World-Model-Inspired Flicker State Modeling for Burst Flicker Removal

Bowen Tang **Tao Wang** **Xin Yu** **Wenhan Luo** **Kaihao Zhang**[✉] **Bo Li** **Min Zhang**

[1]HITSZ [2]vivo BlueImage Lab [3]The University of Queensland [4]HKUST [5]ANU

Abstract

Burst imaging under unstable illumination suffers from flicker, a periodic degradation that modulates local brightness differently across adjacent frames, creating frame-specific flicker states. Existing methods rely on spatial priors or reference-centric fusion without explicitly modeling the evolving flicker state, whose temporal periodicity makes it predictable from adjacent observations by world models that capture latent state transitions. Motivated by this predictive capacity, we propose StateFlicker, a state-aware burst restoration framework that interprets adjacent frames as observations under evolving flicker states. StateFlicker comprises two state-aware designs: Cross-State Visibility (CSV) and Flicker-State Understanding (FSU). CSV leverages complementary visibility across adjacent flicker states to selectively recover content suppressed under the center state. With this cross-state support, FSU constructs dual state pathways through a frozen world model to obtain predicted and observed center states, and distills their state discrepancy into dense features that characterize the center-specific flicker condition for restoration. Extensive experiments on the burst flicker removal dataset show that StateFlicker achieves superior performance over state-of-the-art methods.

1 Introduction

Burst imaging captures multiple short-exposure frames of the same scene, providing temporal observations that are unavailable in ordinary single-shot capture [18, 31, 12]. However, under artificial illumination, short-exposure bursts often suffer from flicker artifacts: AC-powered lighting varies rapidly over time, and row-wise exposure converts such variations into frame-varying brightness bands [22, 32]. These artifacts undermine the quality and reliability of burst observations, limiting their use in dynamic photography [10, 52, 5], HDR imaging [30, 38, 60], video reconstruction [20, 21, 42], and downstream visual analysis. Removing burst flicker is therefore critical to preserving the benefits of short-exposure imaging in artificially lit scenes.

From the perspective of single-image flicker removal, each degraded frame can be corrected independently by modeling illumination formation or learning spatial priors from flickering images [53, 29]. However, such single-frame correction remains inherently ambiguous, as flicker-induced darkness can resemble shadows or naturally low illumination. To exploit temporal evidence, burst restoration methods recover a target frame from neighboring observations, achieving strong results in low-light enhancement [9, 56, 23], denoising [31, 14, 51], deblurring [10, 3, 37], and super-resolution [46, 6, 25]. These methods typically regard neighboring frames as complementary observations and fuse them after motion compensation. This assumption becomes insufficient for burst flicker, where frame differences arise not only from scene content and motion, but also from periodic illumination states that modulate local visibility, creating frame-specific flicker states.

Recent work has begun to address burst flicker removal as a dedicated task. BurstDeflicker [39] establishes a systematic evaluation protocol, and FlickerFormer [40] further tailors a transformer

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

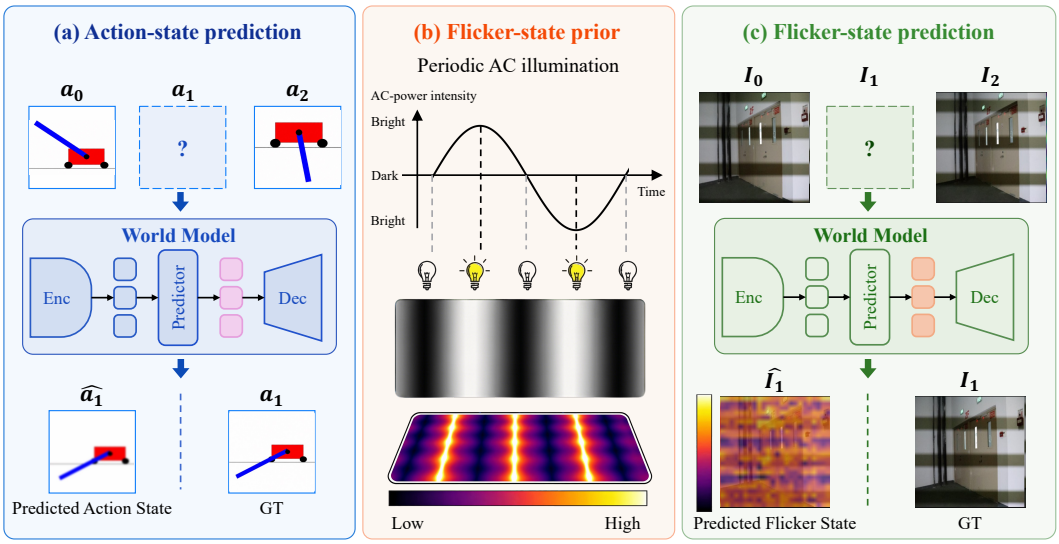


Figure 1: **Motivation of Flicker State**. (a) Pretrained world models predict the latent representation of an unseen state from observed context, given that the states evolve with temporally correlated dynamics [13]. (b) Under AC illumination, burst frames fall on different phases of the lighting cycle, and the resulting flicker degradation forms temporally correlated states across frames. (c) A world model takes adjacent flickering frames I_0 and I_2 as context and predicts the latent of the masked center frame I_1. The resulting heatmap closely aligns with the actual flicker-affected regions.

architecture with flicker-specific priors for periodic patterns and directional banding cues, but still relies on reference-centric fusion to use neighboring frames. Such a formulation mainly captures visible artifact regularities, while leaving the underlying flicker state implicit. However, adjacent frames differ not only in content but also in the flicker state under which observations are made. Distinguishing useful visibility from state-dependent degradation therefore requires explicit modeling of the latent flicker state. Since flicker follows a periodic illumination process, this state evolves in a temporally correlated manner, making it predictable from adjacent observations.

Directly learning this latent flicker state is challenging, as state annotations are unavailable and the degradation is entangled with scene content and motion. The latent predictive ability of pretrained world models, which capture latent state transitions from temporal context, provides a natural solution. As illustrated in Fig. 1(a), these models predict the latent representation of an unseen state from observed context [13, 17, 11]. Since burst flicker follows a periodic illumination process (Fig. 1(b)), this predictive ability naturally applies: a frozen world model predicts the latent representation of a masked center frame from adjacent flickering frames, and the resulting heatmap aligns with flicker-affected regions, as shown in Fig. 1(c). Comparing the predicted and observed center states further exposes the center-specific degradation without requiring explicit state annotations.

Based on this formulation, we propose StateFlicker, a state-aware framework with two complementary designs. Cross-State Visibility (CSV) leverages the fact that flicker states yield different local visibility. Since flicker-suppressed regions in the center frame may remain well-exposed in adjacent frames, CSV estimates a center-conditioned visibility weight for each neighboring observation, selectively recovering content evidence that is suppressed under the center state. Flicker-State Understanding (FSU) constructs dual state pathways through a frozen world model [4, 1] to obtain predicted and observed center states from adjacent-frame context, and distills their state discrepancy into multi-scale dense features that characterize the center flicker condition for restoration. Together, CSV and FSU address two key ambiguities left by reference-centric fusion: which cross-state observations should support content recovery, and what center-state degradation should be suppressed.

The main contributions are summarized as follows:

- We propose StateFlicker, a state-aware framework that formulates flicker as a predictable state and exploits world models for flicker-state understanding in burst flicker removal.

- We introduce Cross-State Visibility (CSV), which leverages complementary visibility across states to recover suppressed content, and Flicker-State Understanding (FSU), which distills the discrepancy between predicted and observed states to characterize flicker condition.
- Extensive experiments on the BurstDeflicker benchmark demonstrate that StateFlicker achieves superior quantitative performance and visual quality over state-of-the-art methods.

2 Related Work

Flicker Removal. Classical flicker removal methods rely on physical modeling or hardware compensation. AC-component analysis [53] and exposure-cycle modeling [24] estimate the illumination variation and correct brightness accordingly. Camera-system methods adapt exposure during capture [36], and circuit-level designs suppress flicker through LED backlight or power-driver control [34, 2, 8]. These approaches assume known illumination conditions or specialized hardware, limiting their applicability to diverse scenes. Learning-based single-image methods relax such constraints by learning spatial restoration mappings from data [29]. However, processing each frame independently makes flicker-darkened bands ambiguous with shadows or naturally dim regions, since temporal evidence from adjacent frames is unavailable to disambiguate the degradation.

Burst flicker removal exploits adjacent frames to resolve this ambiguity. BurstDeflicker [39] establishes the first benchmark for burst flicker removal. FlickerFormer [40] further introduces flicker-specific priors, employing phase correlation for inter-frame alignment, autocorrelation for intra-frame periodicity, and wavelet-based directional attention for stripe localization. These priors explicitly capture the periodic and directional appearance of flicker artifacts. Nevertheless, neighboring frames are still organized through reference-centric feature fusion, which targets observed artifact patterns rather than the underlying illumination process. StateFlicker instead treats each frame as an observation under an evolving flicker state and performs state-level reasoning for restoration.

Burst Image Restoration. Burst imaging leverages multiple short-exposure observations of the same scene to improve image quality [18]. The shared principle across burst methods is to compensate inter-frame motion and then aggregate complementary information. This idea has driven progress in diverse restoration tasks. Noisy observations are merged through learned spatially varying kernels [31] or deep multi-frame networks [14], and visibility in extreme low-light conditions is recovered by aggregating multiple dark frames [9, 56, 23]. Sub-pixel displacements across burst frames enable high-resolution detail recovery [6, 25], while Fourier-domain accumulation [10] and order-invariant aggregation [3] suppress camera shake in burst deblurring. More expressive architectures have also been explored. Hierarchical transformers align features at multiple scales [12], and spatiotemporal attention modules reduce ghosting in HDR reconstruction [30, 38, 60]. Despite their diversity, these methods generally assume that degradations are spatially homogeneous or content-independent within each frame, so that all aligned neighboring pixels contribute positively to the fused result.

Burst flicker does not satisfy this assumption. Flicker introduces periodic illumination states that selectively modulate local visibility across frames, producing spatially non-uniform and temporally varying degradation. The same spatial region can appear well-exposed in one frame and severely suppressed in an adjacent frame, depending on where each frame falls within the illumination cycle. A neighboring frame may therefore reveal content suppressed in the center frame, yet it simultaneously carries its own flicker-corrupted regions. Directly aligning and fusing such frames conflates recoverable visibility evidence with state-dependent degradation.

World Models. A world model learns how the world evolves over time from observed data and predicts future states conditioned on context [16]. Existing approaches span multiple paradigms. State transition approaches compress observations into compact latent states and learn temporal dynamics through recurrent updates for planning and control [16, 17, 50]. Video generation methods model appearance through diffusion or autoregressive architectures to produce high-fidelity visual futures [41, 43]. Embedding prediction methods such as JEPA [4, 33] predict target latent representations from contextual observations without reconstructing pixels, learning a representation space that captures both semantic and temporal structure. More recently, large-scale world foundation models [1] have further demonstrated strong temporal prediction across diverse video domains. A common capability shared across these paradigms is inferring unseen temporal states from partial observations, even when explicit state labels are unavailable.

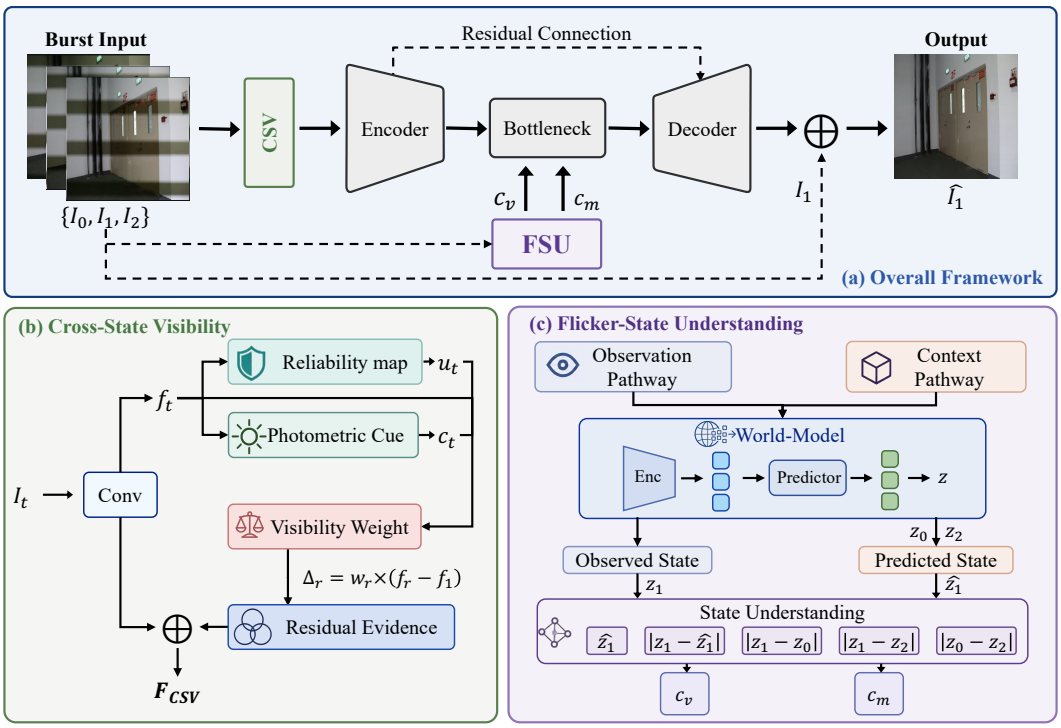


Figure 2: **Overview of StateFlicker**. (a) StateFlicker adopts a U-shaped architecture with CSV and FSU as two state-aware designs. (b) CSV estimates center-conditioned visibility weights from photometric cues and reliability maps, and selectively recovers suppressed content through residual evidence. (c) FSU predicts the hidden center state from adjacent observations via a world model and distills the predicted state into flicker-condition features.

Periodic flicker satisfies the temporal structure required by this capability: the illumination state evolves in a correlated manner across frames, so the degradation condition of each frame is predictable from its neighbors. Motivated by this property, StateFlicker exploits a frozen world model that predicts the hidden center flicker state from adjacent-frame context, and distills the predicted state into dense features that characterize the center-specific flicker condition for restoration.

3 Method

This section first presents an overview of the overall pipeline (Sec. 3.1). Two key components are then introduced within this flow: Cross-State Visibility (CSV), which uses state-conditioned neighboring observations with complementary visibility (Sec. 3.2), and Flicker-State Understanding (FSU), which constructs dual state pathways through a world model and distills the state discrepancy between predicted and observed center states to characterize the flicker condition (Sec. 3.3).

3.1 Overall Pipeline

The overall architecture of StateFlicker is shown in Fig. 2. Given three consecutive flickering frames $\{\mathbf{I}_0, \mathbf{I}_1, \mathbf{I}_2\} \in \mathbb{R}^{H\times W\times 3}$, StateFlicker restores the center frame $\mathbf{I}_1$ through a U-shaped encoder–decoder architecture following [54], which captures hierarchical features while preserving a concise single-output restoration pipeline for residual restoration. The restored image $\hat{\mathbf{I}}_1 = \mathbf{I}_1 + \mathbf{R}$ is obtained by adding the residual map $\mathbf{R} \in \mathbb{R}^{H\times W\times 3}$ to the degraded center frame.

More specifically, the input stage embeds the packed burst through Cross-State Visibility: motivated by complementary visibility across adjacent flicker states, it selects state-conditioned neighboring evidence to form $\mathbf{F}_{\text{CSV}} \in \mathbb{R}^{H\times W\times C}$. Then, FSU constructs dual state pathways through a frozen world model to obtain predicted and observed center states, and distills their state discrepancy into a

global condition vector $\mathbf{c}_v$ and a spatial condition map $\mathbf{c}_m$ that characterize the center flicker condition. The decoder subsequently restores the resolution back to $\mathbb{R}^{H\times W\times C}$ with residual connections.

3.2 Cross-State Visibility (CSV)

Different flicker states yield different local visibility: regions suppressed in the center frame may remain well-exposed in adjacent frames. However, the utility of each neighboring observation is inherently state-dependent, as certain regions may carry their own flicker-corrupted content rather than complementary evidence. Naively concatenating and convolving all frames disregards this distinction and conflates useful visibility evidence with state-dependent degradation. CSV addresses this issue by estimating a center-conditioned visibility weight for each neighboring observation, selectively recovering content that is suppressed under the center flicker state.

Given the packed three consecutive flicker frames, each frame is independently processed by a shared encoder to produce per-frame features $\mathbf{f}_t, t \in \{0, 1, 2\}$. To determine cross-state visibility, both photometric evidence and feature reliability are considered. The former captures flicker-induced intensity modulation and row-wise exposure fluctuation, while the latter estimates whether a local observation is trustworthy and compatible with the center-frame content. Accordingly, a visibility cue and a feature-space reliability map are extracted as:

$$\mathbf{c}_t = C(\mathbf{I}_t), \quad \mathbf{u}_t = U(\mathbf{f}_t) \in \mathbb{R}^{H\times W\times 1}, \tag{1}$$

where $C(\cdot)$ denotes the cue extractor composed of luminance, local contrast, and row-wise luminance deviation. Each neighboring observation is then compared with the center state using their cues, cue differences, relative reliability, and feature compatibility. This formulation favors neighboring evidence where the center feature is reliable, while suppressing structurally incompatible responses caused by motion or misalignment. A center-conditioned visibility weight is estimated as:

$$\mathbf{w}_t = W(\mathbf{f}_1, \mathbf{f}_t, \mathbf{c}_1, \mathbf{c}_t, \mathbf{u}_1, \mathbf{u}_t), \tag{2}$$

where $W(\cdot)$ denotes the visibility estimator. The resulting weight controls the contribution of the neighboring state at each spatial location, suppressing regions that carry their own flicker corruption. Instead of directly injecting neighboring features, CSV adopts a residual form to preserve the center feature as the primary representation and only supplement it with state-dependent corrections. The visibility-weighted residual from each neighbor and the final CSV embedding are computed as:

$$\Delta_t = \mathbf{w}_t \cdot (\mathbf{f}_1 - \mathbf{f}_t), \quad \mathbf{F}_{\text{CSV}} = \mathbf{F}_{\text{base}} + \gamma \cdot A(\mathbf{f}_1, \Delta_0, \Delta_2, \mathbf{w}_0, \mathbf{w}_2), \tag{3}$$

where Δ_t denotes the visibility-weighted feature residual of neighbor t; $\mathbf{F}_{\text{base}} \in \mathbb{R}^{H\times W\times C}$ represents a convolutional embedding of the packed burst; $A(\cdot)$ is a lightweight convolutional network; and γ is a learnable scalar initialized to zero, ensuring that CSV reduces to the standard embedding at initialization. Through this formulation, cross-state visibility differences are explicitly leveraged in early burst fusion, enabling each neighboring observation to contribute content evidence selectively based on its visibility relative to the center flicker state.

3.3 Flicker-State Understanding (FSU)

Building on the local visibility evidence provided by CSV, FSU focuses on the temporal evolution of flicker states across the burst. Neighboring frames are observed under different and complementary illumination conditions, while the center frame corresponds to the target state to be restored. Since pretrained world models [13, 4] have learned to predict latent representations of unseen frames from temporal context, this predictive ability can be naturally repurposed for flicker-state reasoning in burst restoration. Specifically, the predicted center state encodes what the world model expects given the adjacent context, whereas the observed center state reflects the actual flicker-degraded condition. FSU distills the state discrepancy between these two states to characterize the center flicker condition.

Following this formulation, FSU constructs two state pathways using the world model, which tokenizes every two consecutive frames into a temporal tubelet. The observation pathway arranges the input as $[\mathbf{I}_0, \mathbf{I}_2, \mathbf{I}_1, \mathbf{I}_1]$, and $Mid(\cdot)$ extracts the center-observed state $\mathbf{z}_1$. The neighbor-context

pathway uses $[\mathbf{I}_0, \mathbf{I}_0, \mathbf{I}_2, \mathbf{I}_2]$ to derive two adjacent context states without accessing the center frame, from which a lightweight predictor $P(\cdot)$ infers the predicted center state $\hat{\mathbf{z}}_1$. FSU then encodes the pairwise relations among the predicted, observed, and adjacent states into a state representation rather than treating these states directly as restoration features. The two pathways and the state representation are formulated as:

$$\mathbf{z}_1 = Mid\left(\mathcal{M}\left(\mathbf{I}_0, \mathbf{I}_2, \mathbf{I}_1, \mathbf{I}_1\right)\right), \quad \hat{\mathbf{z}}_1 = P\left(\mathcal{M}\left(\mathbf{I}_0, \mathbf{I}_0, \mathbf{I}_2, \mathbf{I}_2\right)\right), \tag{4}$$

$$\mathbf{S}_{\mathrm{FSU}} = [\hat{\mathbf{z}}_1,\ |\mathbf{z}_1 - \hat{\mathbf{z}}_1|,\ |\mathbf{z}_1 - \mathbf{z}_0|,\ |\mathbf{z}_1 - \mathbf{z}_2|,\ |\mathbf{z}_0 - \mathbf{z}_2|], \tag{5}$$

where $\mathcal{M}$ denotes the frozen world model, and $[\cdot]$ denotes channel-wise concatenation. The first difference term captures the discrepancy between the observed and predicted center states, and the remaining terms encode how the center observation relates to its adjacent states. FSU projects $\mathbf{S}_{\mathrm{FSU}}$ into a global condition vector $\mathbf{c}_v$ that summarizes sequence-level flicker dynamics and a spatial condition map $\mathbf{c}_m$ that preserves local state variation. These two conditions are then fused with the encoded feature to provide multi-scale state-level degradation cues for decoding.

4 Experiments

4.1 Experimental Settings

Datasets. To comprehensively evaluate StateFlicker, experiments are conducted on the BurstDeflicker benchmark [39]. Following FlickerFormer [40], each input consists of three consecutive flickering frames, and the output is the restored center frame. All methods are trained and evaluated on the same official training and testing splits to ensure fair comparison.

Implementation Details. StateFlicker adopts a 3-level encoder-decoder architecture with a base channel dimension of 32. The model is optimized using AdamW with a learning rate of 3×10^{-4} for $300,000$ iterations. The training objective combines L_1, perceptual, and FFT losses with weights of 1.0, 0.01, and 0.01, respectively.

4.2 Quantitative Comparison

StateFlicker is compared with 16 methods on the BurstDeflicker benchmark [39] in Tab. 1. All methods receive the same three-frame input and are grouped by their original design focus: general image restoration, burst restoration, HDR reconstruction, and flicker-specific approaches. General restoration methods such as Restormer [54] and AST [57] achieve around 30.6 dB PSNR through powerful spatial modeling, yet their architectures are not designed to exploit the state-dependent visibility differences across adjacent flicker frames. Burst methods such as Burstormer [12] and SAFNet [27] are built for multi-frame processing but assume spatially homogeneous degradations, which fails under periodic flicker. Their PSNR scores remain below 29.5 dB, lower than the best general baselines, as the state-dependent visibility variation of flicker is not modeled. HDR methods including HDRTransformer [30] and FPro [59] reach 30.0–30.6 dB by handling exposure variation, while the spatially varying periodic pattern of flicker falls outside their design scope. Among flicker-aware methods, FlickerFormer [40] narrows this gap by incorporating periodicity and directional banding priors, achieving 31.226 dB

Table 1: Quantitative results on burst flicker removal dataset [39]. The best scores are highlighted in **bold**.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
Stripformer$_{\text{ECCV22}}$ [45]	29.223	0.892	0.058
Uformer$_{\text{CVPR22}}$ [47]	30.544	0.910	0.056
Restormer$_{\text{CVPR22}}$ [54]	30.630	0.917	0.055
HDRTransformer$_{\text{ECCV22}}$ [30]	30.031	0.918	0.054
Retinxformer$_{\text{ICCV23}}$ [7]	29.598	0.899	0.055
FFTformer$_{\text{CVPR23}}$ [26]	29.478	0.895	0.050
Burstormer$_{\text{CVPR23}}$ [12]	29.439	0.910	0.056
FBANet$_{\text{ICCV23}}$ [49]	29.459	0.896	0.052
MambaIR$_{\text{ECCV24}}$ [15]	29.478	0.904	0.060
SAFNet$_{\text{ECCV24}}$ [27]	29.223	0.892	0.058
FPro$_{\text{ECCV24}}$ [59]	30.551	0.910	0.051
AST$_{\text{CVPR24}}$ [57]	30.646	0.918	0.050
AFUNet$_{\text{ICCV25}}$ [28]	28.922	0.903	0.066
RT-XNet$_{\text{ICIP25}}$ [19]	29.718	0.909	0.058
HINT$_{\text{ICCV25}}$ [58]	30.336	0.916	0.046
Flickerformer$_{\text{CVPR26}}$ [40]	31.226	0.920	0.045
StateFlicker (ours)	**32.116**	**0.921**	**0.044**

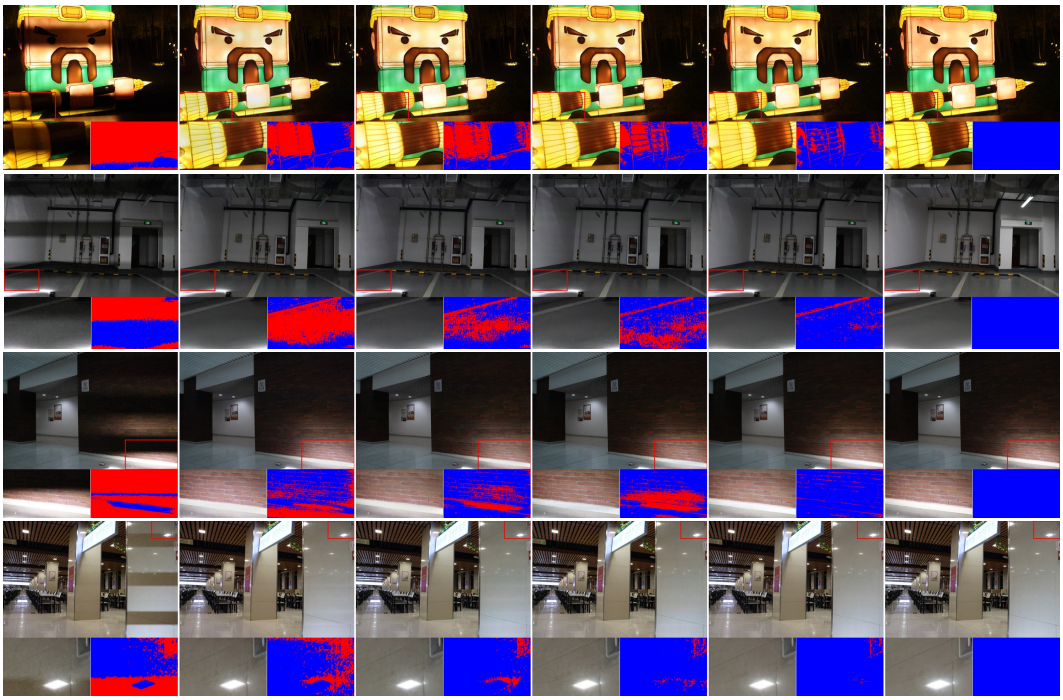


Figure 3: **Qualitative comparison of burst flicker removal**. StateFlicker delivers superior restoration quality across scenes. Cropped regions marked by red boxes are enlarged for visual comparison. In difference maps, red pixels denote high-error regions identified under a fixed threshold.

PSNR and the best LPIPS [55] of 0.045. StateFlicker obtains the highest PSNR, SSIM [48], and LPIPS [55] of 32.116 dB, 0.921, and 0.044, surpassing FlickerFormer by 0.890 dB in PSNR. The consistent improvement over both generic and flicker-specific methods indicates that explicit flicker-state modeling provides complementary benefits beyond artifact-level spatial priors.

4.3 Qualitative Results

Qualitative comparisons are presented in Fig. 3. Following FlickerFormer [40], the difference maps highlight regions with reconstruction errors above a fixed threshold in red, making local restoration failures easier to inspect. HDRTransformer [30] and AST [57] recover part of the suppressed intensity. However, evident state-dependent distortions remain in their outputs. For the first scene, both methods amplify illuminated surfaces and produce horizontal artifacts within the red-box region. In the parking-garage scene, large red responses appear around the light source and along the bright floor boundary. This indicates inaccurate reconstruction under abrupt local illumination changes. Similar degradations are observed on the hallway wall. The brick texture is over-smoothed or contaminated by residual banding. Although FlickerFormer [40] reduces large-scale flicker artifacts, high-error regions remain in several local areas. These errors appear around the floor highlight in the garage scene, the brick-wall crop in the hallway scene, and the ceiling area in the lobby scene.

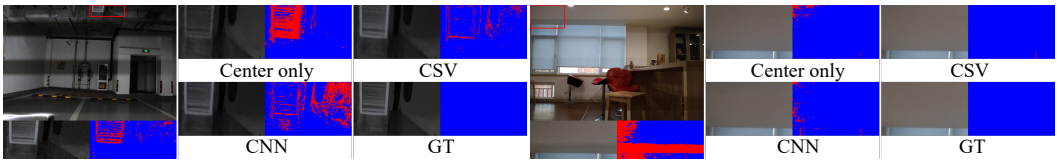
By comparison, StateFlicker yields cleaner restorations with fewer error responses. It better preserves structures in the first crop, reduces errors near the garage light and floor boundary, recovers more coherent hallway bricks with less residual banding, and maintains smoother lobby illumination. Residual flicker near image corners and boundaries is also suppressed more effectively, where competing methods still show localized red responses. These improvements come from the state-aware design: CSV exploits complementary visibility across flicker states to recover center-suppressed content, while FSU characterizes the center-specific flicker condition through world-model prediction.

Table 2: Overall ablation of state-aware parts.

Base	CSV	FSU	PSNR ↑	SSIM ↑	LPIPS ↓
✓			31.554	0.918	0.054
✓	✓		31.775	0.920	0.051
✓	✓	✓	**32.116**	**0.921**	**0.044**

Table 3: Ablation on exploiting adjacent-frame states for cross-state visibility modeling.

Center only	CNN	CSV	PSNR ↑	SSIM ↑	LPIPS ↓
✓			27.468	0.896	0.076
	✓		31.554	0.918	0.054
		✓	**31.775**	**0.920**	**0.051**



(a) Effect of cross-state visibility in burst embedding, comparing center-only, CNN, CSV, and GT.



(b) Effect of feature source for FSU, comparing DINO, VideoMAE, world-model features, and GT.



(c) Effect of the state condition in FSU, comparing random, averaged states, the full FSU condition, and GT.

Figure 4: Visual ablation results of StateFlicker, covering three design aspects: (a) cross-state visibility for selective burst embedding, (b) feature-source selection for flicker-state prediction, and (c) state-condition construction for restoration. Red pixels in the error maps denote high-error regions identified under a fixed threshold. Zoom in for better view.

4.4 Ablation Study

The proposed state-aware designs are analyzed through quantitative results and visual error maps. As shown in Tab. 2, both CSV and FSU contribute to the improvement. The two modules address complementary ambiguities in burst flicker removal. CSV replaces generic multi-frame aggregation with visibility-aware evidence selection at the embedding stage, while FSU provides a center-state degradation cue from temporal context. Together, they decouple cross-state evidence reliability from center-specific flicker characterization, which are entangled in standard multi-frame processing.

Effect of cross-state visibility. Three burst observation strategies are compared in Tab. 3. The center-only baseline has no neighboring reference, leaving flicker-suppressed regions indistinguishable from naturally dark content. The CNN baseline packs all three frames into a multi-channel input, providing access to complementary content but aggregating it uniformly without distinguishing well-exposed neighboring regions from flicker-corrupted ones. CSV instead estimates a center-conditioned per-pixel visibility weight from photometric cues and feature-space reliability. This weight controls the contribution of each neighbor in a residual form, keeping the center content primary.

The visual comparison in Fig. 4 (a) supports this analysis. Center-only restoration leaves broad red error regions in flicker-suppressed areas, confirming that the missing content cannot be hallucinated from a single degraded observation. The CNN baseline reduces some errors by leveraging neighboring content. However, structured residual responses remain where the neighboring frames are themselves flicker-corrupted, indicating that indiscriminate aggregation propagates state-dependent artifacts into the output. CSV produces cleaner error maps closer to GT, especially around locally suppressed or boundary regions. This improvement is consistent across scenes with different flicker intensities,

Table 4: Ablation on feature sources of FSU.

Feature source	PSNR ↑	SSIM ↑	LPIPS ↓
DINO [35]	31.117	0.915	0.056
VideoMAE [44]	31.833	0.918	0.050
World Model [4]	**32.116**	**0.921**	**0.044**

Table 5: Ablation on the state condition of FSU.

Condition	PSNR ↑	SSIM ↑	LPIPS ↓
Random	31.730	0.917	0.053
Avg	31.768	0.916	0.053
FSU	**32.116**	**0.921**	**0.044**

suggesting that the visibility weighting adapts to varying cross-state complementarity rather than following a fixed aggregation pattern.

Effect of flicker-state features. FSU depends on a feature source that can support center-state prediction from neighboring observations. DINO, VideoMAE, and world-model features are compared in Tab. 4, with the restoration network and CSV unchanged. DINO features [35] are extracted by a pretrained encoder that processes each frame independently. They therefore capture per-frame appearance rather than the continuous evolution of flicker states required by the predict-and-compare formulation of FSU. VideoMAE features [44] are learned by reconstructing masked tubelets within a video clip. Although they capture temporal continuity, they are optimized to recover observed content rather than to infer state transitions toward an unobserved frame. World-model features [4] instead jointly encode consecutive frames and learn temporally continuous state transitions. This property makes them well suited for predicting the center flicker state from adjacent observations.

The difference is visualized in Fig. 4 (b). DINO features leave dense red errors in flicker-affected regions. VideoMAE reduces these errors but still leaves scattered responses, whereas the world-model variant suppresses them more effectively and approaches GT. Lacking either temporal continuity or a predictive objective, DINO and VideoMAE fail to infer the expected center state and its deviation. These results verify that FSU benefits from modeling state transitions toward an unseen frame, rather than from per-frame structure or generic temporal reconstruction.

Effect of flicker-state condition. Three state conditions are compared in Tab. 5 by replacing only the state representation $\mathbf{S}_{\mathrm{FSU}}$, with all other settings fixed. Random features serve as a control for additional conditioning capacity. Since they encode no flicker information, they fail to improve over the Base+CSV setting in Tab. 2. Avg averages the two adjacent states and bypasses the predictor $P(\cdot)$. This average summarizes neighboring observations but does not reveal how the center frame deviates from them, and thus brings no measurable gain. The full FSU instead predicts the expected center state from adjacent context without accessing the center frame. Comparing this prediction with the observed state isolates the center-specific flicker deviation, yielding consistent improvements across all metrics. The benefit of FSU therefore stems from predictive state reasoning rather than extra capacity or direct temporal aggregation.

The visual comparison in Fig. 4 (c) supports this analysis. Random and Avg leave scattered red responses along structural edges, such as the exit sign and the chair frame. Neither extra conditioning capacity nor a direct summary of adjacent states thus provides usable cues. The full FSU further suppresses these residual errors and yields maps closer to GT, suggesting that the predicted-vs-observed discrepancy identifies the center-specific degradation to be suppressed.

5 Conclusion

This paper presents StateFlicker, a state-aware burst restoration framework that formulates flicker degradation as a predictable temporal state and exploits the predictive capacity of world models for burst flicker removal. The proposed Cross-State Visibility (CSV) module leverages complementary visibility across adjacent flicker states, selectively recovering content suppressed under the center state through residual evidence from neighboring observations. The proposed Flicker-State Understanding (FSU) module constructs dual state pathways through a frozen world model to obtain predicted and observed center states, and distills their state discrepancy into dense features that characterize the center-specific flicker condition for restoration. Extensive experiments on the BurstDeflicker benchmark demonstrate that StateFlicker achieves state-of-the-art quantitative performance and visual quality, validating that flicker-state modeling benefits restoration beyond spatial priors.

Acknowledgment

This work was supported in part by the NSFC Science Fund Program for Young Scientists Fund (Grant No. 62406089), in part by the Guangdong Basic and Applied Basic Research Foundation (2025A1515011361), in part by the Shenzhen Science and Technology Program (JCYJ20240813110459017), in part by the Guangdong Provincial Key Laboratory of Artificial General Intelligence and Large Models (2026B1212010005), in part by the National Natural Science Foundation of China (Grant No. 62372480).

References

[1] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

[2] Hyun-A Ahn, Seong-Kwan Hong, and Oh-Kyong Kwon. A highly accurate current LED lamp driver with removal of low-frequency flicker using average current control method. *IEEE Transactions on Power Electronics*, 2017.

[3] Miika Aittala and Frédo Durand. Burst image deblurring using permutation invariant convolutional neural networks. In *ECCV*, 2018.

[4] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

[5] Guangrui Bai, Hailong Yan, Wenhai Liu, Yahui Deng, and Erbao Dong. Towards lightest low-light image enhancement architecture for mobile devices. *Expert Systems with Applications*, page 129125, 2026.

[6] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Deep burst super-resolution. In *CVPR*, 2021.

[7] Yuanhao Cai, Hao Bian, Jing Lin, Haoqian Wang, Radu Timofte, and Yulun Zhang. Retinexformer: One-stage retinex-based transformer for low-light image enhancement. In *ICCV*, 2023.

[8] Ignacio Castro, Aitor Vazquez, Manuel Arias, Diego G Lamar, Marta M Hernando, and Javier Sebastian. A review on flicker-free AC–DC LED drivers for single-phase and three-phase AC power grids. *IEEE Transactions on Power Electronics*, 2019.

[9] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *CVPR*, 2018.

[10] Mauricio Delbracio and Guillermo Sapiro. Burst deblurring: Removing camera shake through fourier burst accumulation. In *CVPR*, 2015.

[11] Jingtao Ding, Yunke Zhang, Yu Shang, Yuheng Zhang, Zefang Zong, Jie Feng, Yuan Yuan, Hongyuan Su, Nian Li, Nicholas Sukiennik, et al. Understanding world or predicting future? a comprehensive survey of world models. *ACM Computing Surveys*, 2025.

[12] Akshay Dudhane, Syed Waqas Zamir, Salman Khan, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Burstormer: Burst image restoration and enhancement transformer. In *CVPR*, 2023.

[13] Daniel Freeman, David Ha, and Luke Metz. Learning to predict without looking ahead: World models without forward prediction. In *NeurIPS*, 2019.

[14] Clément Godard, Kevin Matzen, and Matt Uyttendaele. Deep burst denoising. In *ECCV*, 2018.

[15] Hang Guo, Jinmin Li, Tao Dai, Zhihao Ouyang, Xudong Ren, and Shu-Tao Xia. Mambair: A simple baseline for image restoration with state-space model. In *ECCV*, 2024.

[16] David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *NeurIPS*, 2018.

[17] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *ICLR*, 2019.

[18] Samuel W Hasinoff, Dillon Sharlet, Ryan Geiss, Andrew Adams, Jonathan T Barron, Florian Kainz, Jiawen Chen, and Marc Levoy. Burst photography for high dynamic range and low-light imaging on mobile cameras. *ACM TOG*, 2016.

[19] Raman Jha, Adithya Lenka, Mani Ramanagopal, Aswin Sankaranarayanan, and Kaushik Mitra. Rt-x net: Rgb-thermal cross attention network for low-light image enhancement. In *ICIP*, 2025.

[20] Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In *CVPR*, 2018.

[21] Meiguang Jin, Zhe Hu, and Paolo Favaro. Learning to extract flawless slow motion from blurry videos. In *CVPR*, 2019.

[22] Ali Kanj, Hugues Talbot, and Raoul Rodriguez Luparello. Flicker removal and superpixel-based motion tracking for high speed videos. In *ICIP*, 2017.

[23] Ahmet Serdar Karadeniz, Erkut Erdem, and Aykut Erdem. Burst photography for learning to enhance extremely dark images. *IEEE TIP*, 2021.

[24] Minwoong Kim, Kurt Bengtson, Lisa Li, and Jan P Allebach. Reducing flicker due to ambient illumination in camera captured images. In *Color Imaging XVIII: Displaying, Processing, Hardcopy, and Applications*, 2013.

[25] Sanghyun Kim, Minjung Lee, Woohyeok Kim, Deunsol Jung, Jaesung Rim, Sunghyun Cho, and Minsu Cho. Burst image super-resolution with base frame selection. In *CVPR*, 2024.

[26] Lingshun Kong, Jiangxin Dong, Jianjun Ge, Mingqiang Li, and Jinshan Pan. Efficient frequency domain-based transformers for high-quality image deblurring. In *CVPR*, 2023.

[27] Lingtong Kong, Bo Li, Yike Xiong, Hao Zhang, Hong Gu, and Jinwei Chen. Safnet: Selective alignment fusion network for efficient hdr imaging. In *ECCV*, 2024.

[28] Xinyue Li, Zhangkai Ni, and Wenhan Yang. Afunet: Cross-iterative alignment-fusion synergy for hdr reconstruction via deep unfolding paradigm. *ICCV*, 2025.

[29] Xiaodan Lin, Yangfu Li, Jianqing Zhu, and Huanqiang Zeng. Deflickercyclegan: Learning to detect and remove flickers in a single image. *IEEE TIP*, 2023.

[30] Zhen Liu, Yinglong Wang, Bing Zeng, and Shuaicheng Liu. Ghost-free high dynamic range imaging with context-aware transformer. In *ECCV*, 2022.

[31] Ben Mildenhall, Jonathan T Barron, Jiawen Chen, Dillon Sharlet, Ren Ng, and Robert Carroll. Burst denoising with kernel prediction networks. In *CVPR*, 2018.

[32] Naomi J Miller, Felipe A Leon, Jianchuan Tan, and L Irvin. Flicker: A review of temporal light modulation stimulus, responses, and measures. *Lighting Research & Technology*, 2023.

[33] Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mido Assran, Koustuv Sinha, Mike Rabbat, Yann LeCun, Nicolas Ballas, and Adrien Bardes. V-jepa 2.1: Unlocking dense features in video self-supervised learning. *arXiv preprint arXiv:2603.14482*, 2026.

[34] Ehsan Nadernejad, Claire Mantel, Nino Burini, and Søren Forchhammer. Flicker reduction in LED-LCDs with local backlight. In *MMSPW*, 2013.

[35] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[36] SangHyun Park, GyuWon Kim, and JaeWook Jeon. The method of auto exposure control for low-end digital camera. In *ICACT*, 2009.

[37] Fidel Alejandro Guerrero Peña, Pedro Diamel Marrero Fernández, Tsang Ing Ren, Jorge de Jesus Gomes Leandro, and Ricardo Massahiro Nishihara. Burst ranking for blind multi-image deblurring. *IEEE TIP*, 2019.

[38] Tianheng Qiu, Qi Wu, Yuchun Dong, Shenglin Ding, Xuan Huang, Hu Wei, and Guanghua Pan. Recursive multi-exposure alignment with spatiotemporal decoupling for efficient burst hdr and restoration. In *CVPR*, 2025.

[39] Lishen Qu, Zhihao Liu, Shihao Zhou, Yaqi Luo, Jie Liang, Hui Zeng, Lei Zhang, and Jufeng Yang. Burstdeflicker: A benchmark dataset for flicker removal in dynamic scenes. In *NeurIPS*, 2025.

[40] Lishen Qu, Shihao Zhou, Jie Liang, Hui Zeng, Lei Zhang, and Jufeng Yang. It takes two: A duet of periodicity and directionality for burst flicker removal. *arXiv preprint arXiv:2603.22794*, 2026.

[41] Zhongwei Ren, Yunchao Wei, Xun Guo, Yao Zhao, Bingyi Kang, Jiashi Feng, and Xiaojie Jin. Videoworld: Exploring knowledge learning from unlabeled videos. In *CVPR*, 2025.

[42] Wang Shen, Wenbo Bao, Guangtao Zhai, Li Chen, Xiongkuo Min, and Zhiyong Gao. Blurry video frame interpolation. In *CVPR*, 2020.

[43] Robbyant Team, Zelin Gao, Qiuyu Wang, Yanhong Zeng, Jiapeng Zhu, Ka Leong Cheng, Yixuan Li, Hanlin Wang, Yinghao Xu, Shuailei Ma, et al. Advancing open-source world models. *arXiv preprint arXiv:2601.20540*, 2026.

[44] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *NeurIPS*, 2022.

[45] Fu-Jen Tsai, Yan-Tsung Peng, Yen-Yu Lin, Chung-Chi Tsai, and Chia-Wen Lin. Stripformer: Strip transformer for fast image deblurring. In *ECCV*, 2022.

[46] Roger Y Tsai and Thomas S Huang. Multiframe image restoration and registration. *Multiframe image restoration and registration*, 1984.

[47] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *CVPR*, 2022.

[48] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 2004.

[49] Pengxu Wei, Yujing Sun, Xingbei Guo, Chang Liu, Guanbin Li, Jie Chen, Xiangyang Ji, and Liang Lin. Towards real-world burst image super-resolution: Benchmark and method. In *ICCV*, 2023.

[50] Philipp Wu, Alejandro Escontrela, Danijar Hafner, Pieter Abbeel, and Ken Goldberg. Daydreamer: World models for physical robot learning. In *Conference on robot learning*, 2023.

[51] Hailong Yan, Junjian Huang, and Tingwen Huang. Igdnet: Zero-shot robust underexposed image enhancement via illumination-guided and denoising. *IEEE Transactions on Artificial Intelligence*, 2025.

[52] Hailong Yan, Ao Li, Xiangtao Zhang, Zhe Liu, Zenglin Shi, Ce Zhu, and Le Zhang. Mobileie: An extremely lightweight and effective convnet for real-time image enhancement on mobile devices. In *ICCV*, 2025.

[53] Yoonjong Yoo, Jaehyun Im, and Joonki Paik. Flicker removal for cmos wide dynamic range imaging based on alternating current component analysis. *IEEE Transactions on Consumer Electronics*, 2014.

[54] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, 2022.

[55] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[56] Di Zhao, Lan Ma, Songnan Li, and Dahai Yu. End-to-end denoising of dark burst images using recurrent fully convolutional networks. *arXiv preprint arXiv:1904.07483*, 2019.

[57] Shihao Zhou, Duosheng Chen, Jinshan Pan, Jinglei Shi, and Jufeng Yang. Adapt or perish: Adaptive sparse transformer with attentive feature refinement for image restoration. In *CVPR*, 2024.

[58] Shihao Zhou, Dayu Li, Jinshan Pan, Juncheng Zhou, Jinglei Shi, and Jufeng Yang. Devil is in the uniformity: Exploring diverse learners within transformer for image restoration. In *ICCV*, 2025.

[59] Shihao Zhou, Jinshan Pan, Jinglei Shi, Duosheng Chen, Lishen Qu, and Jufeng Yang. Seeing the unseen: A frequency prompt guided transformer for image restoration. In *ECCV*, 2024.

[60] Weiyu Zhou, Tao Hu, Yixu Feng, Duwei Dai, Yu Cao, Peng Wu, Wei Dong, Yanning Zhang, and Qingsen Yan. Flow-guided deformable alignment with channel-wise self-attention reconstruct for efficient burst hdr restoration. In *CVPR*, 2025.