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

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

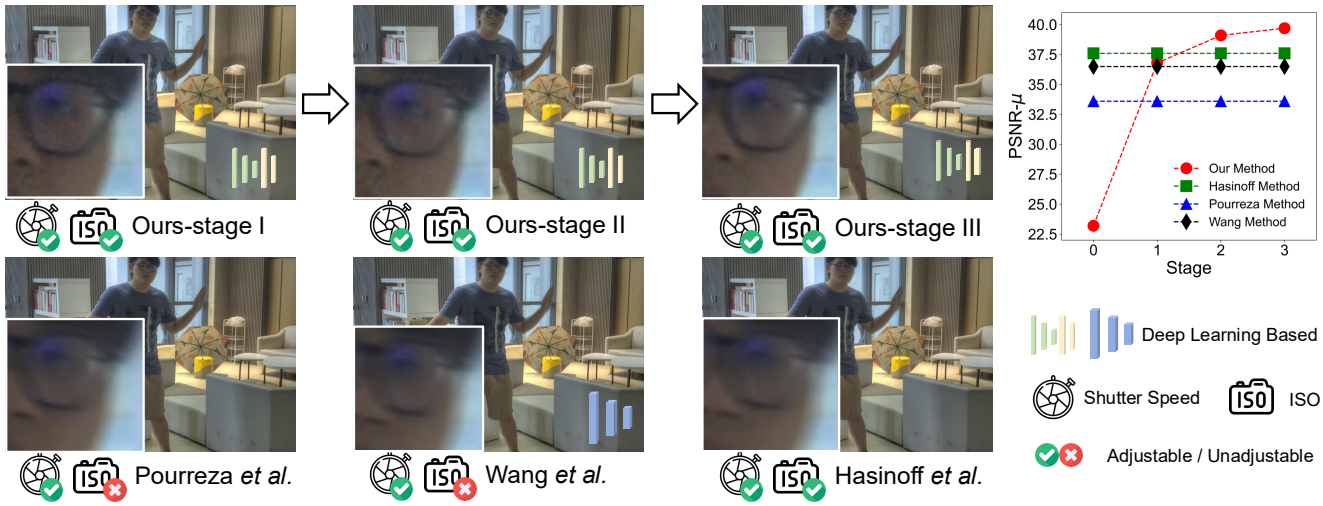


Figure 1. AdaptiveAE takes camera preview images as input and automatically predicts the ISO and shutter speed for each LDR captures for exposure fusion through a 3-stage sequential refinement procedure to achieve an optimal balance between noise level and motion-related problems for high quality HDR capturing in dynamic scenes with deep reinforcement learning. AdaptiveAE achieves PSNR 39.7 on HDRV dataset [26], while baseline methods [6, 21, 32] that either only predicts shutter speed or do not consider motion can only achieve PSNR below 37.6 and has evident motion blur and ghosting artifacts in HDR results.

Abstract

Mainstream high dynamic range imaging techniques typically rely on fusing multiple images captured with different exposure setups (shutter speed and ISO). A good balance between shutter speed and ISO is crucial for achieving high-quality HDR, as high ISO values introduce significant noise, while long shutter speeds can lead to noticeable motion blur. However, existing methods often overlook the complex interaction between shutter speed and ISO and fail to account for motion blur effects in dynamic scenes.

In this work, we propose AdaptiveAE, a reinforcement learning-based method that optimizes the selection of shutter speed and ISO combinations to maximize HDR reconstruction quality in dynamic environments. AdaptiveAE integrates an image synthesis pipeline that incorporates motion blur and noise simulation into our training procedure, leveraging semantic information and exposure histograms. It can adaptively select optimal ISO and shutter speed sequences based on a user-defined exposure time budget, and find a better exposure schedule than traditional solutions. Experimental results across multiple datasets demonstrate that it achieves the state-of-the-art performance.

1. Introduction

High-dynamic-range (HDR) imaging plays a pivotal role in computational photography. To capture an HDR scene,

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

†Corresponding authors.

due to hardware limits, a single capture can only cover a Low Dynamic Range (LDR), and HDR fusion techniques are proposed to combine multiple LDR images with varying exposures to cover a wide dynamic range [7, 10, 14, 15, 17, 19, 23, 35, 36, 40]. The typical way to vary exposure is to change either shutter speed or ISO. This is a challenging process, as a longer shutter speed may increase the signal-to-noise ratio but introduce unpleasant motion blur, a large ISO may increase brightness but also magnify noise, and significant exposure differences can cover a wider dynamic range but also increase the risk of misalignment. Therefore, the choice of exposure values (EVs) for each capture is critical in this process to ensure high-quality HDR results.

Still, very limited work discusses how to choose the optimal exposure levels, particularly in dynamic scenes. Prior works on exposure scheduling mainly focus on static scenes and ignores potential motion blur. Learning-based techniques overlook the intricate interplay between ISO and shutter speed, resulting in suboptimal image quality under varying conditions [6, 21, 32]. Additionally, many existing methods treat ghosting and motion blur as separate, computationally intensive post-processing tasks, which is unsuitable for real-time applications [11, 25, 27, 29, 31, 38, 39].

In this work, we propose *AdaptiveAE*, an efficient exposure control algorithm designed for HDR capturing, which addresses both motion blur and noise during image acquisition. Given the previously captured image, our method optimizes the exposure bracketing strategy for the subsequent capture, based on illumination information and semantic data from previous frames. Unlike previous approaches that treat motion blur as a separate post-processing task, we aim to address it during the capture process.

Designing both efficient and adaptive exposure control is non-trivial, and we resort to reinforcement learning [18] to solve this challenge. Our solution mimics an experienced photographer. At each iteration, the policy network takes the previously captured LDR images as input, together with extracted semantic and illumination information. Given this information, the policy network learns to determine the optimal exposure setup for the following captures, which maximizes the additional information provided by this frame while also reducing the risk of misalignment and motion blur in a dynamic scene. Once the next frame is captured, this newly captured image will be used as input for the subsequent refinement iteration. As shown in Fig. 1 right, the final quality (PSNR) of fusion increases as more images are captured.

Our proposed approach offers several advantages over traditional exposure controls. First, *AdaptiveAE* controls both exposure time and ISO and also adapts to different scenes. As a result, it has a much higher upper bound compared to either a fixed exposure schedule or an adaptive control algorithm that only changes exposure time. As shown in Fig. 1, our method iteratively achieves an optimal balance between noise and blur, resulting in less noise, reduced motion blur, and superior image quality compared to other baseline methods. Second, our method can handle both static and dynamic scenes. In static scenes, it achieves performance comparable to state-of-the-art techniques, and in dynamic ones, it produces visually compelling HDR images with minimal motion blur and ghosting artifacts. Third, our method can automatically choose the best number of frames for HDR imaging. Unlike traditional HDR approaches that often use a fixed number of frames, such as three, our approach provides flexibility to determine whether capturing three or more frames is optimal for certain scenes, balancing image quality and time budget.

Current datasets [3, 8, 9, 13, 26] are inadequate for studying auto-exposure (AE) with simultaneous noise and motion blur considerations. To bridge this gap, we introduce a blur-aware data synthesis pipeline. This novel approach enables the concurrent analysis of blur and noise in AE prediction, thereby enhancing HDR image quality. Our method uniquely integrates these factors, departing from traditional practices that address them separately.

We evaluate AdaptiveAE on established benchmarks, including the DeepHDR Video dataset [3] and the HDRV dataset [26], employing various downstream exposure fusion techniques. Our results demonstrate state-of-the-art performance compared to existing auto-exposure methods. Additionally, comprehensive ablation studies and targeted experiments focusing on motion blur confirm the efficacy of our approach and underscore the critical importance of incorporating blur synthesis into our pipeline. Our method, tested on real-world scenes using a SONY Alpha 7C-II, demonstrates superior noise control and effectively reduces motion blur, outperforming baselines in the visual quality of the fused HDR images.

Current datasets [3, 8, 9, 13, 26] are inadequate for studying auto-exposure (AE) with simultaneous noise and motion blur considerations. To bridge this gap, we introduce a blur-aware data synthesis pipeline. This novel approach allows for concurrent analysis of blur and noise in AE prediction, enhancing HDR image quality. Our method uniquely integrates these factors, departing from traditional practices that address them separately.

We evaluate AdaptiveAE on established benchmarks, including the DeepHDR Video dataset [3] and the HDRV dataset [26], employing various downstream exposure fusion techniques. Our results demonstrate state-of-the-art performance compared to existing auto-exposure methods. Also, comprehensive ablation studies and targeted experiments focusing on motion blur confirm the efficacy of our approach and highlight the critical importance of incorporating blur synthesis within our pipeline. Our method, tested on real-world scenes using a SONY Alpha 7C-II,

demonstrates superior noise control and effectively reduces motion blur, outperforming compared baselines in the visual quality of the fused HDR images.

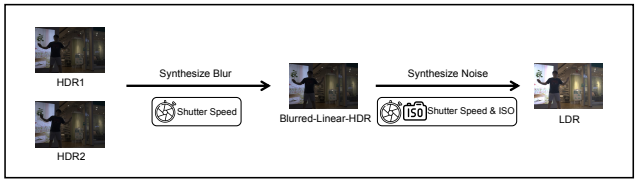
2. Related work

Strategy for exposure bracketing. Determining the optimal set of exposures for multiple-exposure dynamic range imaging is a well-established problem. Most digital cameras allow users to set the compensation ratio for exposure bracketing, while mobile cameras typically impose fixed ratios during automatic exposure bracketing. Heuristic strategies based on the histogram are proposed by [5, 21, 30], to balance single-to-noise ratio (SNR) and saturation. The work by [6] firstly formulates this challenge as a constrained optimization problem in the linear RGB domain, addressing the scenario that follows multiple exposure fusion and precedes tone mapping and denoising. This concise formulation facilitates the use of a straightforward numerical solver. Further extensions of this formulation consider the alignment of multiple input images to account for handshake [24]. In the context of structured light 3D reconstruction, various exposures are likewise treated as an optimization problem [4]. Exposure influences not only the noise but also, to some extent, the tone of the image in the standard RGB domain. Evaluating the final image quality complicates the problem further, as subsequent tone mapping or retouching can significantly alter image quality. Therefore, a neural network is utilized to estimate exposures and fuse multiple exposed images to achieve optimal fidelity in the gamma-corrected domain [12]. Additionally, reinforcement learning is employed by [37] and [32] to assess the rewards on comprehensive image quality after more sophisticated tone adjustment.

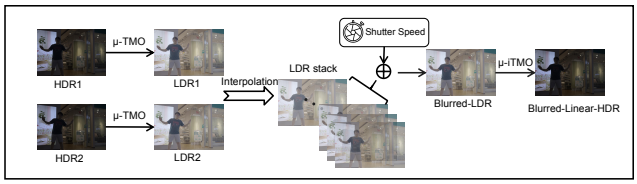
Challenge in dynamic scenes. In static scenes, increasing under-exposures can reduce saturation, and a longer shutter speed improves signal-to-noise ratio in dark areas. However, in dynamic scenes, excessive exposure may cause ghosting artifacts, and a prolonged shutter speed can lead to motion blur. Consequently, heuristic exposure bracketing is usually limited to two or three EV settings [5, 30]. The method by [6] imposes an upper limit on total shutter speed, while [24] addresses handshake motion through image registration but scarcely tackles local object motion. Most approaches do not fully address motion in dynamic scenes, leaving motion blur and ghosting artifacts to be addressed through post-processing.

3. Method

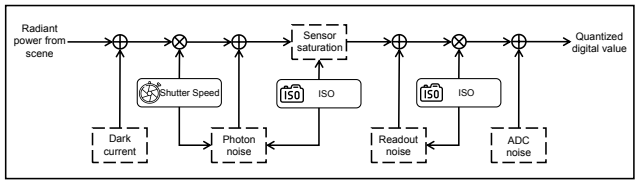
Conventional weighted linear combination methods, such as exposure fusion, provide straightforward SNR estimation but fail in dynamic scenes where moving objects create unquantifiable motion blur and ghosting artifacts due to misalignment. Our approach addresses these challenges during capturing rather than in post-processing, as the latter is shown to yield suboptimal results. We predict exposure-related risks—such as motion blur, ghosting, noise, and saturation—based on a limited number of previously captured frames. Additionally, we employ a sequential strategy for exposure and ISO parameter determination, rather than simultaneously predicting settings for all three LDR images, which reflects the iterative nature of auto-exposure in mobile cameras that enables adaptation to significant brightness transitions.



(a) Our overall image synthesis pipeline.



(b) Our blur synthesis pipeline.



(c) Our noise synthesis pipeline.

Figure 2. Our blur-aware data synthesis pipeline.

3.1. Blur-aware data synthesis pipeline

To simulate capturing in real environments, we designed an image synthesis pipeline to generate realistic motion blur and noise in LDR images from HDR videos in the training dataset, for use in training. Typical exposure settings involve adjusting the exposure value (EV), which is calculated as:

$$\mathrm{EV} = \log_2 \left(\frac{F^2}{T} \times \frac{100}{\mathrm{ISO}} \right), \tag{1}$$

where F denotes the aperture's f-number, ISO represents the ISO sensitivity and T is the exposure time in seconds.

Similar to recent methods [6, 21, 32] concerning high dynamic range capture, we assume that aperture and focus are held constant to prevent changes in defocus. This leaves just two camera settings to manipulate: (1) Shutter speed, which controls the amount of light to collect, and (2) ISO, which determines the sensor gain.

Our pipeline synthesizes motion blur and noise for the ground truth static HDR image based on a specified ISO

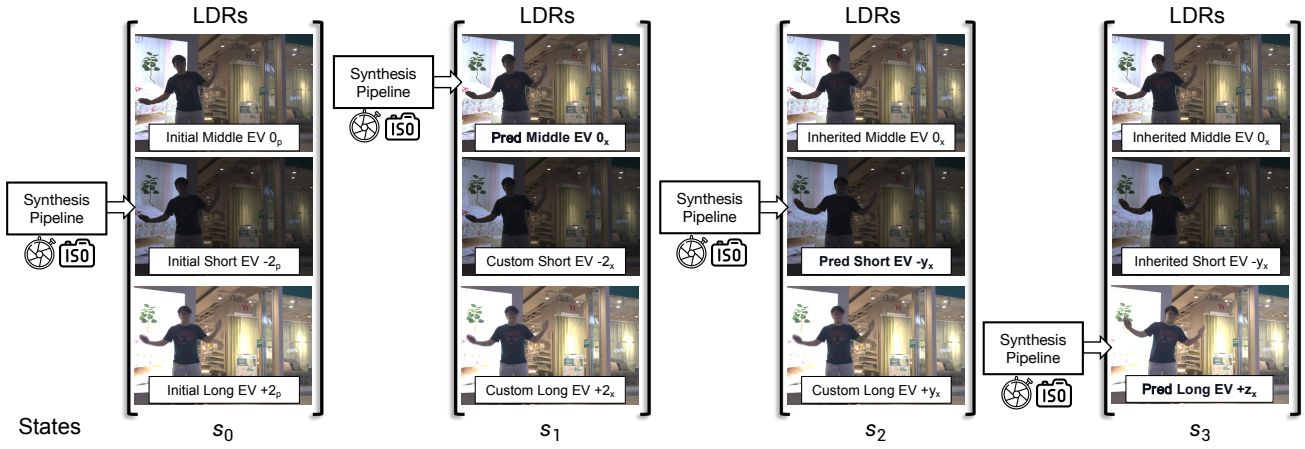


Figure 3. The training scheme of AdaptiveAE. States are defined as the three LDRs synthesized using predicted ISOs and shutter speeds. Starting from s_0 where the three LDRs has EV $\{-2, 0, +2\}$ with arbitrary EV 0 baseline, ISOs and shutter speeds, the agent sequentially predicts, customizes or inherits capturing parameters (*i.e.* ISO and shutter speed) for the next stage and synthesize the corresponding LDR using our image synthesis pipeline. Unlike training, the LDRs will be captured rather than synthesized during inference.

and shutter speed. As depicted in Fig. 2, it uses two consecutive static HDR images as input. Motion blur is first synthesized according to shutter speed to produce a blurred HDR image in linear space. Next, noise is added based on shutter speed and ISO to create an LDR image, reflecting our exposure choice. Note that motion blur should be applied before adding noise, as it influences the raw input by affecting the number and pattern of captured photons during photography.

Synthesizing blur. On-the-shelf training datasets consist of consecutive HDR ground truth of a scene with motion. Now we explain how we simulate motion blur to a frame of HDR f_i^L, as shown in Fig. 2b, which is the i-th HDR ground truth frame in the dataset scene, and the superscript L indicates it is in linear space. We first use μ-law tone-mapping with $\mu = 5000$ to transfer f_i^L and f_{i+1}^L from HDR space to LDR space where the image interpolation algorithm we applied is trained upon, receiving $f_i^{\mathcal{T}}$ and $f_{i+1}^{\mathcal{T}}$, where the superscript $\mathcal{T}$ denotes they are in LDR space. Then we use RIFE [7] to interpolate the them to 256 frames and get the sequence of images $\{f_i^{\mathcal{T}}, s_1^{\mathcal{T}}, s_2^{\mathcal{T}}, \cdots, s_{254}^{\mathcal{T}}, f_{i+1}^{\mathcal{T}}\}$. Then for the selected shutter speed T_j for the j-th LDR $l_j^{\mathcal{T}}$ to take, the blurred HDR b_j^L is simulated as:

$$b_j^L = \mathbf{iTMO}\left(\frac{f_i^{\mathcal{T}} + \sum_{m=1}^{m_j} s_m^{\mathcal{T}}}{m_j}\right), m_j = \left\lceil \frac{256 T_j}{\Delta\tau} \right\rceil, \quad (2)$$

where $\Delta\tau$ denotes the time elapsed from f_i^L is taken to f_{i+1}^L is not yet taken and **iTMO** indicates the inverse μ-law tonemapping function with $\mu = 5000$.

Synthesizing noise. We adopt the noise model mentioned in [6], in which noise is modeled 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:

$$Var(n) = \frac{\Phi T \times \mathrm{ISO}^2}{U^2} + \frac{\sigma_{\mathrm{read}}^2 \times \mathrm{ISO}^2}{U^2} + \sigma_{\mathrm{ADC}}^2, \quad (3)$$

where Φ is the radiance level, T is the shutter speed, U is a camera-dependent variable.

Following our noise model, as shown in Fig. 2c, we can synthesize the corresponding noise with selected ISO and shutter speed to the blurred HDR b_j^L to get the LDR image $l_j^{\mathcal{T}}$. For details, please refer to our supplementary material.

3.2. Problem formulation of AdaptiveAE

Given a scene, the goal of AdaptiveAE, which has access to three initial preview LDR images $\{p_j^{\mathcal{T}}\}_{1,2,3}$ (*i.e.* underexposed, mid-exposed, overexposed) before capturing, is to find an optimal exposure setup (*i.e.* ISO and shutter speed) for LDR capturing, with which the fused HDR output will result in pleasing visual performance.

Our method formulates exposure bracketing as a Markov Decision Process [22], solved via deep reinforcement learning to refine exposure parameters (ISO, shutter speed) sequentially. As illustrated in Fig. 4, the process starts from three LDRs at a default $\{-2_p, 0_p, +2_p\}$ EV spacing relative to an arbitrary reference (subscript p). The refinement then proceeds in stages, as shown in Fig. 3.

First, the agent predicts optimal parameters for the mid-exposed frame, establishing a new 0-EV reference (subscript x). The side frames are then **customized**—their parameters are procedurally set to achieve a symmetric $\{-2_x, 0_x, +2_x\}$ EV bracket. Next, the agent refines the underexposed frame to an EV of $-y_x$. The mid-exposed frame (0_x) is **inherited** (its parameters are reused), while

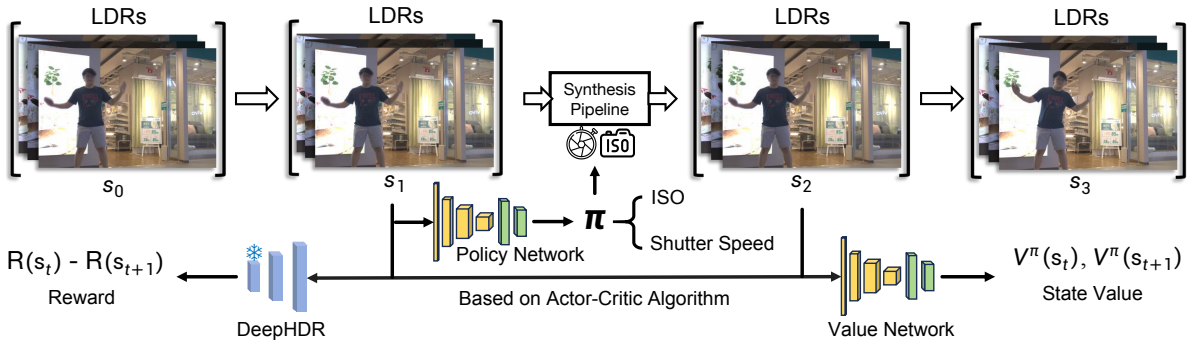


Figure 4. Training pipeline of our method. The ISO and shutter speed prediction process is conceptualized as a Markov Decision Process, where a CNN-based policy network predicts the ISO and shutter speed of the next exposure sets. Concurrently, a CNN-based value network estimates the state value. We leverage our blur-aware image synthesis pipeline to synthesize the predicted LDRs and employ DeepHDR [7] to fuse the predicted LDR images, generating our HDR result and calculating the reward for the current policy. The entire system is optimized using the A3C (Asynchronous Advantage Actor-Critic) method [18].

the overexposed frame is customized to $+y_x$ to maintain symmetry, yielding $\{-y_x, 0_x, +y_x\}$. Finally, the agent predicts the overexposed frame's EV as $+z_x$, creating a potentially asymmetric set $\{-y_x, 0_x, +z_x\}$. This sequential prediction can be extended, as our adopted fusion method [7] handles more than three LDRs. A video providing further details on this process, along with an example of extended exposure bracketing, is available in the supplementary materials.

3.3. Optimization objectives

Let us denote the problem as $P = (S, A)$, where S is a state space and A is an action space. Specifically, in our task, S is the space of the exposure setups (*i.e.*, ISO and shutter speed) for an LDR set that typically contains three LDRs (*i.e.* under-exposed, mid-exposed, and over-exposed), while A is the set of all possible ISO and shutter speed combinations, which is discrete. During training, at stage $s_j = \{(\text{ISO}_{j1}, T_{j1}), (\text{ISO}_{j2}, T_{j2}), (\text{ISO}_{j3}, T_{j3}), \}$, we first find the corresponding HDR ground truth image pairs $(f_i^L, f_{i+1}^L)_{1,2,3}$ and generate the corresponding LDRs $\{l_j^{\mathcal{T}}\}_{1,2,3}$ through our image synthesis pipeline in Fig. 2. The footnote $1, 2, 3$ denotes that the same operation is done for the under-, mid-, and over-exposed LDRs. Taking $\{l_j^{\mathcal{T}}\}_{1,2,3}$ as input, the agent predicts an action $a_j = (\text{ISO}_j, T_j)$, which is expanded to an exposure setup for three LDRs by customizing EV or inheriting from state s_{j-1}, mapping state s_j to state s_{j+1}. Adding a sequence of M LDRs to exposure bracketing corresponds to a trajectory τ of states and actions:

$$\tau = (s_0, a_0, \cdots, s_{M-1}, a_{M-1}, s_M), \tag{4}$$

where s_M is the stopping state. Our goal is to find a policy that maximizes the accumulated reward during the decision-making process. In this paper, the reward function with the j-th action (i.e., corresponding to deciding the exposure setting for the j-th LDR for exposure fusion) is thus written as:

$$r(s_j, a_j) = \mathcal{R}(s_{j+1}) - \mathcal{R}(s_j) - \mathcal{P}(j), \tag{5}$$

where $s_{j+1} = p(s_j, a_j)$, and $\mathcal{R}$ denotes our reward design and $\mathcal{P}$ denotes the L_{step} penalty, detailed in 3.4.

As depicted in Fig. 4, our model comprises a policy network and a value network, both of which utilize a CNN-based architecture. The policy network predicts the optimal ISO and shutter speed for the subsequent exposure, outputting a distribution of action probabilities $\pi(s, \theta)$ for an input image s. Concurrently, the value network $V^\pi(s, \omega)$ estimates the corresponding state value. These networks, with combined parameters $\psi = (\theta, \omega)$, are trained by maximizing our objective $J(\theta)_\psi$ to learn the optimal policy $\pi(s)$. Specifically, to train the policy network and the value network, we apply the A3C (Asynchronous Advantage Actor-Critic) method [18], where the actor is represented by the policy network and the critic is the value network. Network details are in the supplementary materials.

3.4. Reward

When designing the reward function for our system, we considered four key factors: (1) similarity between the fused HDR and the ground truth HDR; (2) quality of important regions in the fused HDR; (3) quality of moving regions in the fused HDR; and (4) a penalty for overly long LDR stacks. Thus, our reward function is:

$$\mathcal{R} = -(P_{\text{construction}} + P_{\text{priority}} + P_{\text{ghost}}). \tag{6}$$

We consider $P_{\text{construction}}$, the L2 loss between our fused HDR and the ground truth, as our major reward component, which is affected by noise and saturation. Note that through the entire sequential decision process, the middle-exposed frame is used as a reference for HDR fusion. Conforming to Eq. (3), during training, noise is synthesized according to the irradiance of ground truth HDR, and during inference, noise is estimated with the irradiance of noisy signals.

P_{priority} represents an L2 loss within the areas in the image masked by an importance mask, which is generated by a saliency predictor [20]. This ensures that the highest quality is maintained in the most significant areas, thereby enhancing the overall visual fidelity where it matters most.

P_{ghost} is also an L2 loss within a masked area, denoting areas with large motions and thus having a higher risk of motion blur- or ghosting-caused HDR quality degradation. The mask is computed by calculating the optical flow using RAFT [28] between the HDR ground truth of the middle-exposed frame (i.e., the reference frame for fusion) and the corresponding HDR f_i^L and selecting the pixels where the mode of the flow vector exceeds a constant threshold K. Normalizing the largest optical flow vector, we empirically set K to 0.2. P_{ghost} guides the agent to deal more carefully with regions that are prone to artifacts caused by motion and is helpful for high-quality HDR capturing, as is verified by the result of our ablation studies in Tab. 2.

$\mathcal{P}(j)$ is a penalty designed to penalize excessively long exposure brackets. Capturing an excessive number of shots, such as 10, even with random exposure settings, can lead to nearly perfect HDR fusion results, but it is time-consuming. Typically, three shots [2, 6, 21] are sufficient to achieve high-quality outcomes. To this end, we incorporate a penalty for taking more than three shots, as follows:

$$\mathcal{P}(j) = \begin{cases} 0 & \text{if } j \leq H \\ \alpha(j-H)^2 & \text{if } j > H \end{cases}, \tag{7}$$

where α is a positive coefficient and H is set to 3.

In this manner, the autonomous agent optimizes exposure parameters by predicting relatively fast shutter speeds for LDR images, particularly for the middle-exposed reference frame, thereby minimizing motion blur while avoiding excessive ISO values that would introduce noise-related degradation. When confronted with potential ghosting artifacts—which emerge from information deficiency in LDR images due to concurrent saturation and motion—the agent adaptively selects EVs that minimize both underexposure and saturation, resulting in a significant reduction of ghosting artifacts in the final reconstruction.

4. Experiments

Experiment details. We use Real-HDRV [26] for training and tested our performance on Real-HDRV [26] and DeepHDRVideo [3]. For HDR fusion, we adopt DeepHDR [33] to generate the HDR image based on the selected exposure bracketing. Additionally, we set the number of LDR frames to 3 for all methods involved. For methods that do not account for changes in ISO values, we set the ISO to 200 as a standard value for most cases, which is rationalized in our supplementary materials. Since RIFE [7] is relatively time-consuming, the image interpolation and blur synthesis step is performed before training. We apply random flipping, rotation, and cropping with 512 $\times$ 512 pixels for data augmentation. Our training dataset consists of a total of 770 scenes, including 440 dynamic and 330 static scenes.

Evaluation metrics. We directly evaluate the performance of the fused HDR results. Similar to previous HDR fusion methods [7, 14, 19], we employ PSNR-μ, SSIM-μ, PU-PSNR, PU-SSIM, and HDR-VDP-2 [16] as evaluation metrics. PSNR-μ and SSIM-μ denote PSNR and SSIM of the fused HDR after μ-law tone-mapping with μ=5000. PU-PSNR and PU-SSIM are computed after perceptually uniform encoding [1]. When computing the HDR-VDP-2 [16], the diagonal display size is 30 inches.

4.1. Results

Results on Real-HDRV dataset. We compared our trained agent's performance with several state-of-the-art HDR exposure bracketing methods, including Pourreza-Shahri *et al*. [21], Hasinoff *et al*. [6], and Wang *et al*. [32]. The first two are non-deep-learning methods that do not consider motion: Pourreza-Shahri *et al*. use K-means clustering to adjust shutter speed based on image brightness, while Hasinoff *et al*. mathematically optimize ISO and shutter speed for the best worst-case SNR. Wang *et al*. utilize reinforcement learning to predict shutter speed for maximizing PSNR, but also ignore motion.

Our method achieves state-of-the-art performance on the HDRV-Test dataset, as shown in Tab. 1. By treating ISO as a variable, it provides flexibility in handling extremely dark scenes. Additionally, it excels in dynamic scenes due to a blur synthesis model and carefully designed rewards. Thus, our model effectively balances noise reduction and motion artifact minimization, delivering high-quality HDR results. Visualization results in Fig. 5 show that deep learning-based exposure fusion models affect EV selection differently. In Hasinoff *et al*.'s setup, three LDRs are equally weighted, but dynamic scenes require a reference image, emphasizing the middle-exposed LDR's quality. This often shifts its EV toward under-exposure to reduce motion blur. Our experiments confirm that if the middle-exposed image is blurry—a common problem in other methods—the fused HDR will also be blurred.

Inference time. Our method takes less than 250 ms for each HDR image, which is acceptable. With average exposure time n ($\leq 30ms$) and prediction time m ($\leq 10ms$), the total execution time is $6n + 3m$ ($\leq 250ms$). This can be further lowered to $6n$ ($\leq 200ms$) adopting concurrency. Note that DeepHDR is only required during training for calculating rewards, and on-camera inference can be performed with a small agent, taking only $3.5ms$ for each frame without optimization. Besides, our method can be further accelerated by on-camera processors when applied to DSLRs. More discussions are in our supplementary materials.

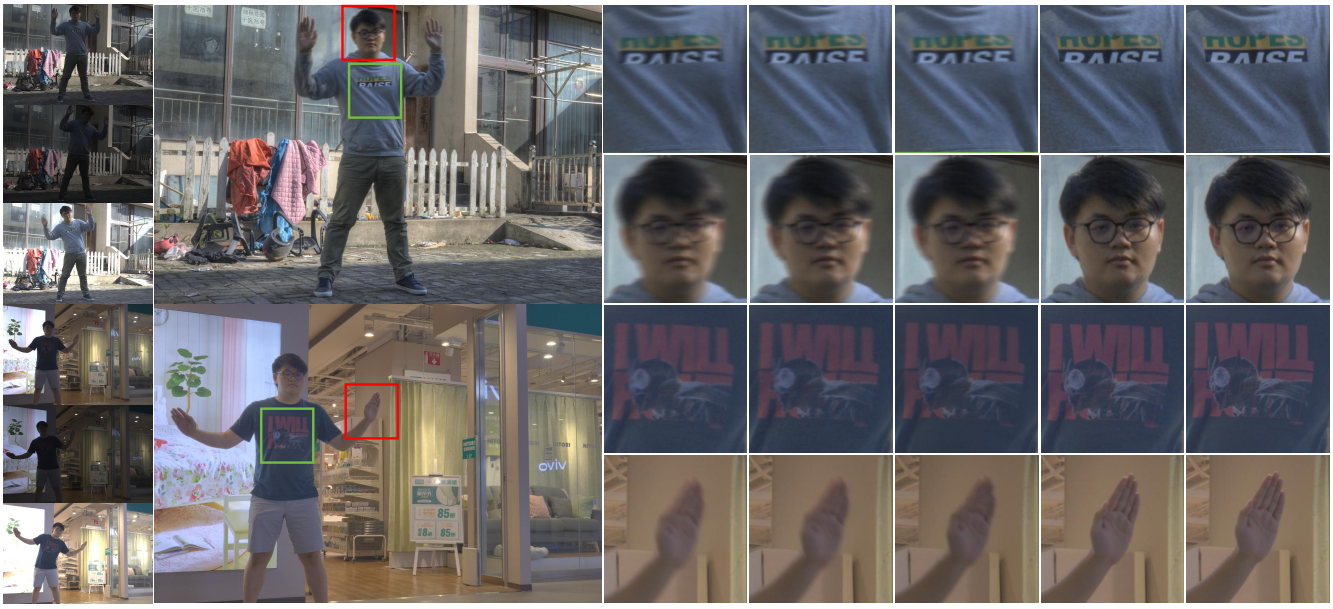


Figure 5. Qualitative comparisons with other auto-exposure methods on HDRV dataset [26]. Left: Predicted LDRs with varying ISO and shutter speed settings and synthesized using our image synthesis pipeline. Middle: Fused HDR image using DeepHDR [7] and tone-mapped using Photomatix. Right: Zoom-in results for tested methods.

Table 1. Comparison of Different Methods. We utilized one preview image and three preview images, respectively, for the compared methods and our method to predict exposure settings for three LDRs. We leveraged DeepHDR [7] for exposure fusion and used the mentioned metrics to evaluate the quality of the fused HDR. **Bold**: The best.

Methods	HDRV [26]					DeepHDRVideo [3]				
	PSNR-μ	SSIM-μ	HDR-VDP-2	PU-PSNR	PU-SSIM	PSNR-μ	SSIM-μ	HDR-VDP-2	PU-PSNR	PU-SSIM
Pourreza *et al*. [21]	33.64	0.8617	54.55	30.61	0.8679	35.57	0.8780	55.67	31.59	0.8791
Hasinoff *et al*. [6]	37.59	0.9052	57.02	32.87	0.8980	38.47	0.9157	58.65	34.45	0.9132
Wang *et al*. [32]	36.46	0.8902	56.09	32.68	0.8933	37.95	0.9019	57.39	33.27	0.9008
Ours	**39.70**	**0.9408**	**59.20**	**34.67**	**0.9465**	**39.81**	**0.9371**	**58.90**	**36.19**	**0.9338**

Table 2. Ablation study of AdaptiveAE on HDRV [26] dataset. Base denotes our model trained with only the step penalty and construction reward. **Bold**: The best.

Model	PSNR-μ	SSIM-μ	PU-PSNR	PU-SSIM
Base	38.21	0.9227	32.68	0.9198
Base+P_{priority}	38.57	0.9261	33.02	0.9239
Base+P_{priority}+P_{ghost}	**39.70**	**0.9408**	**34.67**	**0.9465**

Gap to the best-achievable. For each scene in the test set of Real-HDRV, we iteratively search for the best set of predictions by Gaussian sampling around our initial prediction (50 times per exposure parameter per frame, with a deviation of 20% of the mean for both ISO and shutter speed). Statistics in Tab. 4 show that our method approaches the locally optimal result while being efficient.

Cross datasets test. To test the generalization ability of our model, we also evaluated the performance of our trained agent on DeepHDRVideo [3], as shown in the right of Tab. 1. Our agent exhibits good generalization abilities.

Table 3. Comparison of different exposure fusion methods. We utilize the four auto-exposure methods to predict exposure settings and employ our image synthesis pipeline to create three LDRs. Then, we apply different exposure fusion methods to fuse them and compare the HDR quality on the HDRV dataset [26]. HDR-Transformer [14] is pretrained on HDRV [26] dataset. P: PSNR-μ, S: SSIM-μ. Pou: Pourreza *et al*. [21], Has: Hasinoff *et al*. [6], W: Wang *et al*. [32]. HDR-Trans: HDR-Transformer. **Bold**: The best.

Model	DeepHDR		HDR-GAN		HDR-Trans	
	P-μ	S-μ	P-μ	S-μ	P-μ	S-μ
Pou	33.64	0.8617	35.71	0.8892	35.84	0.8824
Has	37.59	0.9252	38.58	0.9263	39.11	0.9372
W	36.46	0.9002	37.95	0.9169	38.89	0.9210
Ours	**39.70**	**0.9408**	**40.73**	**0.9376**	**41.37**	**0.9478**

Since the DeepHDRVideo dataset only provides ground truth HDR for the middle image in the sequence, we synthesize HDR for the other images using DeepHDR [33].

Cross HDR fusion methods test. In our training scheme,

Table 4. Performance comparison investigating the gap between our method and the optimum on the HDRV-test dataset.

	Ours	Worst	Average	Best
PSNR-μ	39.70	25.76	32.41	39.93
SSIM-μ	0.9408	0.7738	0.8609	0.9412

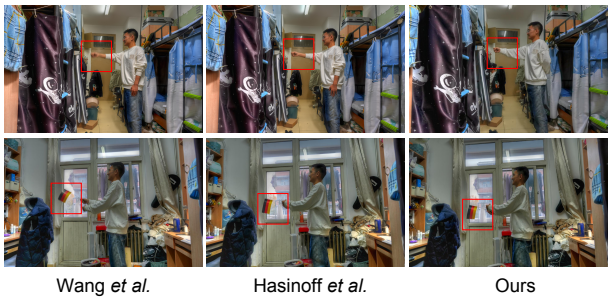


Figure 6. Results of our real capture data. The subject performs steady and repetitive movements, taking shots according to the exposure settings predicted by various methods.

we used DeepHDR for exposure fusion. For further testing, we adopted different exposure methods, including HDR-GAN [19] and HDR-Transformer [14] (pre-trained on the HDRV-dataset [26]), as post-processing exposure fusion methods. All tests are conducted on the HDRV-Test dataset. As shown in Tab. 3, without considering motion before LDR capturing, however powerful an exposure fusion method (*i.e.* HDR-GAN and HDR-Transformer) fails to yield a satisfactory result. Notably, when stronger fusion models are used, the performance gap between our model and traditional models, which do not account for motion, increases. This is because, when blur and ghosting risk are mitigated, post-processing models can effectively manage the remaining challenges. In contrast, ignoring motion makes these challenges more difficult for post-processing models, leading to potential failure.

Results for real capture. We used a SONY Alpha 7C-II to evaluate our model on real-world scenes. Subjects performed steady, repetitive movements while ISO and shutter speed were manually set for each capture, with the aperture fixed at f/2.8. The camera and subject positions remained unchanged during multi-exposure captures. As shown in Fig. 6, our method provides noise control comparable to other baselines while effectively mitigating motion blur, which can impair the visual quality of the fused HDR image; for more results, see our supplementary materials.

4.2. Ablation studies

Effectiveness of penalty item. We conduct ablation studies to validate the effectiveness of our reward design. All quantitative evaluations are conducted on the HDRV-Test dataset. We train our networks using only the construction

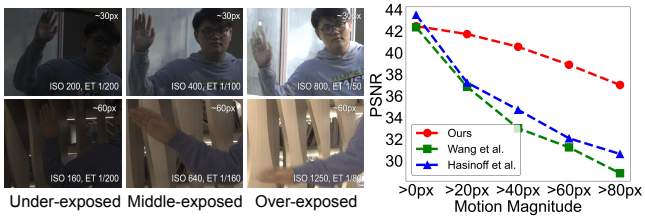


Figure 7. **Left:** Zoom-in details of the moving regions and our predicted ISO and shutter speed (seconds) for each LDR. Up: LDRs for scenes with an average motion level of around 30 pixels, Down: around 60 pixels. **Right:** Comparisons with baseline methods across different motion magnitude ranges. Tested on Real-HDRV [26] dataset.

reward and step penalty as the base model and test the effectiveness of P_{ghost} and P_{priority}. As shown in Tab. 2, these two reward elements enhance our agent's performance by directing the model to focus on semantically important regions and moving areas. The former typically corresponds to the scene's focal point (*e.g.* a person's face), while the latter involves regions at high risk of motion blur and ghosting artifacts. Qualitative evidence is provided in our supplementary materials.

Robustness to dynamic scenes. We offer a pair of examples from the Real-HDRV [26] dataset with around 30 and 60 pixels of regional movement, respectively. As shown in Fig. 7: **Left**, our trained agent tends to predict faster shutter speeds for scenes with stronger movement, validating the responsiveness of our agent in dynamic scenes. We assess our method's performance against previous auto-exposure methods [6, 32] at varying motion magnitudes (Fig. 7: **Right**). To create the evaluation dataset for robustness to different motion levels, we follow HDRFlow [34], using RAFT [28] to process dynamic scenes from HDRV and obtain optical flow maps. We then manually crop these images with reasonable flow predictions, dividing them into 128 $\times$ 128 blocks, and calculate the average motion magnitude for each block. Finally, we evaluate the PSNR of blocks corresponding to different motion magnitudes. As shown in Fig. 7: **Right**, our AdaptiveAE demonstrates greater robustness than other methods as motion magnitude increases.

5. Conclusion

We introduce AdaptiveAE, which optimizes HDR exposure in dynamic settings using deep reinforcement learning, treating exposure bracketing as a Markov Decision Process. It autonomously adjusts ISO and shutter speed for a pre-trained exposure fusion algorithm. Reward systems focus on moving and key regions, minimizing sequence lengths. Experiments show AdaptiveAE outperforms state-of-the-art methods in dynamic scenes while matching top models in static ones, allowing flexible HDR capture. Our analysis of noise models and exposure settings offers insights for future research, with plans to include adjustable apertures.

Acknowledgement. This work was supported by the National Key R&D Program of China No. 2022ZD0160201, Shanghai Artificial Intelligence Laboratory, National Natural Science Foundation of China (Grant No. 62136001, 62088102), Beijing Natural Science Foundation (Grant No. L233024), and Beijing Municipal Science & Technology Commission, Administrative Commission of Zhongguancun Science Park (Grant No. Z241100003524012). PKU-affiliated authors thank openbayes.com for providing computing resources.

References

[1] Maryam Azimi, Vedad Hulusic, Philippe Hanhart, and Touradj Ebrahimi. PU21: A novel perceptually uniform encoding for adapting existing quality metrics for HDR. In *Proceedings of the Picture Coding Symposium (PCS)*, pages 1–5, 2021. 6

[2] Neil Barakat, A. Nicholas Hone, and Thomas E. Darcie. Minimal-bracketing sets for high-dynamic-range image capture. *IEEE Transactions on Image Processing*, 17(10):1864–1875, 2008. 6, 1

[3] Guanying Chen, Chaofeng Chen, Shi Guo, Zhetong Liang, Kwan-Yee K. Wong, and Lei Zhang. HDR video reconstruction: A coarse-to-fine network and a real-world benchmark dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2502–2511, 2021. 2, 6, 7, 1

[4] Wenyuan Chen, Xingjian Liu, Changhai Ru, and Yu Sun. Automated exposures selection for high dynamic range structured-light 3-D scanning. *IEEE Transactions on Industrial Electronics*, 70(7):7428–7437, 2023. 3

[5] Natasha Gelfand, Andrew Adams, Sung Hee Park, and Kari Pulli. Multi-exposure imaging on mobile devices. In *Proceedings of the 18th ACM International Conference on Multimedia*, pages 823–826, 2010. 3

[6] Samuel W Hasinoff, Frédo Durand, and William T Freeman. Noise-optimal capture for high dynamic range photography. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 553–560, 2010. 1, 2, 3, 4, 6, 7, 8

[7] Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. Real-time intermediate flow estimation for video frame interpolation. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2022. 2, 4, 5, 6, 7, 1, 3

[8] Nima Khademi Kalantari and Ravi Ramamoorthi. Deep high dynamic range imaging of dynamic scenes. *ACM Transactions on Graphics (TOG)*, 36(4):144:1–144:12, 2017. 2

[9] Nima Khademi Kalantari, Eli Shechtman, Connelly Barnes, Soheil Darabi, Dan B. Goldman, and Pradeep Sen. Patch-based high dynamic range video. *ACM Transactions on Graphics (TOG)*, 32(6):202:1–202:10, 2013. 2

[10] Lingtong Kong, Bo Li, Yike Xiong, Hao Zhang, Hong Gu, and Jinwei Chen. SAFNet: Selective alignment fusion network for efficient HDR imaging. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2024. 2

[11] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. DeblurGAN-v2: Deblurring (orders-of-magnitude) faster and better. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8878–8887, 2019. 2

[12] Jieyu Li, Ruiwen Zhen, and Robert L. Stevenson. A lightweight exposure bracketing strategy for HDR imaging without access to camera RAW. *Electronic Imaging*, 2023 (15):116–1–116–6, 2023. 3

[13] Shuaizheng Liu, Xindong Zhang, Lingchen Sun, Zhetong Liang, Hui Zeng, and Lei Zhang. Joint HDR denoising and fusion: A real-world mobile HDR image dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13966–13975, 2023. 2

[14] Zhen Liu, Yinglong Wang, Bing Zeng, and Shuaicheng Liu. Ghost-free high dynamic range imaging with context-aware transformer. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2022. 2, 6, 7, 8

[15] Kede Ma, Zhengfang Duanmu, Hanwei Zhu, Yuming Fang, and Zhou Wang. Deep guided learning for fast multi-exposure image fusion. *IEEE Transactions on Image Processing*, 29:2808–2819, 2020. 2

[16] Rafał Mantiuk, Kil Joong Kim, Allan G. Rempel, and Wolfgang Heidrich. HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Transactions on Graphics (TOG)*, 30(4):1–14, 2011. 6

[17] Tom Mertens, Jan Kautz, and Frank Van Reeth. Exposure fusion: A simple and practical alternative to high dynamic range photography. *Computer Graphics Forum*, 28(1):161–171, 2009. 2

[18] Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. arXiv preprint:1602.01783, 2016. 2, 5

[19] Yuzhen Niu, Jianbin Wu, Wenxi Liu, Wenzhong Guo, and Rynson W. H. Lau. HDR-GAN: HDR image reconstruction from multi-exposed LDR images with large motions. *IEEE Transactions on Image Processing*, 30:3885–3896, 2021. 2, 6, 8

[20] Junting Pan, Cristian Canton Ferrer, Kevin McGuinness, Noel E. O'Connor, Jordi Torres, Elisa Sayrol, and Xavier Giró-i Nieto. SalGAN: Visual saliency prediction with generative adversarial networks. arXiv preprint:1701.01081, 2017. 6

[21] Reza Pourreza-Shahri and Nasser Kehtarnavaz. Exposure bracketing via automatic exposure selection. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 320–323, 2015. 1, 2, 3, 6, 7

[22] Martin L. Puterman. Markov decision processes. In *Handbooks in Operations Research and Management Science*, chapter 8, pages 331–434. North-Holland, Amsterdam, 1990. 4

[23] Linhao Qu, Shaolei Liu, Manning Wang, and Zhijian 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*, pages 2126–2134, 2022. 2

[24] Kalpana Seshadrinathan, Sung Hee Park, and Oscar Nestares. Noise and dynamic range optimal computational imaging. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 2785–2788, 2012. 3

[25] Ziyi Shen, Wenguan Wang, Xiankai Lu, Jianbing Shen, Haibin Ling, Tingfa Xu, and Ling Shao. Human-aware motion deblurring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5572–5581, 2019. 2

[26] Yong Shu, Liquan Shen, Xiangyu Hu, Mengyao Li, and Zihao Zhou. Towards real-world HDR video reconstruction: A large-scale benchmark dataset and a two-stage alignment network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2879–2888, 2024. 1, 2, 6, 7, 8, 3

[27] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8174–8182, 2018. 2

[28] Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2020. 6, 8

[29] Fu-Jen Tsai, Yan-Tsung Peng, Chung-Chi Tsai, Yen-Yu Lin, and Chia-Wen Lin. BANet: A blur-aware attention network for dynamic scene deblurring. *IEEE Transactions on Image Processing*, 31:6789–6799, 2022. 2, 1, 3

[30] Peter van Beek. Improved image selection for stack-based HDR imaging. *Electronic Imaging*, 2019(4):581–1–581–6, 2019. 3

[31] Channarayapatna Shivaram Vijay, Chandramouli Paramanand, Ambasamudram Narayanan Rajagopalan, and Rama Chellappa. Non-uniform deblurring in HDR image reconstruction. *IEEE Transactions on Image Processing*, 22 (10):3739–3750, 2013. 2

[32] Zhouxia Wang, Jiawei Zhang, Mude Lin, Jiong Wang, Ping Luo, and Jimmy S. Ren. Learning a reinforced agent for flexible exposure bracketing selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1820–1828, 2020. 1, 2, 3, 6, 7, 8

[33] Shangzhe Wu, Jiarui Xu, Yu-Wing Tai, and Chi-Keung Tang. Deep high dynamic range imaging with large foreground motions. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2018. 6, 7

[34] Gangwei Xu, Yujin Wang, Jinwei Gu, Tianfan Xue, and Xin Yang. HDRFlow: Real-time HDR video reconstruction with large motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24851–24860, 2024. 8

[35] Qingsen Yan, Dong Gong, Qinfeng Shi, Anton van den Hengel, Chunhua Shen, Ian Reid, and Yanning Zhang. Attention-guided network for ghost-free high dynamic range imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1751–1760, 2019. 2

[36] Qingsen Yan, Lei Zhang, Yu Liu, Yu Zhu, Jinqiu Sun, Qinfeng Shi, and Yanning Zhang. Deep HDR imaging via a non-local network. *IEEE Transactions on Image Processing*, 29: 4308–4322, 2020. 2

[37] Runsheng Yu, Wenyu Liu, Yasen Zhang, Zhi Qu, Deli Zhao, and Bo Zhang. DeepExposure: Learning to expose photos with asynchronously reinforced adversarial learning. In *Advances in Neural Information Processing Systems 31*, pages 8377–8386, 2018. 3

[38] Lu Yuan, Jian Sun, Long Quan, and Heung-Yeung Shum. Image deblurring with blurred/noisy image pairs. *ACM Transactions on Graphics (TOG)*, 26(3):1:1–1:10, 2007. 2

[39] Kaihao Zhang, Wenhan Luo, Yiran Zhong, Lin Ma, Björn Stenger, Wei Liu, and Hongdong Li. Deblurring by realistic blurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2737–2746, 2020. 2

[40] Zhilu Zhang, Haoyu Wang, Shuai Liu, Xiaotao Wang, Lei Lei, and Wangmeng Zuo. Self-supervised high dynamic range imaging with multi-exposure images in dynamic scenes. arXiv preprint:2310.01840, 2023. 2

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.

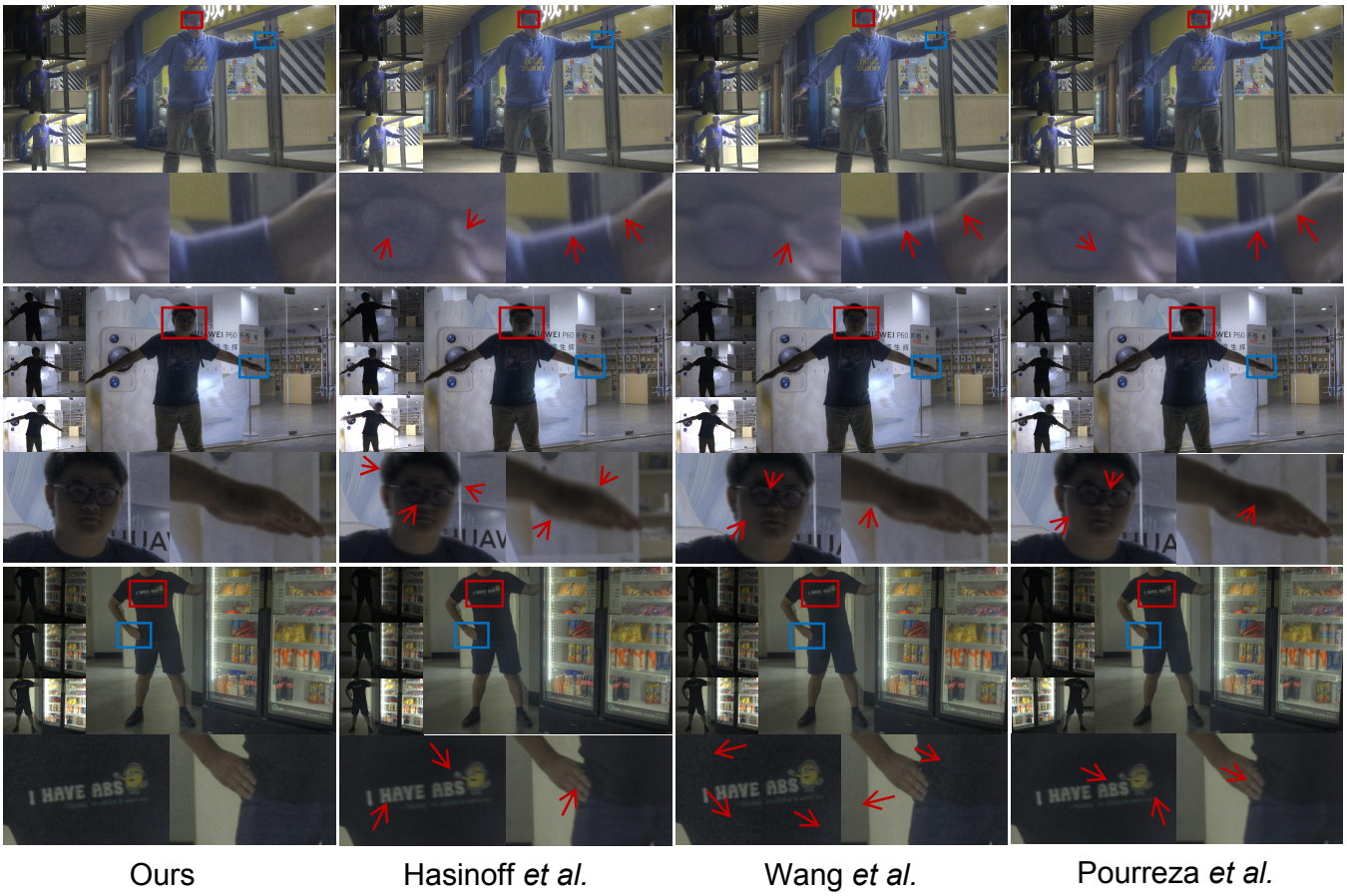


Figure A1. Additional qualitative comparisons with other auto-exposure methods on HDRV dataset [26]. Upper left: Predicted LDRs with varying ISO and shutter speed settings and synthesized using our image synthesis pipeline. Upper right: Fused HDR image using DeepHDR [7] and tone-mapped using Photomatix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

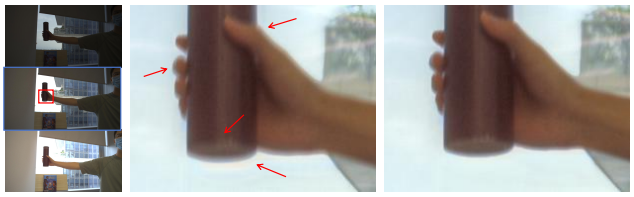


Figure A2. Effectiveness of ghost penalty.

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.

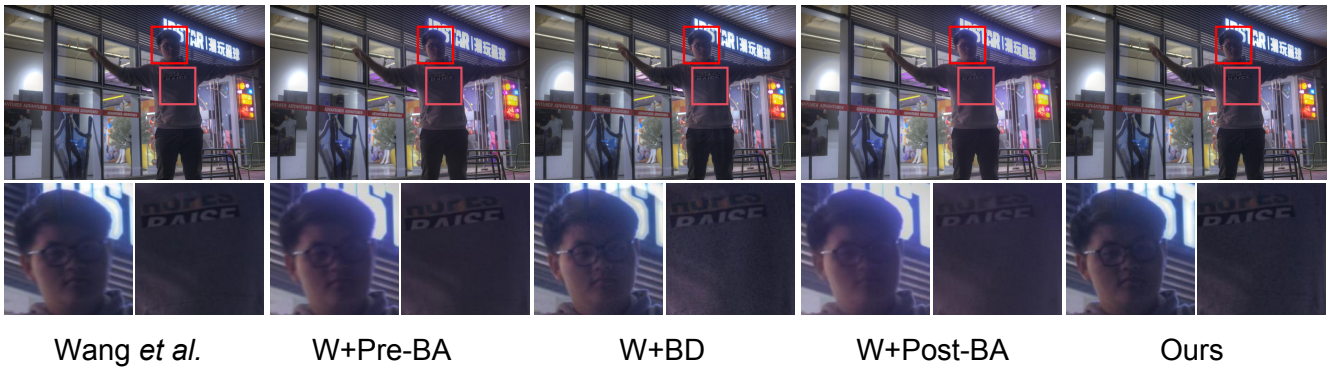


Figure A3. Necessity of considering motion blur during LDR capturing. We compared our method on the HDRV dataset [26] with Wang *et al*. [32] combined with different post-processing deblurring methods. W denotes Wang *et al*., Pre-BA denotes applying BANet [29] to LDRs for fusion. BD denotes DeepHDR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