which comes from the combined effect of amplifier and quantization. Hence, for pixels below the saturation level, we have:

Table A2. Ablation study of the impact of ISO on fixed-ISO methods on HDRV [26] dataset. W denotes Wang *et al*. [32] and W-optimal denotes setting the fixed ISO of Wang *et al*. [32] to optimal value for SNR cross all available ISOs. **Bold**:best.

Model	PSNR-μ	SSIM-μ	PU-PSNR	PU-SSIM
W	36.46	0.8902	32.68	0.8933
W-optimal	37.64	0.9033	33.01	0.9058
Ours	**39.70**	**0.9208**	**34.67**	**0.9465**

$$Var(n) = \frac{\Phi T \times \mathrm{ISO}^2}{U^2} + \frac{\sigma_{\mathrm{read}}^2 \times \mathrm{ISO}^2}{U^2} + \sigma_{\mathrm{ADC}}^2. \quad \text{(A2)}$$

Note that the rationality of modeling ADC noise as independent of ISO lies in the fact that the quantization process, which could be represented by $q(x)$ in the following equation:

$$q(x) = \min(\lfloor x + 0.5 \rfloor, ADU), \quad \text{(A3)}$$

where ADU (Analog-to-Digital Units) denotes the maximum value that can be recorded by the sensor, for a target camera that records scenes as b-bits raw images, $ADU = 2^b - 1$, this q function is independent of ISO settings. The post-amplifier noise is also naturally independent of the foreground imaging settings.

Following our noise model, we can synthesize the corresponding noise with the decided ISO and shutter speed to the blurred HDR b_j^L, thereby obtaining the LDR image $l_j^{\mathcal{T}}$. This noise model facilitates the synthesis of LDR images with various ISO and shutter speed settings. Moreover, it accurately simulates the actual noise that arises in photography, helping our model to exhibit good generalization abilities on various datasets and real data.

Denoting the entire image synthesis process, which consists of motion blur synthesis and adding noise, as $\mathcal{S}$, and the corresponding LDR output as $l_j^{\mathcal{T}}$, we have:

$$l_j^{\mathcal{T}} = \mathcal{S}(f_i^{\mathcal{T}}, f_{i+1}^{\mathcal{T}}, (\text{ISO}_j, T_j)). \tag{A4}$$

where ISO_j and T_j are bracketed to denote that they are a pair of camera settings.

C. Network details

The architecture of our proposed AdaptiveAE network comprises two primary components: a Policy Network and a Value Network. The Policy Network is responsible for producing two output layers: one with 24 units for ISO selection and another with 19 units for shutter speed selection. Specifically, the ISO space consists of 24 possible settings—{50, 64, 80, 100, 125, 160, 200, 250, 320, 400, 500, 640, 800, 1000, 1250, 1600, 2000, 2500, 3200, 4000, 5000, 6400, 8000, 10000}—and the shutter speed space contains 19 possible values—{1/30, 1/40, 1/50, 1/60, 1/80, 1/100, 1/125, 1/160, 1/200, 1/250, 1/320, 1/400, 1/500, 1/640, 1/800, 1/1000, 1/1250, 1/1600, 1/2000}. The Policy Network employs softmax activation functions for both outputs, providing probability distributions over possible ISO and shutter speed configurations. In contrast, the Value Network outputs a single-unit layer, which estimates the state value. To ensure non-negative outputs, the Value Network incorporates a ReLU activation function. The separation of the Policy and Value Networks facilitates efficient decision-making by modeling both action distribution and state evaluation independently, allowing the system to adapt effectively to varying exposure conditions.

C.1. Semantic feature branch

The semantic feature branch leverages pre-trained AlexNet features, initially with a dimensionality of 4096. We apply this branch to the median-exposed LDR image from the input set. These semantic representations are transformed using a two-layer fully connected architecture. The first layer comprises 1024 neurons, while the second has 256 neurons, both with ReLU activation.

C.2. Irradiance feature branch

The irradiance feature branch processes exposure information from multiple LDR images by extracting histograms from each LDR image separately and concatenating them along the channel dimension. This multi-exposure histogram data is processed through three sequential 1D convolutional layers: the first with 128 filters, the second with 256 filters, and the third with 512 filters, all using a kernel size of 4 and a stride of 4. Following this, two fully connected layers with 1024 and 256 neurons, respectively, process the features, maintaining ReLU activation throughout.

C.3. Stage encoding branch

The stage encoding branch introduces a temporal dimension to the network by encoding both the current exposure iteration and the total planned exposures. It processes a two-dimensional input (current stage, total stages) through two layers: the first with 32 neurons and the second with 64 neurons, both activated by ReLU functions. This enhancement allows the network to adapt its strategy based on the remaining exposure budget.

C.4. Feature fusion mechanism

Features from the multiple LDR inputs, semantic, irradiance, and stage encoding branches are concatenated and processed through two fusion layers for comprehensive integration. The first fusion layer includes 512 neurons, followed by a second layer with 256 neurons, both using ReLU activation. This thorough fusion of features equips the network with the capacity to synthesize multi-modal information, thereby enhancing predictive accuracy.

Despite accepting multiple LDR inputs directly into each processing branch, the architecture maintains computational efficiency with approximately 7-8 million parameters, achieving inference times under 10 milliseconds, making it suitable for real-time applications in computational photography and image signal processing.