applied sciences

MDPI

Article

Dynamic Multi-Exposure HDR Reconstruction via Dual-Branch Base-Detail Collaboration

Qin Zhou [1,2], Min Chen [1,2], Feifan Cai [1,2], Zihao Zhang [1,2] and Youdong Ding [1,2,*]

[1] Shanghai Film Academy, Shanghai University, Shanghai 200072, China; qin1016@shu.edu.cn (Q.Z.); mailforcm@shu.edu.cn (M.C.); feifan@shu.edu.cn (F.C.); zhcheung@shu.edu.cn (Z.Z.)
[2] Shanghai Engineering Research Center of Motion Picture Special Effects, Shanghai University, Shanghai 200072, China
* Correspondence: ydding@shu.edu.cn

Abstract

Dynamic multi-exposure high dynamic range (HDR) image reconstruction remains challenging because it must preserve globally consistent luminance and structure while recovering fine-grained local textures from low dynamic range (LDR) inputs corrupted by saturation, under-exposure, and motion-induced artifacts. Existing CNN-based methods are effective at local detail restoration but remain limited in global context modeling, whereas Transformer-based methods improve long-range interaction but can still weaken local-detail refinement. Current hybrid designs suggest that the two representation types are complementary, but they do not fully address branch specialization, cross-branch collaboration, and local-feature reliability control. To address this gap, we propose a dual-branch Transformer-CNN framework with a base branch built on Window-based Residual Transformer Blocks (WRTBs), a detail branch equipped with Detail-Aware Gating (DAG) for reliability-aware local refinement, and Bidirectional Cross-Branch Fusion (BCBF) for stage-wise collaboration between the two branches. Experiments on Kalantari17, the Tel benchmark, and Challenge123 show that the proposed design remains competitive on the standard benchmark, achieving the best HDR-VDP2 and tied best with μ-SSIM on Kalantari17, while yielding clearer gains on the more challenging Tel and Challenge123 benchmarks.

Keywords: HDR imaging; multi-exposed imaging; HDR reconstruction; computer vision

1. Introduction

High dynamic range (HDR) imaging is important for many real-world vision systems, including mobile imaging, surveillance, robotics, medical imaging, and advanced driver assistance. In dynamic scenes, however, multi-exposure HDR reconstruction must fuse observations corrupted by both motion and severe illumination variation. Object or camera motion introduces ghosting artifacts, while saturated highlights and under-exposed shadows remove reliable radiometric evidence. Dynamic HDR reconstruction must, therefore, satisfy two coupled goals at once: preserving globally consistent tone and structure, and restoring locally accurate textures from spatially unreliable observations.

Existing methods address these two goals unevenly. CNN-based HDR reconstruction methods, such as AHDRNet [1] and NHDRRNet [2], are effective at preserving local textures through convolutional priors and attention-guided fusion. Their ability to model long-range structural and luminance dependencies, however, remains limited when large

Check for updates

Academic Editor: Pedro Couto

Received: 31 March 2026
Revised: 17 April 2026
Accepted: 18 April 2026
Published: 23 April 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

motion, severe saturation, or complex scene layouts require broader context. Transformer-based methods improve this aspect. In particular, HDR-Transformer [3] demonstrates the benefit of hierarchical global-context modeling, and SCTNet [4] further strengthens alignment-free semantic consistency in dynamic scenes. However, these methods still primarily rely on a dominant Transformer-style representation and do not explicitly protect reliability-sensitive local detail recovery. Existing hybrid designs, such as HyHDR [5], further support the complementarity of CNN and Transformer representations. Yet, they still leave three practical design questions insufficiently answered for dynamic HDR reconstruction: how to assign clear roles to the base and detail branches, how to make these branches collaborate effectively once separated, and how to prevent unreliable local responses from contaminating that collaboration.

To address this gap, we propose a dual-branch Transformer-CNN framework centered on **Bidirectional Cross-Branch Fusion (BCBF)** for dynamic multi-exposure HDR reconstruction. The framework consists of a Transformer-based base branch and a CNN-based detail branch. The base branch is built on **Window-based Residual Transformer Blocks (WRTBs)** and is responsible for global structure and luminance modeling, whereas the detail branch incorporates **Detail-Aware Gating (DAG)** for reliability-aware local texture recovery. BCBF then enables stage-wise bidirectional exchange between the two branches so that global context and local detail cues can refine each other throughout the network. This design suppresses detail responses corrupted by saturation, under-exposure, and motion before cross-branch interaction. Together, these components couple stable structure modeling with selective detail refinement in a unified reconstruction pipeline.

The main contributions of this work are summarized as follows:

1. **A BCBF-centered cross-branch collaboration mechanism.** We propose Bidirectional Cross-Branch Fusion (BCBF) as a branch-collaboration mechanism for stage-wise bidirectional interaction between the base and detail branches, improving their coordination without explicit optical-flow alignment.
2. **A branch-specialized base-detail architecture for collaboration.** We build a dual-branch Transformer-CNN framework in which WRTB-based base modeling and CNN-based detail refinement are explicitly separated, providing a task-appropriate setting for effective BCBF-based collaboration.
3. **Reliability-aware detail refinement before branch interaction.** We integrate a DAG-based Detail-Aware Gating (DAG) design into the detail branch to suppress unreliable local responses and improve the quality of detail features before cross-branch interaction.
4. **Systematic validation across standard and challenging HDR benchmarks.** We provide benchmark comparisons on Kalantari17 [6], the Tel benchmark [4], and Challenge123 [7], together with complexity profiling and targeted ablations, to analyze the effectiveness of the proposed design on the standard benchmark and under more difficult illumination and motion conditions.

2. Related Work

2.1. Traditional Multi-Exposure Fusion

Early multi-exposure HDR methods mainly reduced ghosting in two ways: rejecting unreliable pixels or explicitly aligning exposures before fusion. Rejection-based methods, such as those of Grosch [8], Pece and Kautz [9], and Jacobs et al. [10], detected motion and suppressed suspicious regions, but they often discarded useful information together with the ghosts. Registration-based approaches estimated correspondences through optical flow [11,12], patch-based optimization [13], or additional brightness and gradient constraints [14]. These methods established the importance of motion handling, but they

also exposed a persistent difficulty: when saturation, under-exposure, occlusion, or large displacement corrupts local evidence, explicit matching itself becomes fragile.

2.2. CNN-Based and Attention-Based Methods

Deep learning substantially improved HDR reconstruction by learning alignment and fusion from data. Flow-guided CNN pipelines, such as that of Kalantari and Ramamoorthi [6], explicitly aligned exposures before merging, while Wu et al. [15] targeted large foreground motion through learned translation. ADNet [16] further introduced attention-guided deformable alignment to improve robustness under dynamic motion. More recent designs, such as SAFNet [7], further emphasized selective use of unreliable regions under motion and saturation. In parallel, attention-based CNNs pursued implicit alignment: AHDRNet [1] used spatial attention to suppress misaligned features, and NHDRRNet [2] introduced non-local interactions to strengthen contextual reasoning. Overall, CNN-based methods remained strong at local texture preservation and artifact suppression, but their broader structural and luminance modeling was still comparatively limited when the scene required more global context.

2.3. Transformer-Based HDR Reconstruction

Transformers were introduced into HDR reconstruction to capture long-range dependencies and global interactions that are difficult to model with purely local convolutions. Selective TransHDR [17] used a ghost-region-aware selective Transformer strategy, HDR-Transformer [3] adopted a hierarchical Transformer architecture, and SCTNet [4] further explored alignment-free reconstruction through semantics-consistent global reasoning. HDR-Transformer is particularly representative of the shift toward hierarchical global-context modeling: it shows that Transformer backbones can better coordinate luminance and structure over large spatial ranges than purely convolutional pipelines. SCTNet further demonstrates that stronger semantic consistency can reduce reliance on explicit alignment in difficult dynamic HDR scenes. These methods clearly establish the value of Transformer representations for global structure and luminance consistency. However, they also leave an important gap. Their main emphasis remains global interaction within a dominant Transformer-oriented stream, so local high-frequency refinement is not explicitly separated from global modeling, nor is unreliable local evidence selectively filtered before later fusion. More recent work also explored stronger semantic priors under degraded inputs [18]. Transformer-heavy designs, therefore, improve global reasoning, but they can still weaken detail preservation when saturation, under-exposure, and motion corruption make local observations unreliable.

2.4. Hybrid Transformer-CNN Designs

Hybrid designs attempted to combine the complementary strengths of Transformers and CNNs, namely, global dependency modeling and local texture-preserving inductive bias. In HDR reconstruction, HyHDR [5] showed the benefit of combining pixel-level and patch-level representations, confirming that global and local cues are, indeed, complementary rather than interchangeable. Mai et al. [19] explored a deep-unrolled low-rank formulation for HDR imaging, AFUNet [20] introduced a deep-unfolding perspective on iterative alignment and fusion, and dual-domain designs further explored complementary spatial and frequency cues [21]. These studies support the value of hybrid modeling, but they do not fully address how the two representation types should collaborate under saturation, under-exposure, and motion corruption. In particular, the interaction between global and local streams often remains implicit, late-stage, or insufficiently specialized. As a result, three design questions remain open for dynamic HDR scenes: how to assign explicit roles to the base and detail representations, how to make them collaborate once separated,

and how to filter unreliable detail evidence before it enters cross-branch interaction. Our method is positioned in this gap.

2.5. Generative and Diffusion-Based Approaches

Beyond regression-based pipelines, generative models aimed to improve perceptual quality in severely degraded regions. HDR-GAN [22] used adversarial learning to synthesize plausible details, while diffusion-based methods such as HDR-V-Diff [23] and UltraFusion [24] introduced stronger generative priors for ambiguous content. Yan et al. [25] further adopted a step-wise strategy that first generated pseudo-static LDR images and then reconstructed HDR with the help of motion-region segmentation [26]. These approaches are relevant because they address missing or unreliable content through stronger priors, but they also shift the emphasis from deterministic reconstruction toward prior-driven synthesis and usually incur higher inference cost than the supervised reconstruction setting considered in this work.

3. Methodology

3.1. Problem Setup

Let $\{L^{(k)}\}_{k=1}^{3}$ denote the input LDR exposure bracket from a dynamic scene, where each frame $L^{(k)} \in \mathbb{R}^{3\times H\times W}$ is associated with exposure time t_k. In the standard dynamic HDR setting considered here, the goal is to recover an HDR image $\hat{H}$ that is aligned to the middle exposure $L^{(2)}$, matches the ground-truth HDR image H_{gt}, and suppresses motion-induced artifacts. We adopt the middle exposure as the reference for three related reasons. First, this choice follows the common benchmark protocol in dynamic multi-exposure HDR reconstruction, where the ground truth and evaluation target are typically aligned to the middle exposure. Second, in a three-exposure bracket, the middle exposure often provides a practical compromise between the other two inputs: it usually retains more shadow information than the short exposure while suffering less saturation and motion blur than the long exposure, making it a relatively stable anchor for geometry and luminance. Third, the present framework is organized around reference-centered reconstruction, in which the base branch maintains global consistency with respect to $L^{(2)}$ and the other exposures mainly provide complementary radiometric and local-detail evidence. This assumption is still conditional rather than universal. It becomes weaker when the middle exposure is itself unreliable because of strong local saturation, severe under-exposure, motion blur, rolling-shutter distortion, or severe parallax. We, therefore, treat it as a standard but bounded design assumption and revisit its failure cases in Section 4.6 and its broader limitations in Section 4.7.

To account for exposure variation, each LDR frame is paired with a linearized HDR proxy and then concatenated to form the network input. Specifically, for each exposure we construct a packed tensor

$$F^{(k)} = \mathrm{Concat}\left(\tilde{H}^{(k)}, L^{(k)}\right) \in \mathbb{R}^{6\times H\times W}, \tag{1}$$

and the final input is

$$X = \mathrm{Concat}\left(F^{(1)}, F^{(2)}, F^{(3)}\right) \in \mathbb{R}^{18\times H\times W}, \tag{2}$$

or $X \in \mathbb{R}^{B\times 18\times H\times W}$ for a mini-batch of size B. The detailed radiometric formulation of $\tilde{H}^{(k)}$ is given in Section 3.3. The network predicts a normalized HDR image $\hat{H}$. A sigmoid output layer constrains $\hat{H} \in [0, 1]$ for stable training and evaluation.

3.2. Overview

Figure 1 illustrates the proposed dual-branch Transformer-CNN framework for HDR reconstruction. The network contains three main parts: a Transformer-based base branch for global structure and luminance modeling, a CNN-based detail branch for reliability-aware detail refinement, and the proposed Bidirectional Cross-Branch Fusion (BCBF) module as the central interaction mechanism between them.

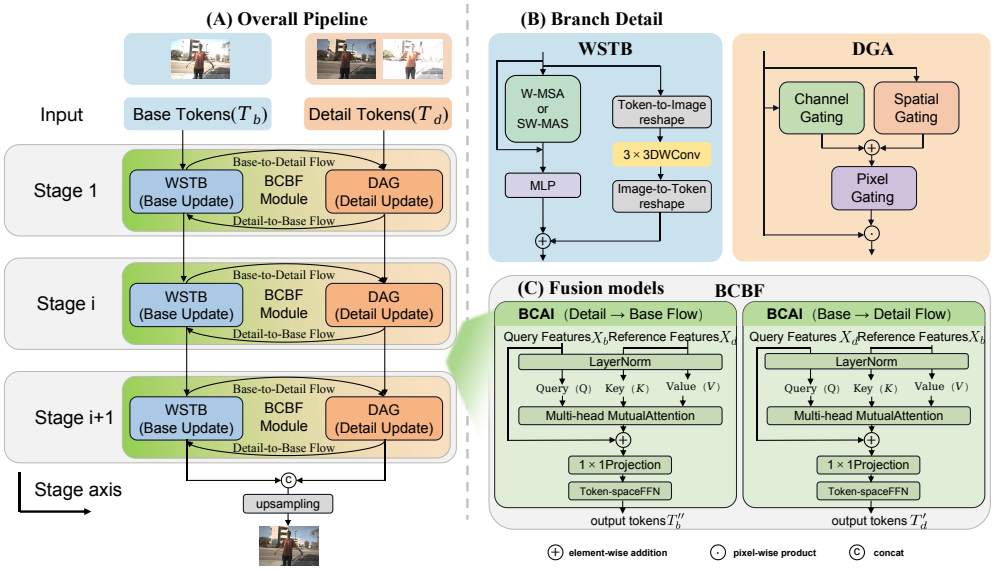


Figure 1. Overall pipeline of the proposed method. The base branch restores global structure with window-based self-attention, the detail branch enhances local textures via Detail-Aware Gating (DAG), and Bidirectional Cross-Branch Fusion (BCBF) exchanges complementary information between the two branches.

Given the packed input tensor X, we first extract shallow exposure-aware features and form an initial representation that preserves both radiometric and local texture cues. This representation is then routed into two processing streams with explicit role separation. The base branch, built on WRTBs, focuses on globally consistent structure and illumination and uses the middle exposure as the primary geometric reference. In parallel, the detail branch focuses on local textures and artifact-sensitive local refinement, with unreliable responses selectively suppressed before they are exchanged with the base branch.

This branch division is motivated by the observation that dynamic HDR reconstruction involves two different, but coupled, subproblems. The first is to maintain globally consistent luminance and scene structure across exposures, which requires long-range context aggregation and stable reference-centered reasoning. The second is to restore local textures and high-frequency transitions, which are more easily corrupted by saturation, under-exposure, motion, and local misalignment. A single shared representation must handle both objectives simultaneously, which can blur the distinction between stable global evidence and locally unreliable detail responses. We, therefore, separate the network into a base branch and a detail branch. The base branch focuses on global luminance and structural consistency, whereas the detail branch focuses on local-detail refinement under reliability control. Under this formulation, DAG improves the quality of local-detail cues before interaction, and BCBF enables the two branch-specialized representations to refine each other without collapsing their distinct roles.

The two branches are updated in parallel and interact at each stage through BCBF. This stage-wise collaboration is the core mechanism of the framework: stable global features from the base branch guide detail recovery, while reliability-filtered detail responses from the detail branch supplement the base representation. After the last interaction stage, the

refined features are aggregated by the reconstruction head to produce the final HDR prediction $\hat{H}$. During training, branch-specific predictions can also be supervised to stabilize the intended base and detail learning objectives.

From a functional perspective, deghosting and detail preservation are achieved through the cooperation of the branch-specialized modules rather than by a single component. The base branch provides stable global luminance and structural guidance, which helps prevent the reconstruction from being dominated by locally corrupted observations. The detail branch focuses on high-frequency textures and edge transitions, but its responses are first filtered by DAG so that unreliable local activations caused by saturation, underexposure, or motion inconsistency are suppressed before later fusion. BCBF then couples the two branches in both directions: base-to-detail interaction injects stable structural context into local refinement, while detail-to-base interaction transfers reliability-filtered local cues back to the global stream. In this way, deghosting is promoted by stable global guidance and the suppression of contaminated local responses, whereas detail preservation is supported by selective local refinement under cross-branch structural constraint.

3.3. Input Representation

We next specify the radiometric encoding used in the network. The design goal is not merely to increase the input dimensionality, but to preserve two complementary forms of evidence that are both necessary for dynamic HDR reconstruction: exposure-normalized radiometric consistency and texture-friendly local appearance. The purpose of this representation is, therefore, to preserve both exposure-aware linear information and image-domain cues so that the later base and detail branches can operate on complementary evidence from the same exposure bracket. We use channel-first notation throughout the paper: each normalized LDR frame is denoted by $L^{(k)} \in \mathbb{R}^{3\times H\times W}$. For 16-bit TIFF inputs, values are scaled by 2^{16} and clamped to $[0, 1]$. The exposure time of the k-th frame is denoted by $t_k > 0$. In Kalantari17 [6], the exposure file stores log2-style values e_k, and we convert them to exposure scales via $t_k = 2^{e_k}$.

Linearization. We apply gamma correction and exposure normalization to convert each LDR input into an approximate linear-domain representation:

$$\tilde{H}^{(k)} = \frac{\left(L^{(k)}\right)^{\gamma}}{t_k + \varepsilon},$$

where γ is the gamma parameter (we use $\gamma = 2.2$) and ε is a small constant for numerical stability (we use $\varepsilon = 10^{-8}$, consistent with the implementation).

Channel packing. For each exposure, we concatenate the linear proxy $\tilde{H}^{(k)}$ and the original gamma-domain LDR $L^{(k)}$ along the channel dimension:

$$F^{(k)} = \mathrm{Concat}\left(\tilde{H}^{(k)}, L^{(k)}\right) \in \mathbb{R}^{6\times H\times W}.$$

The final network input is formed by stacking the three exposures as

$$X = \mathrm{Concat}\left(F^{(1)}, F^{(2)}, F^{(3)}\right) \in \mathbb{R}^{18\times H\times W},$$

and the batched input is fed into the network in tensor layout $X \in \mathbb{R}^{B\times 18\times H\times W}$. This 18-channel design is structured rather than redundant: each exposure contributes six channels, namely, a three-channel linearized HDR proxy and a three-channel original LDR image. The resulting packed representation allows the network to access both exposure-aware radiometric evidence and texture-preserving image-domain cues for every input frame. Intuitively, $\tilde{H}^{(k)}$ provides exposure-aware linear cues that are useful for global structure

and luminance modeling, whereas $L^{(k)}$ retains texture-friendly local contrast and noise characteristics that are useful for detail recovery. This packed representation, therefore, supports the later role separation in a natural way. The base branch can emphasize exposure-normalized radiometric evidence for luminance and structural reasoning, whereas the detail branch can focus more directly on local contrast, edges, and texture statistics retained in the original LDR channels. At the same time, stacking the under-exposed, middle-exposed, and over-exposed observations enables the network to compare complementary evidence across exposures, so that missing shadow details, saturated highlight regions, and reference-centered structures can be handled within a unified input space.

3.4. Transformer Base Branch

The Transformer base branch serves as the global-context branch of the framework. It is intended to model the component of HDR reconstruction that is relatively stable at the scene level, namely, global luminance relationships, long-range structural dependencies, and reference-centered layout consistency. These properties are difficult to preserve with purely local operators when exposure variation and motion affect large image regions. In later cross-branch interaction, the base branch, therefore, provides stable contextual guidance rather than direct local-detail enhancement. To maintain a stable reference for geometry and luminance, it primarily processes the **middle-exposure embedding**.

3.4.1. Window-Based Residual Transformer Block (WRTB)

Each stage of the base branch employs a **Window-based Residual Transformer Block (WRTB)**. This block combines alternating regular and shifted window-based self-attention with an image-domain residual refinement path. Its purpose is twofold: to expand the receptive field and to preserve globally coherent structure while compensating for local discontinuities caused by window partitioning. Formally, let $\mathbf{T}_b \in \mathbb{R}^{H\times W\times C}$ denote the feature tokens of the base branch. The WRTB updates these tokens through a series of window-based self-attention layers followed by residual image-domain refinement:

$$\mathbf{T}_b^{(i+1)} = \mathbf{T}_b^{(i)} + \Phi_i(\mathbf{T}_b^{(i)}) \tag{3}$$

where Φ_i represents the composite operation of alternating window-based attention and a residual 3×3 depth-wise convolutional refinement in the image domain.

Figure 2 illustrates the WRTB design used in the base branch.

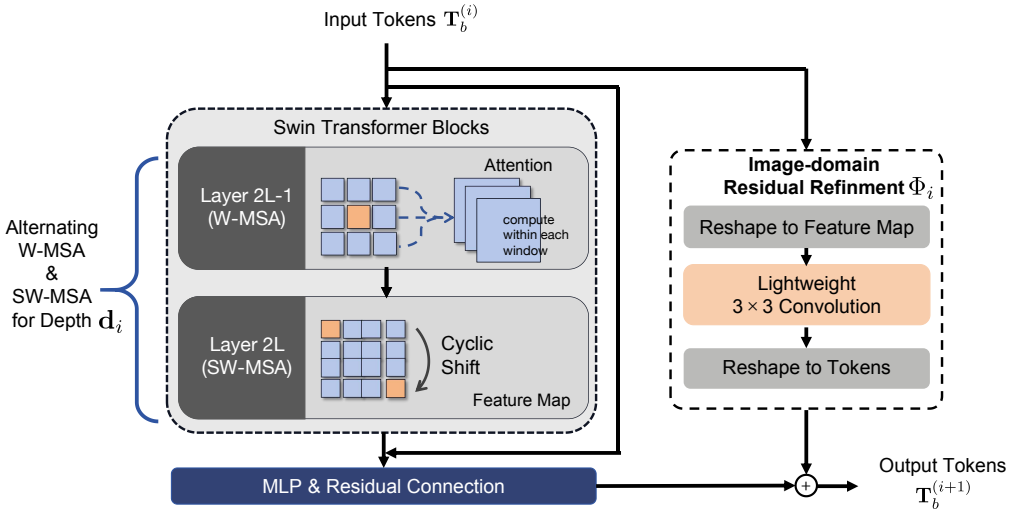


Figure 2. Illustration of the Window-based Residual Transformer Block (WRTB) in the base branch. Alternating W-MSA and SW-MSA expands cross-window interactions, while a residual 3×3 depth-wise convolution improves spatial consistency.

3.4.2. Window-Based Self-Attention Mechanisms

Using only fixed non-overlapping windows would keep the computation efficient. It would also restrict information exchange to within each local window. In dynamic HDR reconstruction, this is undesirable because global structure, luminance continuity, and motion-related context often span across window boundaries. To alleviate this limitation while keeping the computation tractable, the WRTB alternates between **Regular Window-based Multi-Head Self-Attention (W-MSA)** and **Shifted Window-based Multi-Head Self-Attention (SW-MSA)**:

- **W-MSA:** The feature map is partitioned into non-overlapping windows of size $M \times M$. For the feature representation within each window $\mathbf{X}_w \in \mathbb{R}^{M^2 \times C}$, the attention is computed as follows:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}} + \mathbf{B}\right)\mathbf{V} \tag{4}$$

  where $\mathbf{B}$ is the learnable relative position bias. This local constraint reduces the complexity from quadratic $O((HW)^2)$ to linear $O(HW \cdot M^2)$ relative to the image resolution.
- **SW-MSA:** To introduce cross-window connections, a cyclic shift of $\lfloor M/2 \rfloor$ pixels is applied before partitioning. A masking mechanism $\mathcal{M}$ is then used during the softmax operation so that self-attention is only calculated within sub-regions belonging to the same original spatial context.

This shifted-window self-attention scheme follows the Swin Transformer design [27]. It is also inspired by efficient hierarchical Transformer variants for super-resolution such as HiT-SR [28]. In our framework, alternating W-MSA and SW-MSA preserves computational efficiency while progressively establishing interactions across neighboring windows. This is important for maintaining globally coherent structure and luminance relationships under dynamic exposure variation.

The detailed computational procedure for a single WRTB stage is summarized in Algorithm 1.

Algorithm 1: Forward pass of WRTB stage

Input: Input tokens $\mathbf{T} \in \mathbb{R}^{H \times W \times C}$, Window size M, Depth D
Output: Updated tokens $\mathbf{T}_{out}$

for $l \leftarrow 1$ **to** D **do**
  $\mathbf{T}_{norm} \leftarrow \text{LayerNorm}(\mathbf{T})$;
  if l *is odd* **then**
    $\mathbf{T} \leftarrow \mathbf{T} + \text{W-MSA}(\mathbf{T}_{norm})$;
  else
    $\mathbf{T} \leftarrow \mathbf{T} + \text{SW-MSA}(\mathbf{T}_{norm})$;
  $\mathbf{T} \leftarrow \mathbf{T} + \text{MLP}(\text{LayerNorm}(\mathbf{T}))$;

$\mathbf{X} \leftarrow \text{ReshapeToMap}(\mathbf{T})$ $\mathbf{X}_{refined} \leftarrow \text{DWConv}_{3\times3}(\mathbf{X})$;
$\mathbf{T}_{out} \leftarrow \mathbf{T} + \text{ReshapeToTokens}(\mathbf{X}_{refined})$;
return $\mathbf{T}_{out}$;

3.4.3. Residual Refinement and Implementation

Following the attention layers, the base branch incorporates an **Image-domain Residual Refinement** path inside each WRTB. The tokens are reshaped back to the spatial domain

and processed by a residual 3×3 depth-wise convolution. No additional normalization is used in this path. The goal is to compensate for the local continuity loss that may arise from window-based partitioning. This refinement stabilizes the image-domain representation after global interaction rather than introducing another heavy transformation block. For inputs of arbitrary resolution, we use reflection padding to ensure compatibility with the fixed window size M.

3.5. CNN Detail Branch

In contrast to the base branch, the CNN detail branch is designed for the component of reconstruction that is locally variant and degradation-sensitive. Fine textures, edges, and small structural transitions are more vulnerable to clipping, under-exposure, motion corruption, and local inconsistency. The detail branch, therefore, focuses on selective recovery of high-frequency information under spatially heterogeneous reliability. Some regions still contain recoverable textures, whereas others are degraded by clipping, under-exposure, or motion corruption. Detail enhancement, therefore, cannot be applied indiscriminately. It must be coupled with suppression of contaminated responses before they participate in cross-branch collaboration. The key component of this branch is the **Detail-Aware Gating (DAG)** module, which acts as a branch-specialized refinement unit. Rather than introducing a fundamentally new attention primitive, DAG organizes established spatial, channel, and pixel gating operations into a soft reliability estimator for detail features [29,30]. DAG should, therefore, be understood as a feature-domain soft reliability estimator rather than as an explicit physical degradation detector or a hard masking module. Its purpose is not to classify local regions into a binary reliable/unreliable partition, but to continuously re-weight detail responses according to their estimated usefulness for later reconstruction and cross-branch interaction.

The gating mechanism inside DAG is inspired by DEA-Net for single-image dehazing [30]. The analogy is not tied to image formation, but to feature reliability. In multi-exposure HDR reconstruction, extreme illumination causes clipping in saturated highlights and severe noise in under-exposed shadows, as discussed in Section 3.3. These degraded regions tend to exhibit compressed gradients, reduced local variance, and weaker inter-channel discrimination. Their activations are then dominated by nuisance components such as clipping plateaus or sensor noise rather than recoverable textures. We, therefore, use DAG as a generic soft reliability estimator for the detail branch: it preserves informative detail responses and down-weights unreliable evidence before later BCBF-based interaction. Viewed from the standard middle-exposure reference setting, DAG helps integrate complementary cues from the under-exposed and over-exposed inputs and transfer only the reliable detail information to the reference-centered reconstruction pathway, rather than treating all exposure-specific responses as equally trustworthy.

Figure 3 illustrates the DAG design in the detail branch.

The operation begins by reshaping the detail tokens $T_d^{(i)} \in \mathbb{R}^{B\times N\times C}$ into a standard feature map $D^{(i)} \in \mathbb{R}^{B\times C\times H\times W}$ without loss of information. The DAG then computes a content-adaptive gate by aggregating complementary reliability cues. Specifically, the spatial gate A_s is extracted from pooled channel statistics:

$$A_s = \mathrm{Conv}_{7\times 7}\Big([\mathrm{AvgPool}_c(D), \mathrm{MaxPool}_c(D)]\Big). \tag{5}$$

The channel gate A_c is estimated through a squeeze-and-excitation style bottleneck [31]:

$$A_c = \mathrm{MLP}\big(\mathrm{GAP}(D)\big). \tag{6}$$

Here, $A_s \in \mathbb{R}^{B\times 1\times H\times W}$ and $A_c \in \mathbb{R}^{B\times C\times 1\times 1}$, where GAP denotes global average pooling and $[\cdot,\cdot]$ denotes channel-wise concatenation. After broadcasting them to a common shape, we obtain the coarse guidance map

$$P_1 = A_s + A_c. \tag{7}$$

To achieve fine-grained selection, the final pixel gate P_2 is computed by concatenating P_1 with the original input D and applying an efficient depth-wise convolution (DWConv) followed by a sigmoid activation σ, in the spirit of pixel-wise attention modulation [32]:

$$P_2 = \sigma\Big(\mathrm{DWConv}_{7\times 7}([D, P_1])\Big). \tag{8}$$

These three components play complementary roles in reliability estimation. The spatial gate estimates where local evidence is likely to be reliable, the channel gate estimates which feature responses are globally informative, and the pixel gate performs fine-grained reliability modulation conditioned on both. The gated output $\hat{D}^{(i)} = D^{(i)} \odot P_2$ can, therefore, be viewed as a reliability-weighted detail representation. Informative high-frequency structures receive larger responses, whereas features associated with clipping, low-SNR darkness, or local inconsistency are attenuated. Placing this gating unit right before token re-embedding ensures that later cross-branch interaction receives filtered local evidence rather than indiscriminately enhanced features.

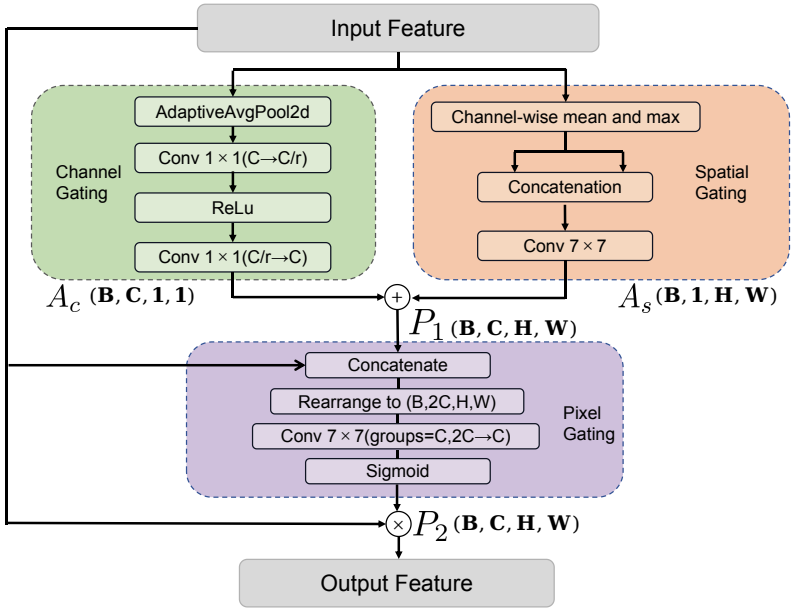


Figure 3. Detail-Aware Gating (DAG) module in the detail branch. The gate aggregates channel attention, spatial attention, and pixel attention cues, and then adaptively modulates the input feature map for robust detail refinement.

3.6. Bidirectional Cross-Branch Fusion (BCBF)

The base and detail branches must exchange information during feature extraction rather than only at the reconstruction head. We introduce **Bidirectional Cross-Branch Fusion (BCBF)** for this purpose. BCBF exchanges complementary feature types stage by stage: the base branch provides stable structural context, while the detail branch provides reliability-filtered detail refinement cues. At the core of BCBF is the **Bidirectional Cross-Attention Interaction (BCAI)** module, which manages reciprocal information exchange between the two branches. Figure 4 illustrates the overall architecture of the proposed BCBF module and its bidirectional interaction pathway.

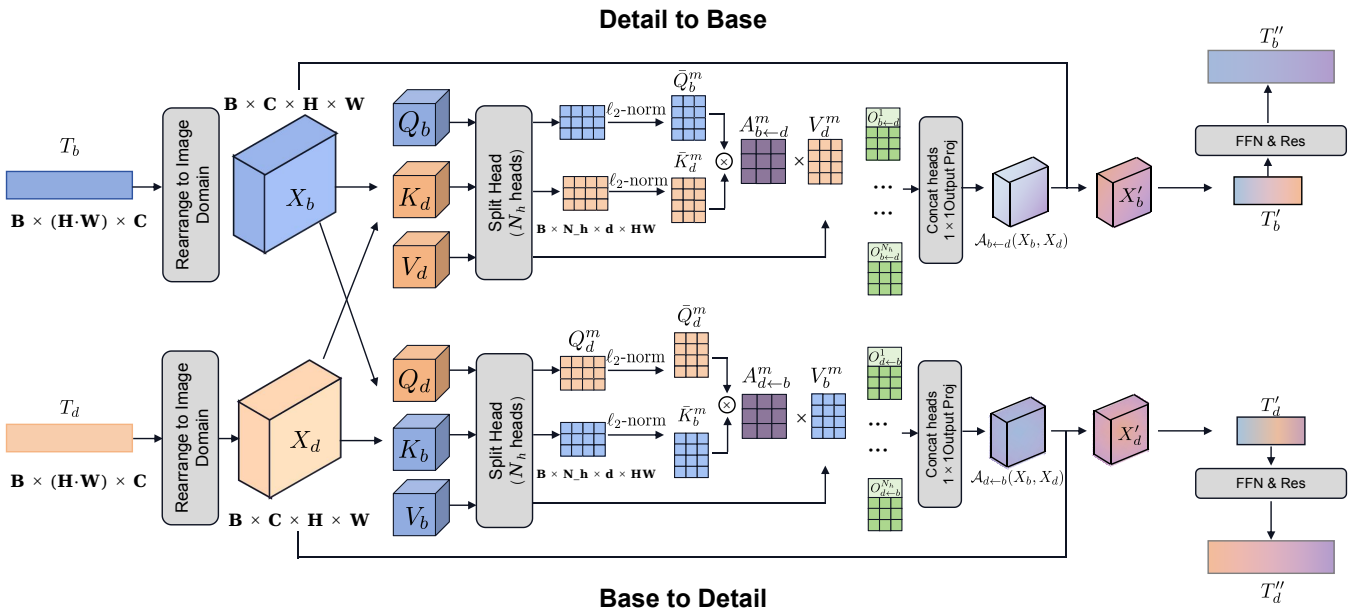


Figure 4. Architecture of the proposed Bidirectional Cross-Branch Fusion (BCBF) module. The diagram illustrates the bidirectional exchange between the base and detail branches, including the two directional updates, channel-subspace mutual attention, residual fusion, and subsequent FFN refinement.

Given the base tokens $T_b^{(i)}$ and detail tokens $T_d^{(i)}$ at stage i, we first reshape them into image-domain features $X_b, X_d \in \mathbb{R}^{B\times C\times H\times W}$. For the directional update $X_d \to X_b$, BCAI applies layer normalization and three 1×1 projections. The query is generated from X_b, and the key-value tensors are generated from X_d, following the standard query-key-value attention formulation [33]. With N_h attention heads and per-head channel dimension $d = C/N_h$, the projected features for head m are reshaped as

$$Q_b^m,\ K_d^m,\ V_d^m \in \mathbb{R}^{d\times HW}.$$

The mutual attention matrix is then defined as

$$A_{b\leftarrow d}^m = \mathrm{Softmax}\Big(\tau_m\, \bar{Q}_b^m (\bar{K}_d^m)^\top\Big) \in \mathbb{R}^{d\times d}, \tag{9}$$

where $\bar{Q}$ and $\bar{K}$ denote ℓ_2-normalized features along the spatial dimension and τ_m is a learnable temperature for head m. The corresponding fused response is

$$O_{b\leftarrow d}^m = A_{b\leftarrow d}^m V_d^m. \tag{10}$$

The outputs of all heads are concatenated and projected by a 1×1 mapping W_o to form the cross-branch residual

$$\mathcal{A}_{b\leftarrow d}(X_b, X_d) = W_o\Big[\mathrm{Concat}\left(\{O_{b\leftarrow d}^m\}_{m=1}^{N_h}\right)\Big]. \tag{11}$$

The directional image-domain update is written as

$$X_b' = X_b + \mathcal{A}_{b\leftarrow d}\big(\mathrm{LN}(X_b), \mathrm{LN}(X_d)\big). \tag{12}$$

Applying the same operation in the reverse direction yields

$$X_d' = X_d + \mathcal{A}_{d\leftarrow b}\big(\mathrm{LN}(X_d), \mathrm{LN}(X_b)\big), \tag{13}$$

and this bidirectional formulation motivates the term mutual attention. Let $\Pi(\cdot)$ denote the image-to-token projection. After mapping the updated image-domain features back to tokens,

$$T_b' = \Pi(X_b'), \qquad T_d' = \Pi(X_d'), \tag{14}$$

each branch is further refined by a feed-forward network (FFN):

$$T_b'' = T_b' + \mathrm{FFN}\big(\mathrm{LN}(T_b')\big), \qquad T_d'' = T_d' + \mathrm{FFN}\big(\mathrm{LN}(T_d')\big). \tag{15}$$

Intuitively, the directional update $X_d \to X_b$ allows local-detail features to supplement the globally oriented base representation. The reverse update $X_b \to X_d$ injects structurally stable context into the detail branch. The interaction is bidirectional because the two branches are role-specialized but functionally interdependent. The base branch benefits from localized detail cues, and the detail branch benefits from stable global guidance. As shown later in the ablation study, removing cross-branch interaction causes a substantial performance drop, and restricting the interaction to only one direction also yields weaker results than the full bidirectional design.

This formulation differs from standard cross-attention in two important respects. In conventional cross-attention, the query features from one stream interact with the key-value features from another stream through pair-wise similarity over spatial tokens, so that each token is updated by aggregating information from matched positions in the other stream. First, standard cross-attention typically constructs an $HW \times HW$ affinity matrix to model token-to-token correspondence [33], much like dense pair-wise interactions in non-local matching [34]. In multi-exposure HDR reconstruction, such dense spatial matching can be unstable under saturation, under-exposure, and local motion, where one-to-one correspondence is often ambiguous or unreliable. In contrast, our mutual attention normalizes each head along the spatial dimension and computes a $d \times d$ affinity in channel subspace. The design, therefore, emphasizes the transfer of complementary feature types between branches rather than explicit token-level matching.

Second, with fixed channel width, the dominant cost of BCBF scales approximately linearly with HW, whereas standard token-based cross-attention incurs quadratic spatial complexity. This makes BCBF better suited to high-resolution HDR fusion when the goal is cross-branch collaboration rather than explicit geometric alignment. The trade-off is that the proposed interaction does not model direct spatial correspondence and may, therefore, be less effective under large non-rigid motion, severe parallax, or rolling-shutter distortions.

Table 1 summarizes the practical distinction. Conventional cross-attention is more natural when feature interaction can be formulated as reliable spatial matching. By contrast, BCBF is designed for the present branch-specialized HDR setting, where the key requirement is to exchange reliable global and local evidence without forcing dense token-level alignment in regions corrupted by saturation, darkness, or motion. As shown later in the fusion replacement ablation, this task-level distinction is not only conceptual but also reflected in performance: replacing BCBF with standard cross-attention leads to weaker reconstruction quality under the same branch framework.

Under this design, BCBF serves as the mechanism that couples stable deghosting-oriented guidance from the base branch with reliability-filtered detail cues from the local branch.

The BCBF module is inserted after each stage-wise update, following the WRTB in the base branch and the DAG-based update in the detail branch. This iterative fusion process allows the network to progressively calibrate global illumination and local textures in a tightly coupled manner rather than deferring branch interaction to the end of the network.

Relative to prior efficient channel-domain attention designs in image restoration [35], BCBF differs from the closest prior designs in three ways: (1) it is used for cross-branch interaction between the base and detail streams rather than for feature mixing within a single branch; (2) it is applied in a bidirectional and stage-wise manner, so the two branch-specialized representations refine each other progressively throughout the network; and (3) it operates on heterogeneous branches, where the detail branch is first filtered by DAG so that the exchanged information is more reliable under saturation, under-exposure, and motion corruption. This branch-specialized decomposition perspective is also related to neighboring dual-branch fusion designs explored in adjacent image-fusion tasks [36]. BCBF should, therefore, be understood as an HDR-oriented branch-collaboration design built on efficient channel-subspace interaction, rather than as a fundamentally new attention primitive.

Table 1. Compact comparison between conventional cross-attention and the proposed BCBF interaction.

Aspect	Conventional Cross-Attention	BCBF
Affinity domain	Spatial token affinity over $HW \times HW$ relations	Channel-subspace affinity over $d \times d$ relations
Main goal	Token-level correspondence and feature aggregation across positions	Bidirectional exchange of complementary feature types between the base and detail branches
Suitable scenario	Relatively reliable spatial matching across inputs or feature streams	Unreliable local observations with complementary branch features under motion and exposure corruption
Computational behavior	Quadratic in the number of spatial tokens	Approximately linear in spatial resolution for fixed channel width
Main limitation	Can be confused by saturation, under-exposure, occlusion, or motion when token correspondence is ambiguous	Does not explicitly model geometric correspondence under severe deformation

3.7. Loss Function

The proposed network is supervised by a composite objective function that targets the final HDR reconstruction $\hat{H}$ together with auxiliary branch outputs, specifically the base-branch prediction $\hat{H}_b$ and the detail-branch prediction $\hat{H}_d$. The goal of this design is to align supervision with branch roles rather than to treat all outputs identically. The final prediction is optimized for overall HDR reconstruction quality. The base branch is regularized toward stable global structure learning. The detail branch is encouraged to preserve local texture gradients.

To align the linear HDR data with human visual perception and ensure numerical stability during optimization, all predictions and the ground-truth image H_{gt} are mapped to the tonemapped domain using the μ-law function $T(\cdot)$:

$$T(Z) = \frac{\log(1+\mu Z)}{\log(1+\mu)}, \quad \text{with } \mu = 5000. \tag{16}$$

Here, Z denotes any HDR image in the linear domain, including $\hat{H}$, $\hat{H}_b$, $\hat{H}_d$, or H_{gt}.

In our default training setting, the primary reconstruction loss on the final output is a μ-law ℓ_1 term:

$$\mathcal{L}_r = \left\| T(\hat{H}) - T(H_{gt}) \right\|_1. \tag{17}$$

This term serves as the main supervision signal for the overall HDR reconstruction. To stabilize the learning of global structure and consistent luminance in the Transformer-based base branch, we further introduce an auxiliary base reconstruction loss $\mathcal{L}_{\text{base}}$:

$$\mathcal{L}_{\text{base}} = \left\| T(\hat{H}_b) - T(H_{gt}) \right\|_1, \tag{18}$$

which penalizes deviations in the tonemapped output of the base branch from the target image. To encourage the detail branch to recover local texture structure more faithfully, we also impose a detail-aware loss $\mathcal{L}_{\text{detail}}$ on $\hat{H}_d$ by penalizing the difference in gradient magnitudes:

$$\mathcal{L}_{\text{detail}}(\hat{H}_d, H_{gt}) = \left\| G(\hat{H}_d) - G(H_{gt}) \right\|_1, \tag{19}$$

where $G(X) = \sqrt{\nabla_x(X)^2 + \nabla_y(X)^2 + \epsilon}$ denotes the Sobel-based gradient magnitude operator. The overall training objective is subsequently formulated as a weighted combination of these terms:

$$\mathcal{L} = \mathcal{L}_r + \lambda_{\text{base}} \mathcal{L}_{\text{base}} + \lambda_{\text{detail}} \mathcal{L}_{\text{detail}}, \tag{20}$$

where λ_{base} and λ_{detail} are hyperparameters controlling the influence of the auxiliary base and detail-aware supervision, respectively. Both terms are treated as auxiliary regularizers rather than primary objectives, so their weights are kept modest to avoid overwhelming the main reconstruction loss. In the default configuration used in this paper, we set $\lambda_{\text{base}} = 0.1$ and $\lambda_{\text{detail}} = 0.1$, which matches the implemented `DualBranchLoss`. An auxiliary-weight sensitivity study is provided in Appendix A.1 (Table A1).

3.8. Complexity Analysis

Complexity overview. We report the complexity of the proposed model to make the architectural trade-offs explicit rather than to claim minimum latency. The purpose of this section is to explain where the cost comes from in the branch-specialized design; the empirical accuracy–cost comparison is deferred to Section 4.4. Although training uses `img_size` = 128, efficiency is reported under the patch-based inference setting for full-resolution testing, where each forward pass takes an input of size $1 \times 18 \times 256 \times 256$, yielding 1.33 M parameters and 22.13 GMACs. The computation is dominated by the window-based self-attention and MLPs in the base branch. The CNN-based DAG operators and the cross-branch interaction add comparatively moderate overhead.

Scaling with resolution. Assuming a window size of M and a feature dimension of C, the window-based attention mechanism incurs a per-layer complexity of approximately $O(HW \cdot M^2 \cdot C)$ (ignoring constants and head partitioning). For a fixed M, this scales linearly with the spatial resolution. By contrast, global attention scales as $O((HW)^2)$, which is impractical for high-resolution HDR images. The cross-branch fusion primarily uses 1×1 projections and head-wise mutual attention in the image domain, so its cost is also approximately linear in HW. The DAG module employs pooled attention cues and depth-wise convolution (Section 3.5), again yielding linear complexity in HW with a controllable bottleneck governed by the channel reduction ratio r.

Profiling perspective. The reported complexity should, therefore, be interpreted as a transparency measure for the structural cost of the model rather than as a standalone efficiency claim. The architecture also exposes clear design knobs, such as the channel reduction ratio r, for analyzing how branch choices affect both reconstruction behavior and computational cost.

4. Experiments

4.1. Datasets

We evaluate the proposed method on three multi-exposure HDR benchmarks: Kalantari17 [6] (74 training/15 test scenes), the Tel benchmark [4] (108 training/36 test scenes), and Challenge123 [7] (96 training/27 test scenes). These datasets serve different evaluation roles. Kalantari17 is the standard benchmark for dynamic HDR reconstruction. The Tel benchmark extends the evaluation to darker scenes and more diverse illumination. Challenge123 is the strictest setting, with stronger motion and more severe saturation. Across all datasets, we use the same preprocessing conventions for our method, including gamma encoding ($\gamma = 2.2$), to improve comparability across experiments.

4.2. Implementation Details

The proposed method is implemented in PyTorch 2.0.1+cu117 with TorchVision 0.15.2+cu117 based on the BasicSR 1.4.2 framework. Unless otherwise stated, the default model uses four stages per branch with $C = 60$, $N_h = 6$, window size $M = 8$, tokenization patch size 1, and DAG reduction ratio $r = 8$. Although training uses `img_size` = 128, efficiency is reported under the patch-based inference setting for full-resolution testing, where each forward pass takes an input of size $1 \times 18 \times 256 \times 256$, yielding 1.33 M parameters and 22.13 GMACs. For the later accuracy–complexity comparison, all methods are profiled under the same setting, namely, the same input size $1 \times 18 \times 256 \times 256$, the same NVIDIA A100 (80 GB) GPU, batch size 1, and the same timing protocol with CUDA synchronization.

Unless otherwise stated, we train a separate model for each benchmark on its official training split with the same architecture, loss, and optimization recipe. The reported cross-dataset results are, therefore, not obtained by zero-shot transfer of a single model. The default training schedule uses Adam with an initial learning rate of 1×10^{-4} for **800,000** iterations. During training, we use random flips and rotations. During validation, we use a 512×512 center crop for faster monitoring.

For full-resolution testing, each scene is partitioned into non-overlapping 256×256 patches, reflection-padded when necessary, and then stitched back to the original resolution without overlap. Unless otherwise stated, all reported results of our method are obtained under this unified inference and evaluation pipeline. All experiments are conducted on NVIDIA A100 (80 GB) GPUs, with an average training speed of approximately **0.715 s/iter** including data loading and backpropagation. Table 2 summarizes the main training and module hyperparameters used throughout the paper. For baseline methods, we follow a three-level protocol to make the result source explicit. If an official implementation and publicly released checkpoints are available, we reproduce the official pipeline and run inference using the released checkpoints. If no public checkpoints are available but the method can be reproduced reliably, we train the model following the original paper and then evaluate it under the same test pipeline. Only when a method cannot be reproduced reliably do we quote the numbers reported in the original paper, in which case the protocol differences should be interpreted accordingly.

Table 2. Main training and module hyperparameters used by the default model unless otherwise stated.

Parameter	Value
Framework	PyTorch 2.0.1+cu117 + TorchVision 0.15.2+cu117 + BasicSR 1.4.2
Input preprocessing	Three-exposure input with gamma encoding ($\gamma = 2.2$); each exposure contributes a linearized HDR proxy and the original LDR image
Stages per branch	4

Table 2. *Cont.*

Parameter	Value
Embedding dimension C	60
BCBF attention heads N_h	6
WRTB window size M	8
Tokenization patch size	1×1
DAG reduction ratio r	8
Training crop size	128×128
Inference patch size	256×256, non-overlapping stitching
Batch size	4 per GPU
Optimizer	Adam with $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay $= 0$
Initial learning rate	1×10^{-4}
Learning-rate schedule	MultiStepLR with milestones at 200 k, 400 k, and 600 k iterations; decay factor 0.5
Total training iterations	800,000
Auxiliary loss weights	$\lambda_{\text{base}} = 0.1$, $\lambda_{\text{detail}} = 0.1$
Data augmentation	Random horizontal/vertical flips and 90° rotations
Validation setting	512×512 center crop for faster monitoring
Hardware	NVIDIA A100 (80 GB) GPU

4.3. Evaluation Metrics

We report PSNR and SSIM [37,38] on the HDR outputs, together with their μ-law variants (μ-PSNR and μ-SSIM), to better reflect perceptual differences across a wide dynamic range [6]. Using the same μ-law compressor $T(\cdot)$ defined in Section 3.7, μ-PSNR and μ-SSIM are computed on $T(\hat{H})$ and $T(H_{gt})$ with $\mu = 5000$. Unless stated otherwise, we apply no border cropping during evaluation, and all metrics are computed on RGB channels rather than luminance only. We also report HDR-VDP2 [39] as a complementary perceptual quality indicator. This unified metric pipeline is used throughout our datasets and ablations for all methods evaluated by checkpoint inference or our own re-training/testing, while the small number of results quoted from original papers still reflect their respective evaluation protocols.

4.4. Comparison with State-of-the-Art

Full-resolution test evaluation. This comparison evaluates whether the proposed branch-specialized design remains effective across datasets with different motion and illumination characteristics. Kalantari17 [6] serves as the canonical benchmark, where the main question is whether the method remains competitive under the standard setting. The Tel benchmark tests darker scenes and more diverse illumination. Challenge123 is the strictest benchmark, with larger displacement and stronger saturation. For the main tables, each reported result of our method comes from the model trained on the corresponding dataset split rather than from zero-shot transfer of a single Kalantari17 model. We evaluate the full-resolution test sets with patch-based inference and stitching using 256×256 non-overlapping patches. Unless explicitly noted otherwise, our model follows the unified preprocessing and metric pipeline described above. Baseline results are obtained in the following order of preference: reproduced inference with publicly released checkpoints, our own re-training and testing when reliable reproduction is possible, and quotation from the original paper only when reproduction is not reliable.

Table 3 shows that the proposed method remains competitive on the standard Kalantari17 benchmark [6] benchmark. Although it does not rank first on every distortion-oriented metric, it achieves the best HDR-VDP2 (66.86) and ties the best μ-SSIM (0.9919) while maintaining competitive PSNR, μ-PSNR, and SSIM. These results indicate that the proposed dual-branch design improves structural and perceptual consistency without sacrificing reconstruction fidelity. This behavior is consistent with the complementary roles of

the base branch for global luminance and structure modeling and the detail branch for local texture recovery.

Table 3. Main comparison on Kalantari17. Bold values indicate the best result in each column, and underlined values indicate the second-best.

Method	PSNR	μ-PSNR	SSIM	μ-SSIM	HDR-VDP2
DHDR [15]	41.38	43.30	0.9862	0.9898	60.50
AHDR [1]	41.74	43.96	0.9872	0.9912	62.30
NHDRR [2]	41.42	42.41	0.9861	0.9887	61.21
ADNet [16]	41.88	44.37	0.9892	<u>0.9917</u>	66.02
APNT [40]	41.61	43.94	0.9879	0.9898	64.05
HyHDR [5]	42.47	<u>44.64</u>	0.9894	0.9915	66.05
HDR-GAN [22]	41.57	43.92	0.9865	0.9905	65.45
DiffHDR [41]	41.46	44.23	0.9885	0.9911	65.52
HDR-Transformer [3]	42.18	44.32	0.9884	0.9916	66.03
SCTNet [4]	42.27	44.38	0.9890	0.9914	66.08
SAFNet [7]	**43.18**	**44.66**	**0.9901**	**0.9919**	<u>66.73</u>
Ours	<u>42.65</u>	<u>44.58</u>	<u>0.9898</u>	**0.9919**	**66.86**

Figure 5 is consistent with the quantitative trend. Our method better preserves boundary continuity and local luminance transitions in motion-sensitive regions, with more confined signed deviations from the GT.

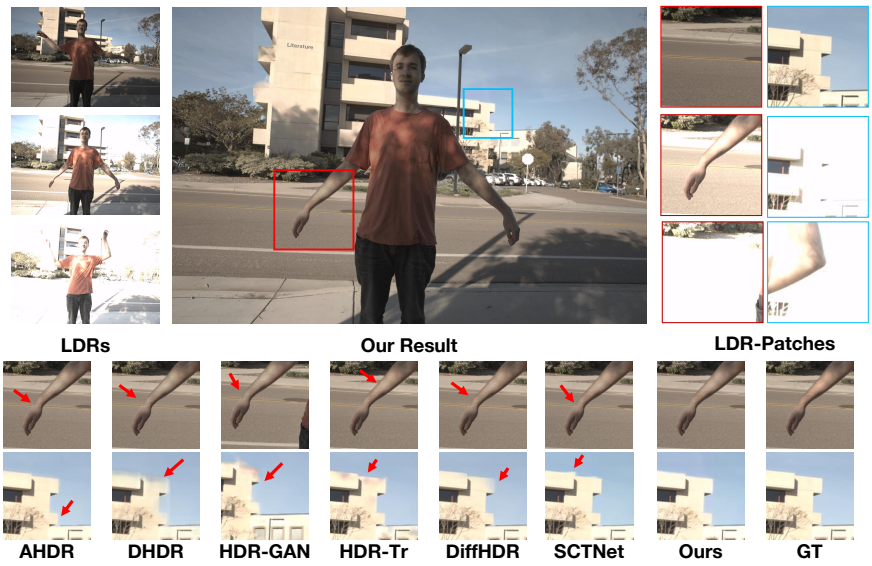


Figure 5. Qualitative comparison on Kalantari17 (all HDR results are tonemapped for visualization). The red and blue boxes indicate the two regions enlarged in the zoomed-in patches, corresponding to two different LDR patch groups, and the arrows highlight motion-sensitive boundaries and thin structures that are prone to ghosting. This example represents a standard dynamic HDR scene, where exposure differences and object motion coexist around high-contrast edges. Difference maps between each reconstructed result and the GT image are also shown; warmer colors indicate positive differences with larger magnitude, cooler colors indicate negative differences with larger magnitude, and intermediate tones indicate relatively small differences. Compared with representative baselines, the proposed method suppresses ghosting more effectively while preserving structural continuity and local luminance transitions, yielding closer visual agreement with the ground truth.

On the Tel benchmark, Table 4 shows a clearer overall advantage. Our method achieves the best values on all five reported metrics. Relative to SCTNet, it improves PSNR by +0.32 dB, μ-PSNR by +0.76 dB, and HDR-VDP2 by +0.42. These gains suggest that

exposure-normalized base modeling and reliability-aware detail filtering are particularly effective for darker scenes and more diverse illumination, where preserving contrast transitions and suppressing local artifacts become more difficult.

Table 4. Main comparison on the Tel benchmark introduced by Tel et al. [4] (36 test scenes). Bold values indicate the best result in each column, and underlined values indicate the second-best.

Method	PSNR	μ-PSNR	SSIM	μ-SSIM	HDR-VDP2
DHDR [15]	45.13	41.82	0.9925	0.9829	67.09
AHDR [1]	46.60	<u>43.05</u>	0.9951	<u>0.9872</u>	68.80
NHDRR [2]	39.61	36.68	<u>0.9956</u>	0.9590	65.41
HDR-GAN [22]	44.87	41.71	0.9949	0.9832	69.57
HDR-Transformer [3]	46.35	42.39	0.9948	0.9844	69.23
DiffHDR [41]	45.63	42.18	0.9946	0.9841	69.88
SCTNet [4]	<u>47.51</u>	42.55	0.9952	0.9850	<u>70.66</u>
Ours	**47.83**	**43.31**	**0.9959**	**0.9876**	**71.08**

Figure 6 shows the same tendency under difficult illumination changes. The proposed method better preserves highlight structure and weak textures, and its signed difference maps exhibit narrower warm/cool deviation regions.

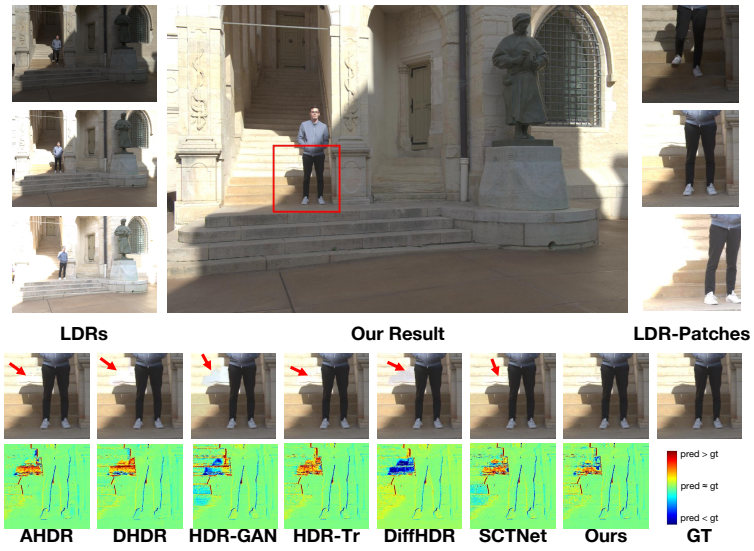


Figure 6. Qualitative comparison on Tel (all HDR results are tonemapped for visualization). The red box indicates the region enlarged in the zoomed-in patches, which highlights challenging areas under dark or spatially varying illumination, including weak-texture regions, contrast transitions, and local highlight-shadow boundaries. Difference maps between each reconstructed result and the GT image are also shown; warmer colors indicate positive differences with larger magnitude, cooler colors indicate negative differences with larger magnitude, and intermediate tones indicate relatively small differences. This example illustrates a difficult low-illumination case in which reliable recovery requires both robust global exposure reasoning and selective local-detail preservation. Compared with representative baselines, the proposed method preserves fine textures and contrast transitions more faithfully, yielding closer visual agreement with the ground truth.

Table 5 further highlights the robustness of the proposed method on Challenge123. Our method achieves the best value on every reported metric and improves PSNR by +0.54 dB over SAFNet. Because Challenge123 contains larger inter-frame motion and stronger saturation than Kalantari17, these results support the roles of DAG and BCBF under severe dynamic interference: DAG suppresses unreliable local responses, while BCBF promotes cross-branch interaction without enforcing dense spatial matching. The

wider margin, therefore, indicates improved structural coherence and artifact suppression in difficult scenes.

Table 5. Main comparison on the Challenge123 dataset [7] (27 test sequences). Bold values indicate the best result in each column, and underlined values indicate the second-best.

Method	PSNR	μ-PSNR	SSIM	μ-SSIM	HDR-VDP2
DHDR [15]	27.13	38.34	0.9680	0.9832	54.03
AHDR [1]	27.55	39.70	0.9669	0.9877	54.57
HDR-GAN [22]	27.86	39.25	0.9672	0.9887	54.61
HDR-Transformer [3]	28.82	40.86	0.9747	0.9897	54.63
DiffHDR [41]	28.67	40.36	0.9746	0.9883	54.60
SCTNet [4]	28.78	40.88	<u>0.9750</u>	<u>0.9901</u>	54.35
SAFNet [7]	<u>28.93</u>	<u>40.94</u>	0.9740	0.9885	<u>54.84</u>
Ours	**29.47**	**41.60**	**0.9774**	**0.9905**	**54.96**

Figure 7 further supports this conclusion in the most challenging setting. Our method better maintains edge continuity in severely displaced regions, while the signed difference maps remain more confined around the GT.

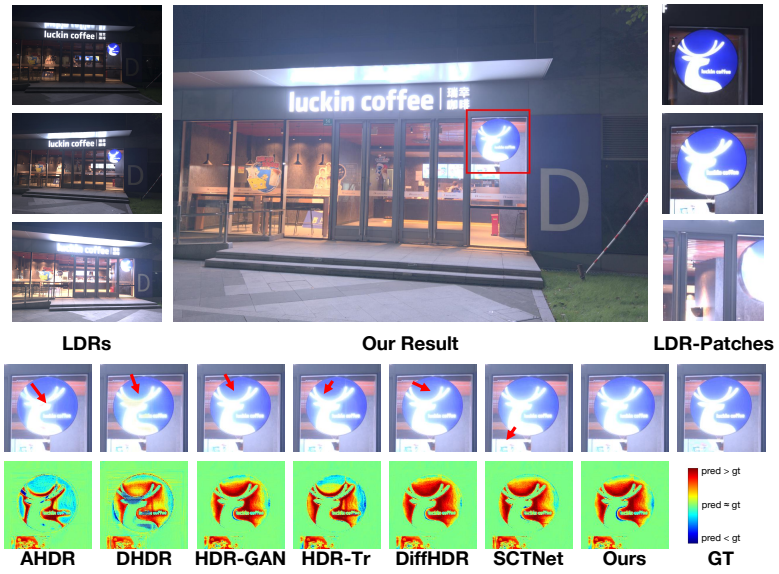


Figure 7. Qualitative comparison on Challenge123 (all HDR results are tonemapped for visualization). The red box indicates the region enlarged in the zoomed-in patches, which highlights severe motion, saturation, and displaced structural details that commonly cause ghosting and local distortion. Difference maps between each reconstructed result and the GT image are also shown; warmer colors indicate positive differences with larger magnitude, cooler colors indicate negative differences with larger magnitude, and intermediate tones indicate relatively small differences. This example is particularly challenging because cross-exposure correspondence is highly ambiguous in strongly degraded regions. Compared with representative baselines, the proposed method suppresses ghost artifacts and restores edge continuity more effectively, yielding closer visual agreement with the ground truth.

Table 6 provides the empirical accuracy–cost comparison of the proposed model. The architecture is not designed primarily for minimum latency. Its value lies in reconstruction quality under challenging dynamic HDR scenarios. Under the unified profiling setup described above, the model has 1.33 M parameters and 22.13 GMACs. Compared with SCTNet, it gains +0.38 dB PSNR and +0.78 HDR-VDP2 with only a moderate increase in parameters and nearly identical latency (449.7 ms vs. 453.0 ms). The extra cost of cross-branch collaboration is, therefore, accompanied by a measurable gain in reconstruction quality.

Table 6. Accuracy and complexity comparison on Kalantari17. The reported accuracy metrics are consistent with those detailed in Table 3. Higher values indicate better performance for PSNR, SSIM and HDR-VDP2, whereas lower values are preferred for Params./GMACs/Latency. Params./GMACs/Latency are profiled under the unified setting described in Section 4.2. Bold values indicate the best result in each accuracy column, and underlined values indicate the second-best.

Method	PSNR	μ-PSNR	SSIM	μ-SSIM	HDR-VDP2	Params. (M)	GMACs	Latency (ms)
AHDR [1]	41.74	43.96	0.9872	0.9912	62.30	1.52	17.2	81.2
HDR-GAN [22]	41.57	43.92	0.9865	0.9905	65.45	2.88	7.8	164.6
HDR-Transformer [3]	42.18	44.32	0.9884	<u>0.9916</u>	66.03	1.48	19.9	380.8
SCTNet [4]	<u>42.27</u>	<u>44.38</u>	<u>0.9890</u>	0.9914	<u>66.08</u>	0.99	16.4	453.0
Ours	**42.65**	**44.58**	**0.9898**	**0.9919**	**66.86**	1.33	22.13	449.7

4.5. Ablation Studies

To validate the design rationale of the proposed method, we organize the ablation analysis around concrete questions rather than around isolated modules:

- **Is a single representation stream sufficient?** We compare the full model against a single-branch baseline (A0 vs. A2) to test the value of explicit base-detail decomposition.
- **How should the detail branch be modeled?** We compare the full model against a dual-branch variant with a Transformer-based detail branch (A1 vs. A2) to evaluate whether DAG-style filtering is more appropriate for detail refinement.
- **Is cross-branch interaction necessary?** We compare the full model against the setting without cross-branch interaction (A2 vs. A3).
- **Is bidirectional interaction better than unidirectional interaction?** We compare the full model against the two unidirectional interaction variants (A2 vs. A4/A5).
- **Is BCBF more effective than simpler or more conventional fusion replacements?** We compare BCBF against identity mapping, concatenation followed by a 1×1 convolution, and a standard cross-attention block.
- **How sensitive is DAG to capacity and placement?** We further analyze the effect of the DAG reduction ratio and placement strategy through the r sweep and stage-placement study.

Core ablations (structure and cross-branch interaction).

The main quantitative results for branch structure and cross-branch interaction are summarized in Table 7.

Table 7. Core ablations on Kalantari17. Bold values indicate the best result in each column.

ID	Arch.	Detail	Interaction	Stages	r	PSNR	μ-PSNR	SSIM	μ-SSIM
A0	single	—	—	—	—	41.887	44.035	0.9892	0.9918
A1	dual	WRTB	bi	all	—	42.344	44.432	0.9885	0.9917
A2	dual	DAG	bi	all	8	**42.653**	**44.581**	**0.9898**	**0.9919**
A3	dual	DAG	none	all	8	38.699	43.334	0.9819	0.9906
A4	dual	DAG	uni (d→b)	all	8	42.308	44.479	0.9892	0.9918
A5	dual	DAG	uni (b→d)	all	8	42.114	44.580	0.9882	0.9917
A6	dual	DAG	bi	last-2	8	42.270	44.416	0.9888	0.9918

WRTB: Window-based Residual Transformer Block (used in the base branch for all dual settings). DAG: CNN-based detail-aware gating. r: `detail_cnn_reduction`.

DAG capacity and placement.

The sensitivity results for the detail branch are summarized in Table 8 for the reduction-ratio sweep and in Table 9 for the stage-placement study.

Table 8. Detail-branch capacity sweep via `detail_cnn_reduction` (full model uses $r = 8$). Bold values indicate the best result in each column.

Setting	PSNR	μ-PSNR	SSIM	μ-SSIM
$r = 4$	42.128	44.495	0.9887	0.9918
$r = 8$ (ours)	**42.653**	**44.581**	**0.9898**	**0.9919**
$r = 16$	42.486	44.440	0.9886	0.9917

Table 9. DAG placement study (full model uses DAG in all stages). Bold values indicate the best result in each column.

Setting	PSNR	μ-PSNR	SSIM	μ-SSIM
All stages (ours)	**42.653**	**44.581**	**0.9898**	**0.9919**
Last-2 stages	42.270	44.416	0.9888	0.9918
First-2 stages	41.981	44.344	0.9884	0.9917

Discussion. Relative to the corrected single-branch baseline (A0), the full model (A2) gains $+0.77$ dB in PSNR and $+0.55$ dB in μ-PSNR. This result supports explicit base-detail decomposition over a single shared stream.

Compared with the dual-branch Transformer-detail variant (A1), the full model gains $+0.31$ dB in PSNR and $+0.15$ dB in μ-PSNR. This result favors DAG-style detail refinement over using another Transformer-style detail stream.

We next analyze the necessity of cross-branch interaction. Removing interaction entirely (A3) reduces PSNR by 3.95 dB, which shows that the two branches are not merely additive. They require explicit collaboration. The unidirectional variants (A4 and A5) retain part of the gain, but neither matches the full bidirectional design (A2). Detail-to-base supplementation and base-to-detail guidance are, therefore, both useful.

The fusion replacement study in Table 10, together with the qualitative examples in Figure 8, shows why BCBF is needed beyond generic branch fusion. Identity mapping performs worst, which confirms that cross-branch interaction itself is essential. Replacing BCBF with concatenation followed by a 1×1 convolution recovers part of the gain, but still under-performs compared to the full model. Standard cross-attention performs even worse in this setting. This result is consistent with the formulation-level comparison in Table 1. Conventional cross-attention is designed around token-level spatial correspondence, which is more appropriate when local matching is relatively reliable. In dynamic HDR reconstruction, however, saturation, under-exposure, occlusion, and motion often make such correspondence ambiguous at exactly the regions where branch interaction matters most. BCBF instead performs channel-subspace exchange between the base and detail branches, allowing structurally stable guidance and reliability-filtered local cues to complement each other without enforcing dense spatial matching. The ablation, therefore, provides empirical support for the task-specific motivation of BCBF rather than merely showing that one fusion block happens to score higher.

The component ablation in Appendix A.2 further clarifies the source of this gain. The full DAG design performs best, while removing spatial, channel, or pixel gating all degrades reconstruction quality. Channel and pixel gating contribute the largest drop in μ-PSNR when removed, whereas spatial gating produces a milder but still consistent degradation. DAG, therefore, works as a cooperative reliability estimator: channel and pixel gating provide the main feature-selection effect, while spatial gating supplies complementary regional guidance for difficult exposure-corrupted areas.

Table 10. Cross-branch fusion module replacement study on Kalantari17 under full-resolution, patch-based inference. Bold values indicate the best result in each column.

Fusion module	PSNR	μ-PSNR	SSIM	μ-SSIM
BCBF (mutual attention)	**42.653**	**44.581**	**0.9898**	**0.9919**
Identity (w/o fusion)	38.699	43.334	0.9819	0.9906
Concat + 1×1	41.912	44.510	0.9881	0.9917
Standard cross-attention	40.740	43.210	0.9870	0.9906

As a complementary qualitative analysis, Figure 9 visualizes the high-level fused representation produced after branch fusion and trunk encoding, immediately before the reconstruction head. The feature is shown both as a standalone map and as an overlay on the reference LDR patch to relate the latent response to the scene content. Although the representation is not directly interpretable as an image-domain output, its stronger activations align with structural boundaries, informative textures, and difficult exposure-transition regions. This visualization is intended as qualitative support for the same conclusion drawn from the ablations: the network concentrates reconstruction-relevant evidence before final HDR synthesis.



Figure 8. Qualitative comparison of cross-branch fusion module replacements on Kalantari17. Representative local patches are shown to compare the visual behavior of BCBF, identity mapping, concatenation followed by a 1×1 convolution, and standard cross-attention in terms of ghost suppression, edge continuity, and texture recovery.

Finally, we analyze how DAG capacity and placement affect the final result. Tables 8 and 9 show that, among $r \in \{4, 8, 16\}$, the default setting $r = 8$ achieves the best overall performance. Both a larger bottleneck ($r = 4$) and a more aggressively compressed setting ($r = 16$) lead to slightly inferior reconstruction quality. The placement study shows the same pattern: applying DAG in all stages gives the best performance, whereas restricting it to fewer stages reduces performance. This supports keeping DAG as a default part of the detail branch rather than treating detail filtering as an optional add-on.

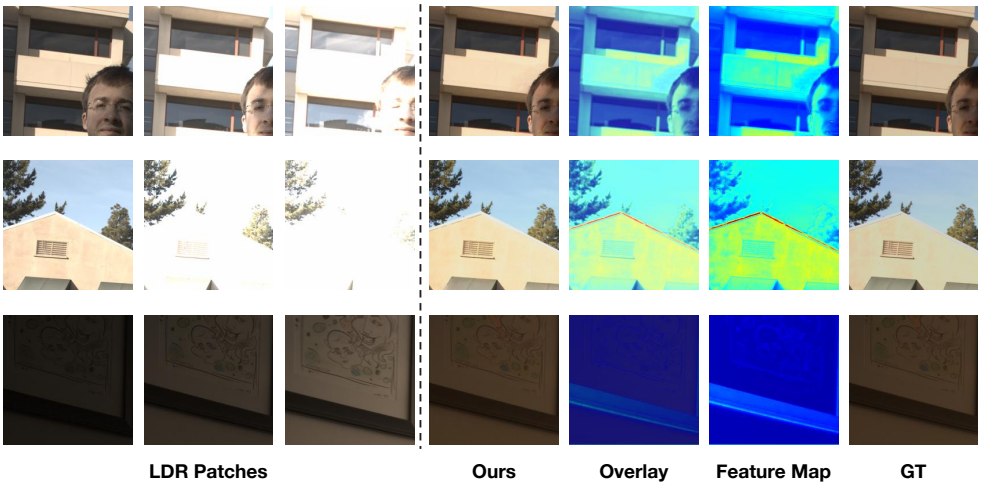


Figure 9. Visualization of the final fused representation before reconstruction on Kalantari17. From left to right: three input LDR patches, the tonemapped result of our method, the feature visualization overlaid on the reference LDR patch, the corresponding high-level fused representation, and the tone-matched GT patch. The visualized feature is extracted after branch fusion and trunk encoding, immediately before the reconstruction head, and reflects a high-level latent representation that concentrates on structurally and photometrically informative regions.

4.6. Failure Cases and Boundary Conditions

Although the proposed method performs strongly across the three benchmarks, some scenes still produce visible failure patterns. Figure 10 presents two representative examples. In the first case, the middle exposure is itself corrupted by motion or poor local exposure, which leads to unstable local structure and residual artifacts in the zoomed region. In the second case, severe non-rigid motion or strong parallax leads to visible errors around thin structures, motion boundaries, and depth discontinuities.

These examples illustrate the boundary conditions of the reference-centered setting introduced in Section 3.2. The reconstruction is most reliable when the middle exposure remains reasonably informative, the non-reference exposures still contain complementary evidence, and the local corruption is difficult but not completely degenerate. More specifically, the reference-centered design is most suitable when three conditions are approximately satisfied: (1) the middle exposure still preserves the main scene geometry and a usable amount of radiometric information, (2) the under-exposed and over-exposed inputs contain complementary evidence for shadows and highlights, and (3) the inter-exposure inconsistency is moderate enough that the non-reference observations can still be interpreted relative to a common anchor. In contrast, the assumption becomes fragile when the middle exposure simultaneously loses geometric and radiometric reliability, for example, because of strong local saturation, severe under-exposure, motion blur, or rolling-shutter distortion, or when no single exposure serves as a stable anchor under large parallax or highly non-rigid motion. Under these conditions, the limitation is not merely a difficult example but a structural weakness of the fixed-reference formulation. Because the present network organizes global modeling and cross-branch interaction around the middle exposure, errors in the middle exposure can propagate into both the global representation and the later detail recovery pathway. The complementary exposures can still provide partial correction, but they cannot fully replace a missing or unreliable anchor when the middle exposure itself is severely degraded. These cases should, therefore, be interpreted as practical limits of the current formulation rather than isolated anomalies.

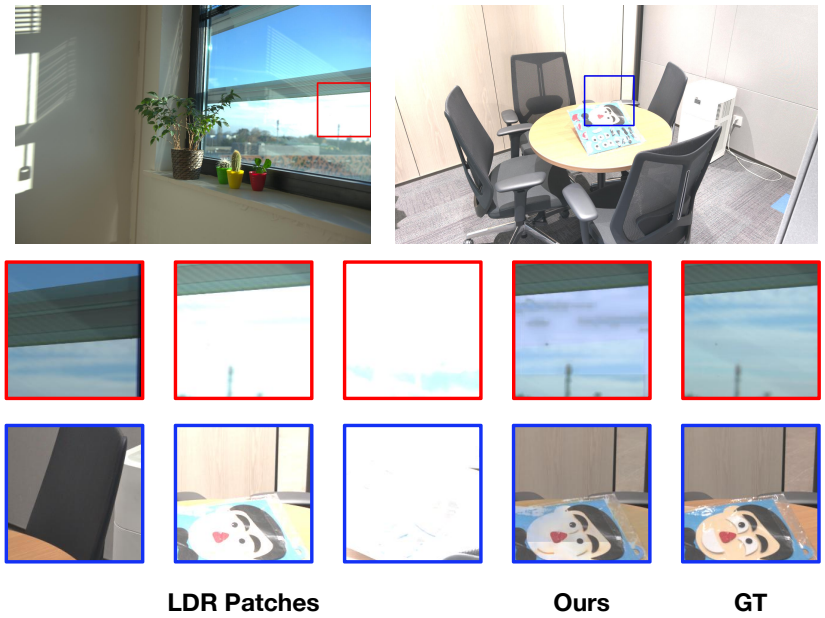


Figure 10. Representative failure cases of the proposed method. For each case, the full reference LDR image is shown in the overview panel, and the lower row provides zoomed-in patches of the reference LDR, the tonemapped result of our method, and the tone-matched GT. The red box marks the Tel example, whereas the blue box marks the Challenge123 example. The first example shows a scene in which the middle exposure is itself unreliable, leading to unstable local reconstruction. The second example shows a scene with severe non-rigid motion or strong parallax, where residual artifacts remain around thin structures and motion boundaries.

4.7. Limitations and Future Outlook

The current framework still has several methodological limitations. First, it assumes a relatively stable middle-exposure frame, so the current formulation remains tied to a fixed-reference reconstruction strategy. This assumption is standard in dynamic HDR reconstruction and is often reasonable for three-exposure brackets, but it becomes restrictive once the middle exposure is itself corrupted by strong local saturation, severe underexposure, motion blur, rolling-shutter distortion, or large parallax. Future work should, therefore, investigate adaptive reference selection at both the image level and the region level. One possible direction is to estimate reference confidence and dynamically choose the most trustworthy exposure as the structural anchor. Another is confidence-aware soft weighting across exposures, so that different spatial regions can rely on different inputs according to their local reliability. A more general direction is reference-free fusion, in which the network aggregates complementary evidence without depending on a single predefined anchor.

Second, BCBF is designed for cross-branch collaboration rather than explicit geometric alignment. This choice is appropriate for unstable local observations, but it can limit robustness under extreme non-rigid motion, large displacement, or thin-structure corruption. Incorporating motion priors, correspondence confidence, or lightweight geometry-aware guidance is, therefore, a promising direction.

Third, the current design is not optimized primarily for minimum latency. Its value lies in reconstruction quality under challenging dynamic HDR scenarios, and the reported complexity is intended to make the trade-off transparent. Future work may explore lighter branch configurations, more efficient interaction patterns, or stronger multi-scale and saturation-aware reliability modeling without compromising the collaborative behavior of the network.

Finally, the current study focuses on single-image HDR reconstruction from exposure brackets. Extending the framework to temporally consistent HDR video remains a promising direction, where cross-frame consistency constraints and temporal fusion mechanisms could further improve stability in real-world dynamic scenes.

5. Conclusions

This paper addresses dynamic multi-exposure HDR reconstruction from the perspective of base-detail collaboration. The proposed framework combines a WRTB-based base branch, a DAG-based detail branch, and BCBF for stage-wise cross-branch interaction. Rather than relying on a single representation stream, the model coordinates global structure and luminance modeling with reliability-aware local-detail refinement.

Experiments on Kalantari17, the Tel benchmark, and Challenge123 show a consistent pattern. The method remains competitive on the standard Kalantari17 benchmark and yields clearer gains on the more challenging Tel and Challenge123 benchmarks, where illumination variation, dark scenes, large motion, and strong saturation are more prominent. The ablation studies further support the value of explicit branch specialization, BCBF-based cross-branch interaction, and reliability-aware detail filtering, while the complexity analysis clarifies the associated structural trade-off.

The framework also has clear methodological limitations. Its current formulation assumes a stable middle-exposure reference and does not explicitly model severe geometric distortion. Future work will, therefore, focus on improving robustness under these harder conditions, exploring lighter interaction patterns, and extending the framework to temporally consistent HDR video reconstruction.

Author Contributions: Methodology, Q.Z.; software, Q.Z.; validation, Q.Z. and M.C.; formal analysis, M.C.; investigation, Q.Z., M.C. and Z.Z.; data curation, Q.Z.; writing—original draft preparation, Q.Z.; writing—review and editing, M.C., F.C. and Y.D.; supervision, Y.D.; project administration, Y.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used in this study are publicly available from their original sources. The training and inference configurations used to produce the main results are provided in the repository. Additional checkpoints and result files are available from the corresponding author upon reasonable request during peer review.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

BCAI	Bidirectional Cross-Attention Interaction
BCBF	Bidirectional Cross-Branch Fusion
CNN	Convolutional Neural Network
DAG	Detail-Aware Gating
DWConv	Depth-wise Convolution
EMA	Exponential Moving Average

FFN	Feed-Forward Network
FLOPs	Floating-Point Operations
GAP	Global Average Pooling
HDR	High Dynamic Range
HDR-VDP2	High Dynamic Range Visual Difference Predictor 2
LDR	Low Dynamic Range
LN	Layer Normalization
MACs	Multiply–Accumulate Operations
MAP	Maximum a Posteriori
MLP	Multi-layer Perceptron
PSNR	Peak Signal-to-Noise Ratio
SSIM	Structural Similarity Index Measure
SW-MSA	Shifted Window-based Multi-Head Self-Attention
THOP	PyTorch model complexity counter
ViT	Vision Transformer
W-MSA	Window-based Multi-Head Self-Attention
WRTB	Window-based Residual Transformer Block

Appendix A. Additional Ablations

Appendix A.1. Auxiliary Loss Weight Sensitivity

Table A1. Sensitivity study of auxiliary loss weights on Kalantari17 under full-resolution, patch-based inference. Here, λ_{base} controls the auxiliary supervision for the base branch, and λ_{detail} controls the gradient-based auxiliary supervision for the detail branch.

Setting $(\lambda_{\text{base}}, \lambda_{\text{detail}})$	PSNR	μ-PSNR	SSIM	μ-SSIM
$(0, 0)$	42.028	44.342	0.9891	0.9918
$(0.05, 0.05)$	42.460	44.396	0.9887	0.9914
$(0.1, 0.1)$	42.653	44.581	0.9898	0.9919
$(0.2, 0.2)$	42.147	44.384	0.9882	0.9918

Appendix A.2. DAG Component Ablation

Table A2 provides a more detailed view of how the three gating components contribute to DAG. The full DAG design achieves the best values on all reported metrics. Removing spatial gating causes only a limited PSNR drop (-0.063 dB) but a clearer decrease in μ-PSNR (-0.565 dB), suggesting that spatial gating is especially helpful for suppressing exposure-sensitive unreliable regions in the tonemapped domain. Removing channel gating leads to the largest PSNR drop (-0.450 dB), indicating that channel-wise modulation is a major factor in selecting informative detail responses. Removing pixel gating produces the largest μ-PSNR drop (-1.059 dB), which shows that fine-grained location-specific filtering is also essential when local evidence is corrupted by clipping, darkness, or motion.

Table A2. Component ablation of the Detail-Aware Gating (DAG) module on Kalantari17 under full-resolution, patch-based inference.

Setting	PSNR	μ-PSNR	SSIM	μ-SSIM
Full DAG	42.653	44.581	0.9898	0.9919
w/o spatial gating	42.590	44.016	0.9894	0.9916
w/o channel gating	42.203	43.535	0.9890	0.9912
w/o pixel gating	42.369	43.522	0.9891	0.9915

These component-level results are consistent with the main-text DAG studies. The capacity sweep shows that $r = 8$ performs better than both $r = 4$ and $r = 16$, which suggests that the gain does not come simply from increasing the size of the gating module. The placement study further shows that applying DAG in all stages performs best, with last-2 stages still outperforming first-2 stages. Taken together, these results support the interpretation that DAG is most effective as a moderate-capacity, multi-stage reliability estimator: channel and pixel gating provide the dominant feature-selection effect, while spatial gating offers complementary regional guidance, and this combination remains useful throughout the detail branch rather than only at the network input.

Appendix A.3. Additional Feature Visualizations

Complementary to the Kalantari17 example in Figure 9, we further visualize the final fused representation before reconstruction on the Tel benchmark and Challenge123. In both cases, the displayed feature is extracted after branch fusion and trunk encoding, immediately before the reconstruction head. These examples serve as supplementary evidence for the learned high-level latent representation under different illumination regimes. The Tel example corresponds to a relatively dark scene with low overall brightness and weak local contrast. Even there, the fused representation still concentrates on informative structures and faint texture boundaries, which is consistent with the strong quantitative performance on difficult low-illumination scenes. The Challenge123 example corresponds to a brighter and more strongly saturated scene. In that case, the representation remains focused on high-contrast edges, structural contours, and difficult exposure-transition regions. Together, these visualizations support the same conclusion as the main text: the proposed base-detail collaboration remains effective in both darker and brighter dynamic HDR scenes.

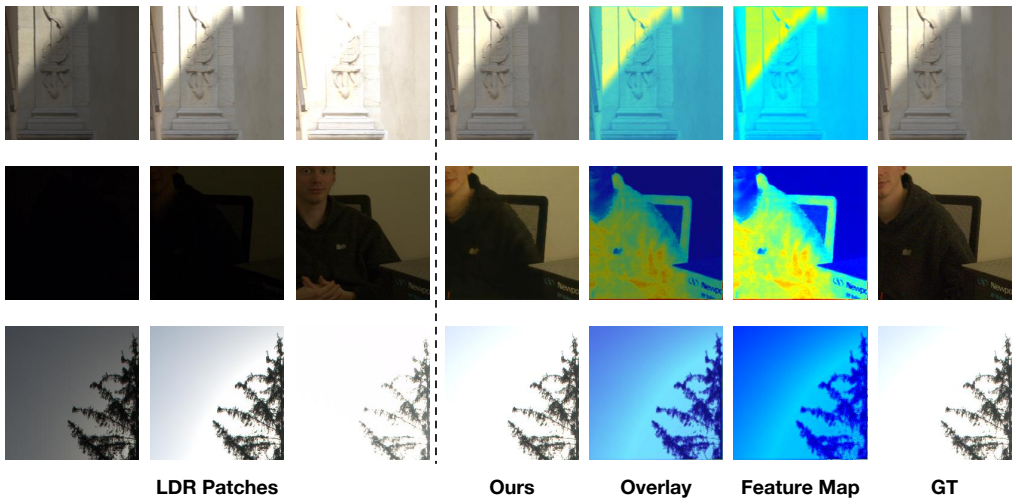


Figure A1. Additional visualization of the final fused representation before reconstruction on the Tel benchmark. This example corresponds to a relatively dark scene. From left to right: three input LDR patches, the tonemapped result of our method, the feature visualization overlaid on the reference LDR patch, the corresponding high-level fused representation, and the tone-matched GT patch. The visualization shows that the learned representation still emphasizes informative structures and subtle texture boundaries under low illumination.

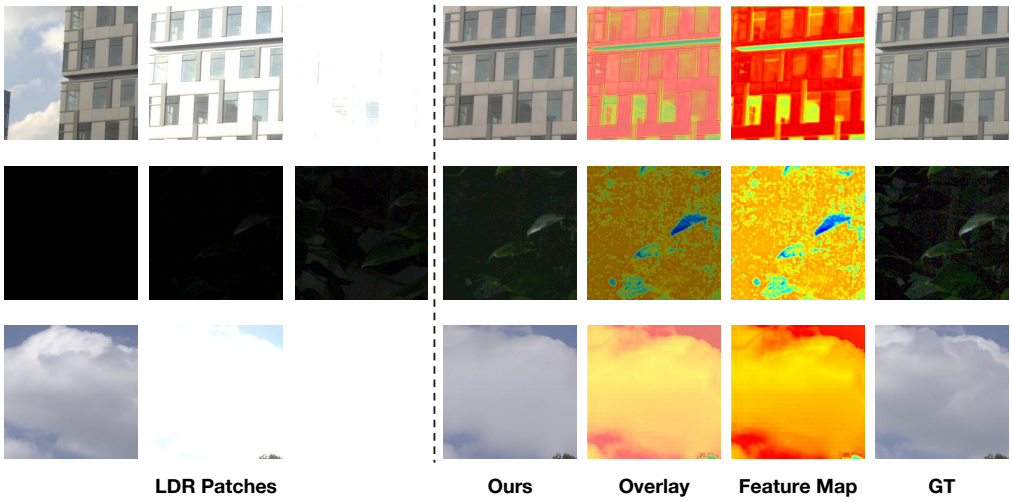


Figure A2. Additional visualization of the final fused representation before reconstruction on Challenge123. This example corresponds to a brighter scene with stronger saturation. From left to right: three input LDR patches, the tonemapped result of our method, the feature visualization overlaid on the reference LDR patch, the corresponding high-level fused representation, and the tone-matched GT patch. The visualization indicates that the learned representation remains focused on high-contrast edges and difficult exposure-transition regions despite the presence of bright highlights.

Appendix A.4. One-Iteration Training Pseudocode

Algorithm A1: One training iteration of the dual-branch Transformer-CNN model. Here, $T(\cdot)$ denotes the μ-law compressor (with $\mu = 5000$); $G(\cdot)$ denotes a detail operator (e.g., image gradients); $\gamma_g = 2.2$ is the gamma exponent; and ε is a small constant

Input: Exposure bracket $\{L^{(k)}, t_k\}_{k=1}^{3}$, ground truth H_{gt}, model f_θ, ablation/config switch $\mathcal{C}$

Output: Updated parameters θ

for $k \leftarrow 1$ **to** 3 **do**

$\tilde{H}^{(k)} \leftarrow (L^{(k)})^{\gamma_g} / (t_k + \varepsilon)$;

$F^{(k)} \leftarrow \mathrm{Concat}(\tilde{H}^{(k)}, L^{(k)})$;

$X \leftarrow \mathrm{Concat}(F^{(1)}, F^{(2)}, F^{(3)})$;

$(\hat{H}, \hat{H}_b, \hat{H}_d) \leftarrow f_\theta(X; \mathcal{C})$;

$\mathcal{L}_r \leftarrow \|T(\hat{H}) - T(H_{gt})\|_1$;

$\mathcal{L}_{\text{base}} \leftarrow \|T(\hat{H}_b) - T(H_{gt})\|_1$;

$\mathcal{L}_{\text{detail}} \leftarrow \|G(\hat{H}_d) - G(H_{gt})\|_1$;

$\mathcal{L} \leftarrow \mathcal{L}_r + \lambda_{\text{base}} \mathcal{L}_{\text{base}} + \lambda_{\text{detail}} \mathcal{L}_{\text{detail}}$;

$\theta \leftarrow \mathrm{AdamStep}\big(\theta, \nabla_\theta \mathcal{L}\big)$;

References

1. Yan, Q.; Gong, D.; Shi, Q.; Van Den Hengel, A.; Shen, C.; Reid, I.; Zhang, Y. Attention-Guided Network for Ghost-Free High Dynamic Range Imaging. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 1751–1760. [CrossRef]
2. Yan, Q.; Gong, D.; Zhang, Y. Deep HDR Imaging via a Non-Local Network. *IEEE Trans. Image Process.* **2020**, *29*, 4308–4322. [CrossRef] [PubMed]
3. Liu, Z.; Wang, Y.; Zeng, B.; Liu, S. Ghost-Free High Dynamic Range Imaging with Context-Aware Transformer. In *Proceedings of the Computer Vision–ECCV 2022*; Springer Nature: Cham, Switzerland, 2022; pp. 344–360. [CrossRef]

4. Tel, S.; Wu, Z.; Zhang, Y.; Heyrman, B.; Demonceaux, C.; Timofte, R.; Ginhac, D. Alignment-Free HDR Deghosting with Semantics Consistent Transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 12790–12799. [CrossRef]
5. Yan, Q.; Chen, W.; Zhang, S.; Zhu, Y.; Sun, J.; Zhang, Y. A Unified HDR Imaging Method with Pixel and Patch Level. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 22211–22220.
6. Kalantari, N.K.; Ramamoorthi, R. Deep High Dynamic Range Imaging of Dynamic Scenes. *ACM Trans. Graph.* **2017**, *36*, 144. [CrossRef]
7. Kong, L.; Li, B.; Xiong, Y.; Zhang, H.; Gu, H.; Chen, J. SAFNet: Selective Alignment Fusion Network for Efficient HDR Imaging. In *Proceedings of the Computer Vision—ECCV 2024*; Springer Nature: Cham, Switzerland, 2024; pp. 256–273. [CrossRef]
8. Grosch, T. Fast and Robust High Dynamic Range Image Generation with Camera and Object Movement. In *Proceedings of the Vision, Modeling, and Visualization (VMV)*; Kobbelt, L., Kuhlen, T., Aach, T., Westermann, R., Eds.; Springer: Heidelberg, Germany, 2006; pp. 277–284.
9. Pece, F.; Kautz, J. Bitmap Movement Detection: HDR for Dynamic Scenes. In Proceedings of the Conference on Visual Media Production (CVMP 2010), London, UK, 17–18 November 2010; pp. 1–8. [CrossRef]
10. Jacobs, K.; Loscos, C.; Ward, G. Automatic High-Dynamic Range Image Generation for Dynamic Scenes. *IEEE Comput. Graph. Appl.* **2008**, *28*, 84–93. [CrossRef] [PubMed]
11. Bogoni, L. Extending Dynamic Range of Monochrome and Color Images through Fusion. In Proceedings of the 15th International Conference on Pattern Recognition (ICPR), Barcelona, Spain, 3–8 September 2000; pp. 3007–3016. [CrossRef]
12. Kang, S.B.; Uyttendaele, M.; Winder, S.; Szeliski, R. High Dynamic Range Video. *ACM Trans. Graph.* **2003**, *22*, 319–325. [CrossRef]
13. Sen, P.; Kalantari, N.K.; Yaesoubi, M.; Darabi, S.; Goldman, D.B.; Shechtman, E. Robust Patch-Based HDR Reconstruction of Dynamic Scenes. *ACM Trans. Graph.* **2012**, *31*, 203. [CrossRef]
14. Hu, J.; Gallo, O.; Pulli, K.; Sun, X. HDR Deghosting: How to Deal with Saturation? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, OR, USA, 23–28 June 2013; pp. 1163–1170. [CrossRef]
15. Wu, S.; Xu, J.; Tai, Y.W.; Tang, C.K. Deep High Dynamic Range Imaging with Large Foreground Motions. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 120–135. [CrossRef]
16. Liu, Z.; Lin, W.; Li, X.; Rao, Q.; Jiang, T.; Han, M.; Fan, H.; Sun, J.; Liu, S. ADNet: Attention-Guided Deformable Convolutional Network for High Dynamic Range Imaging. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 463–470.
17. Song, J.W.; Park, Y.I.; Kong, K.; Kwak, J.; Kang, S.J. Selective TransHDR: Transformer-Based Selective HDR Imaging Using Ghost Region Mask. In *Proceedings of the Computer Vision—ECCV 2022*; Springer: Cham, Switzerland, 2022; pp. 288–304. [CrossRef]
18. Hu, T.; Wu, L.; Dong, W.; Wu, P.; Sun, J.; Xu, X.; Yan, Q.; Zhang, Y. Boosting HDR Image Reconstruction via Semantic Knowledge Transfer. *IEEE Trans. Image Process.* **2026**, *35*, 1910–1922. [CrossRef] [PubMed]
19. Mai, T.T.N.; Lam, E.Y.; Lee, C. Deep Unrolled Low-Rank Tensor Completion for High Dynamic Range Imaging. *IEEE Trans. Image Process.* **2022**, *31*, 5774–5787. [CrossRef] [PubMed]
20. Li, X.; Ni, Z.; Yang, W. AFUNet: Cross-iterative Alignment-Fusion Synergy for HDR Reconstruction via Deep Unfolding Paradigm. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Honolulu, HI, USA, 19–20 October 2025; pp. 10666–10675.
21. Zhou, W.; Yang, Y.; Hu, T.; Hui, P.; Jin, J.; Cao, Y.; Yan, Q.; Zhang, Y. High Dynamic Range Imaging via Spatial-Frequency Interaction. *IEEE Trans. Circuits Syst. Video Technol.* **2026**. [CrossRef]
22. Niu, Y.; Wu, J.; Liu, G.; Guo, W.; Lau, R.W.H. HDR-GAN: HDR Image Reconstruction from Multi-Exposed LDR Images with Large Motions. *IEEE Trans. Image Process.* **2021**, *30*, 3885–3896. [CrossRef] [PubMed]
23. Guan, Y.; Xu, R.; Yao, M.; Gao, R.; Wang, L.; Xiong, Z. Diffusion-Promoted HDR Video Reconstruction. In *Proceedings of the Computer Vision—ECCV 2024 Workshops*; Springer: Cham, Switzerland, 2025; pp. 20–38. [CrossRef]
24. Chen, Z.; Wang, Y.; Cai, X.; You, Z.; Lu, Z.; Zhang, F.; Guo, S.; Xue, T. UltraFusion: Ultra High Dynamic Imaging using Exposure Fusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 11–15 June 2025; pp. 16111–16121. [CrossRef]
25. Yan, Q.; Hu, T.; Sun, Y.; Tang, H.; Zhu, Y.; Dong, W.; Van Gool, L.; Zhang, Y. From Dynamic to Static: Stepwisely Generate HDR Image for Ghost Removal. *IEEE Trans. Circuits Syst. Video Technol.* **2025**, *35*, 1409–1421. [CrossRef]
26. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.Y.; et al. Segment Anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 4015–4026.
27. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022. [CrossRef]

28. Zhang, X.; Zhang, Y.; Yu, F. HiT-SR: Hierarchical Transformer for Efficient Image Super-Resolution. In *Proceedings of the Computer Vision—ECCV 2024*; Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G., Eds.; Springer: Cham, Switzerland, 2025; Volume 15098, pp. 483–500. [CrossRef]
29. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In *Proceedings of the Computer Vision—ECCV 2018*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 3–19. [CrossRef]
30. Chen, Y.; Zhang, K.; Zuo, W.; Zhang, L. DEA-Net: Single Image Dehazing Based on Detail-Enhanced Convolution and Content-Guided Attention. *IEEE Trans. Image Process.* **2024**, *33*, 1002–1015. [CrossRef] [PubMed]
31. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141. [CrossRef]
32. Zhao, H.; Kong, X.; He, J.; Qiao, Y.; Dong, C. Efficient Image Super-Resolution Using Pixel Attention. In *Computer Vision–ECCV 2020 Workshops*; Bartoli, A., Fusiello, A., Eds.; Springer: Cham, Switzerland, 2020; Volume 12537, pp. 56–72. [CrossRef]
33. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In *Proceedings of the Advances in Neural Information Processing Systems*; Guyon, I., von Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S.V.N., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
34. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803. [CrossRef]
35. Zamir, S.W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F.S.; Yang, M.H. Restormer: Efficient Transformer for High-Resolution Image Restoration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 5728–5739.
36. Zhao, Z.; Bai, H.; Zhang, J.; Zhang, Y.; Xu, S.; Lin, Z.; Timofte, R.; Van Gool, L. CDDFuse: Correlation-Driven Dual-Branch Feature Decomposition for Multi-Modality Image Fusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 5906–5916. [CrossRef]
37. Wang, Z.; Bovik, A.C. Mean Squared Error: Love It or Leave It? A New Look at Signal Fidelity Measures. *IEEE Signal Process. Mag.* **2009**, *26*, 98–117. [CrossRef]
38. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]
39. Mantiuk, R.; Kim, K.I.; Rempel, A.G.; Heidrich, W. HDR-VDP-2: A Calibrated Visual Metric for Visibility and Quality Predictions in All Luminance Conditions. *ACM Trans. Graph.* **2011**, *30*, 40. [CrossRef]
40. Chen, J.; Yang, Z.; Chan, T.N.; Li, H.; Hou, J.; Chau, L.P. Attention-Guided Progressive Neural Texture Fusion for High Dynamic Range Image Restoration. *IEEE Trans. Image Process.* **2022**, *31*, 2661–2672. [CrossRef] [PubMed]
41. Yan, Q.; Hu, T.; Sun, Y.; Tang, H.; Zhu, Y.; Dong, W.; Van Gool, L.; Zhang, Y. Toward High-Quality HDR Deghosting With Conditional Diffusion Models. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 4011–4026. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.