


Single-Image HDR Reconstruction with Attention-Driven Autoencoder

Cevher RENK¹,

 0009-0000-4023-3683

Serdar ÇİFTÇİ¹

 0000-0001-7074-2876

¹Department of Computer Engineering, Harran University
Sanliurfa, Turkey

cevherrenk@harran.edu.tr, serdarciftci@harran.edu.tr

Abstract

High Dynamic Range (HDR) imaging captures fine details in both bright and dark regions, closely mimicking the sensitivity of human vision. Traditional HDR methods rely on multiple exposures but often face motion artifacts, hardware constraints, and high computational costs, limiting their real-world use. This study proposes an attention-augmented autoencoder (AE) with a U-Net-like structure for reconstructing HDR images from a single Low Dynamic Range (LDR) input. Five attention mechanisms Spatial, Channel, Bottleneck, Squeeze-and-Excitation (SE), and Self-Attention were individually integrated, forming distinct model variants. Experiments on the DrTMO dataset show that Spatial Attention achieves the best performance, improving SSIM from 0.9130 to 0.9310, PSNR from 21.50 dB to 22.30 dB, and reducing LPIPS from 0.1090 to 0.0928. These results highlight the effectiveness of attention mechanisms in enhancing both structural fidelity and perceptual quality for single-image HDR reconstruction.

Keywords: HDR Reconstruction, Autoencoder, Attention Mechanisms

1. Introduction

Visual scenes in the real world are characterized by broad luminance ranges, often containing both intense highlights and deep shadows within the same frame. Capturing this complexity requires imaging systems that exceed the limitations of standard sensors. High Dynamic Range (HDR) techniques have emerged as essential tools in modern vision applications, enabling enhanced detail retention across diverse lighting conditions. Low Dynamic Range (LDR) images often fail to preserve subtle contrast, while HDR imaging offers superior visual fidelity. This makes HDR crucial in areas like digital cinema, medicine, automotive vision, and surveillance [1–3]. Traditional HDR synthesis pipelines are based on fusing multiple exposures of the same scene. Although effective in static environments, these methods struggle with dynamic content, often suffering from motion-induced artifacts and hardware constraints [4–5]. Recent advances in deep learning have introduced multi-exposure HDR models that mitigate ghosting, but they still depend on multiple aligned inputs. As a more practical alternative, single-image HDR reconstruction has gained momentum by eliminating exposure alignment and sensor requirements while maintaining computational efficiency [6–

7]. One of the pioneering models in this domain utilizes a Convolutional Neural Network (CNN) to predict HDR scenes from a single low-exposure LDR image [7]. Encoder-decoder structures, especially autoencoders, have shown promise in extracting compact representations for HDR restoration [8]. Symmetric autoencoder (AE) architectures, such as U-Net, have become popular for minimizing detail loss by enriching feature flow through skip connections [9–10]. However, traditional AE models often fail to assign appropriate attention to semantically important regions in the scene, resulting in color distortion, contrast loss, and the omission of critical details. Recent works have enhanced AE-based models with attention layers, enabling more selective focus on semantically rich image regions. In this context, five distinct attention mechanisms spatial, channel, bottleneck, squeeze-and-excitation, and self-attention have been individually integrated into the proposed model, each forming a separate variant.

In this study, a U-Net like AE architecture is designed by extending a previously proposed baseline model [1]. Five attention mechanisms Spatial, Channel, Bottleneck, SE, and Self-Attention were added individually and evaluated separately. Experimental results on the DrTMO dataset demonstrate that the Spatial Attention variant outperforms the others across SSIM, PSNR, and LPIPS metrics. The main contributions of this work are summarized as follows:

- Five different attention mechanisms (Channel, Spatial, SE, Bottleneck, and Self-Attention) were independently integrated into a U-Net like autoencoder architecture. Each variant was tested separately to complete missing scene details and improve HDR reconstruction quality.
- The integration of spatial and channel attention mechanisms effectively preserved critical visual content in both bright and dark regions of the scene, directly enhancing color consistency, brightness balance, and structural coherence.
- Self-attention, bottleneck, and SE modules helped the model better capture long-range dependencies, compress representations, and enhance channel-wise information flow. These improvements led to higher contextual awareness and better learning capacity.

2. Related Works

HDR imaging is a key topic in image processing. It aims to generate visuals that match human perception by representing both bright and dark areas together. As a result, HDR synthesis

techniques especially with the rise of deep learning have gained substantial attention in tasks such as image generation, quality enhancement, and scene representation [1]. Traditional HDR synthesis methods are typically based on fusing multiple images captured at different exposure levels. While these approaches yield satisfactory results in static scenes, their applicability is limited due to sensitivity to motion, exposure synchronization issues, and hardware dependencies [2-3]. Single-image HDR methods have emerged as practical alternatives, with lower computational cost and no need for extra hardware [4-5]. One of the earliest deep learning-based methods in this domain aimed to reconstruct missing brightness and detail components from a single low-exposure LDR image. Using CNNs, this method enabled more accurate estimation of scene illumination [5]. Following this advancement, autoencoder (AE)-based architectures built on encoder-decoder designs have gained popularity as powerful tools for learning compact scene representations and recovering missing information. In AE-based models, symmetric structures and skip connections allow joint processing of high- and low-level features, improving reconstruction accuracy. In this context, U-Net is widely used in image processing. It can convert abstracted encoder features into detailed representations [6]. Moreover, segmentation-inspired architectures have been employed to capture semantic relationships within a scene, contributing to improved contextual consistency in HDR synthesis [7]. Recent studies have also introduced HDR synthesis approaches guided by scene-derived semantic information, showing significant improvements in terms of color fidelity, structural consistency, and visual realism [8]. These models enable multi-scale contextual representation learning, allowing for the generation of more detailed HDR scenes. Furthermore, AE-based architectures enhanced with attention mechanisms have shown notable improvements in preserving details in semantically dense regions [9]. Some recent approaches have proposed multi-task HDR systems that jointly handle inverse tone mapping and super-resolution. Such frameworks offer flexible and comprehensive solutions by simultaneously addressing various aspects of image quality [10]. Additionally, deep neural networks used for general image enhancement such as contrast stretching and low-light restoration have also provided a foundational basis for HDR reconstruction tasks [11]. In summary, many existing HDR generation models incorporate attention mechanisms such as spatial, channel, squeeze, bottleneck, and self-attention either in a limited capacity or in isolated configurations. In contrast, the proposed method systematically integrates each attention type into the architecture as a separate model variant, enabling a comparative evaluation of their individual contributions. The proposed framework offers deeper architecture and more complete comparisons than previous works.

3. Proposed Method

In this study, we propose a multi-module deep learning architecture enhanced with attention mechanisms for generating high-quality High Dynamic Range (HDR) scenes from a single Low Dynamic Range (LDR) image. The core framework is built upon a U-Net like autoencoder structure [1] and comprises three primary sub-networks: HDR Encoding Network (\mathcal{N}_1), Up-Exposure Network (\mathcal{N}_2), and Down-Exposure Network (\mathcal{N}_3). These sub-networks are designed to simulate virtual exposure diversity of a scene and reconstruct

accurate HDR representations [6]. To address the common limitations observed in the literature, such as contextual awareness deficiency and loss of detail, five different attention mechanisms have been integrated into the architecture: Spatial Attention, Channel Attention, Bottleneck Attention, Squeeze-and-Excitation (SE) Attention and Self-Attention. To analyze their individual contributions, five attention types were applied separately within the encoder and decoder blocks, resulting in modular model variants. The performance of each variant was systematically evaluated [9]. Additionally, the training of the proposed model was guided by four complementary loss functions aimed at preserving both structural and perceptual consistency: reconstruction loss, perceptual loss, total variation loss, and representation loss. This combination of losses improves pixel fidelity while enhancing visual realism and semantic consistency.

3.1. General Structure of the Model

The proposed architecture is designed around three dedicated modules to address information loss associated with exposure variability. Each sub-network targets the recovery of scene details corresponding to different exposure conditions, thereby contributing to the final HDR synthesis.

3.1.1. HDR Encoding Network (\mathcal{N}_1)

The HDR Encoding Network is an encoder-decoder architecture that transforms an input LDR image into an exposure-independent scene representation. Exposure-normalized feature representations are obtained by scaling encoded features relative to their corresponding exposure times (Δt), enabling consistent modeling across varying lighting conditions. The encoder compresses feature information via convolutional layers into latent representations, which the decoder then reconstructs into exposure-sensitive intermediate forms.

Given two differently exposed input images I_1 and I_2 , encoding is performed as follows:

$$\hat{X}_1 = \mathcal{N}_1(I_1) \quad (1)$$

$$\hat{X}_2 = \mathcal{N}_1(I_2) \quad (2)$$

These encoded features are normalized based on their exposure ratios to simulate exposure diversity:

$$Z_1 = \hat{X}_1 \left(\frac{\Delta t_2}{\Delta t_1} \right) \quad (3)$$

$$Z_2 = \hat{X}_2 \left(\frac{\Delta t_1}{\Delta t_2} \right) \quad (4)$$

Here, Δt_i represents the exposure time of the corresponding image. The encoder utilizes Conv + ReLU + BatchNorm layers, while the decoder uses upsampling + convolution + activation structures. Attention modules are integrated to enhance spatial awareness in the encoder and preserve contextual and structural integrity in the decoder.

3.1.2. Up-Exposure Network (\mathcal{N}_2)

The Up-Exposure Network aims to recover details lost in underexposed (dark) regions. It receives as input the exposure-normalized representation \hat{X}_1 , derived from the HDR Encoding Network:

$$\hat{I}_2 = \mathcal{N}_2 = \mathcal{N}_2 \left(\hat{X}_1 \left(\frac{\Delta t_2}{\Delta t_1} \right) \right) \quad (5)$$

This sub-network, based on an encoder-decoder design, enhances dark scene features in the encoder, while the decoder synthesizes brighter representations with improved clarity. Skip-connections similar to those in U-Net help preserve fine-grained detail.

3.1.3. Down-Exposure Network (\mathcal{N}_3)

The Down-Exposure Network addresses information saturation in overexposed regions. It takes as input the normalized high-exposure representation \hat{X}_2 from the HDR Encoding Network:

$$\hat{I}_1 = \mathcal{N}_3 = \mathcal{N}_3 \left(\hat{X}_2 \left(\frac{\Delta t_1}{\Delta t_2} \right) \right) \quad (6)$$

The encoder re-encodes structural features from saturated regions, while the decoder restores balance by integrating these features into the HDR output. U-Net-inspired skip-connections are also employed here.

In summary, the architecture composed of these three modules reconstructs HDR scenes more accurately by utilizing multiple exposure levels generated from a single LDR image. The overall model structure is illustrated in Figure 1.

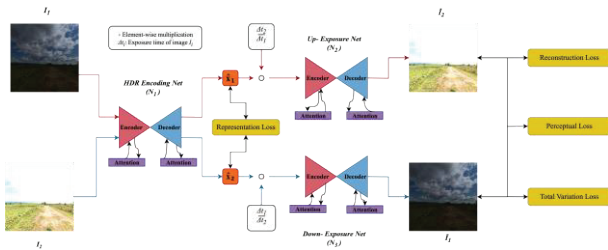


Figure 1. Overall architecture of the HDR Attention Network (HDR-AttNet).

Overview of the proposed HDR-AttNet architecture composed of three interconnected modules HDR Encoding, Up-Exposure, Down-Exposure. Attention modules are highlighted in purple.

3.2. Integration of Attention Mechanisms

In this study, five different attention mechanisms Spatial, Channel, Bottleneck, Squeeze-and-Excitation (SE), and Self-Attention were integrated into the architecture to enhance contextual awareness and reduce detail loss. These mechanisms were individually applied to the encoder and decoder layers, resulting in independent model variants [12]. The five attention mechanisms employed in this study were deliberately selected to represent a diverse set of attention strategies with proven success in various vision-related tasks. Instead of using one attention type, this study examines how each affects single-image HDR reconstruction. Spatial and Channel Attention were chosen due to their widespread effectiveness in enhancing local and global feature representations in convolutional networks. Bottleneck Attention was included to evaluate the role of compact feature modulation, especially relevant in encoder-decoder transitions. SE modules are well-regarded for their lightweight yet powerful channel-wise recalibration, while Self-Attention, inspired by Transformers, offers insights into global dependency modeling. By evaluating these varied mechanisms

within a unified architecture, this study seeks to provide a well-rounded understanding of the practical contributions of attention in HDR synthesis and identify which strategies yield the most significant improvements.

3.2.1. Spatial Attention

The spatial attention mechanism enhances the modeling of structurally prominent regions within the scene by emphasizing local details. It operates by concatenating average and maximum pooled feature maps along the channel axis and applying a convolutional operation. The mechanism is formulated as follows:

$$M_{spatial} = \sigma(f_{3 \times 3}([AvgPool(F); MaxPool(F)])) \quad (7)$$

Here $f_{3 \times 3}$ denotes a 3x3 convolution filter, $[]$ represents concatenation along the channel dimension, and σ is the sigmoid activation function. This design strengthens the representation of spatially salient areas within the scene [13-14].

3.2.2. Channel Attention

The channel attention mechanism models the relative importance of each feature channel to improve color fidelity and brightness accuracy. It utilizes global average pooling to extract channel descriptors and is defined as:

$$Z = \sigma(W_2 \cdot ReLU(W_1 \cdot s)) \quad (8)$$

In this expression, s is the channel-wise pooled descriptor, W_1 and W_2 are the weight matrices of fully connected layers, and σ represents the sigmoid activation function [15].

3.2.3. Bottleneck Attention

Bottleneck attention aims to suppress redundant information by reducing feature dimensionality and emphasizing critical representations. This mechanism is especially effective at encoder-decoder transitions and is expressed as:

$$M_{bottleneck} = \sigma(W_2 \cdot ReLU(W_1 \cdot M)) \quad (9)$$

It provides an efficient way to filter out noise and irrelevant features, ensuring a more focused information flow [16].

3.2.4. Squeeze-and-Excitation (SE) Attention

The SE module adaptively recalibrates channel-wise feature responses by learning the relative importance of each channel. It is mathematically defined as:

$$M_{SE} = M \odot \sigma(f_{dense}(M)) \quad (10)$$

Here, f_{dense} denotes the fully connected network responsible for the squeeze and excitation operations, and \odot represents element-wise multiplication along the channel dimension. This mechanism contributes to improved color consistency and luminance balance [17-18].

3.2.5. Self-Attention

The self-attention mechanism captures long-range dependencies between distant pixels in a scene, thereby enabling global contextual awareness. Based on the Transformer architecture, it is defined by the following equation:

$$\begin{aligned} & \text{Attention}(Q, K, V) \\ &= \text{softmax}\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right) V \end{aligned} \quad (11)$$

Where:

-Q (Query): linear projection of the current spatial location in the feature map

-K (Key): encodes the semantic content of all positions in the feature map

-V (Value): provides the contextual feature vectors used for weighted aggregation

In the context of image data, these matrices are derived via 1×1 convolutions over the input feature map and enable global pixel-level interaction across the spatial domain. Here, Q , K , and V are the query, key, and value matrices, respectively, and d_k is a scaling factor. This approach allows the network to build a more holistic representation by considering the entire scene context [19].

3.3. Loss Functions

In this study, four different loss functions were jointly utilized during training to ensure structural consistency, perceptual accuracy, and smooth reconstructions. Each loss function targets a different aspect of HDR synthesis, contributing to both pixel-level fidelity and global perceptual realism.

3.3.1. Reconstruction Loss

The reconstruction loss evaluates the pixel-wise similarity between the predicted LDR image and the ground truth LDR image with the target exposure. To this end, the ℓ_1 (Mean Absolute Error) is employed:

$$\mathcal{L}_{rec} = \|\hat{I} - I\|_1 \quad (12)$$

Here, \hat{I} denotes the pseudo LDR image generated by the model, while I represents the ground truth LDR image with the target exposure. The pseudo LDR image refers to the model's output that approximates an LDR exposure level, used for training supervision in lieu of ground-truth HDR data. This approach is widely used in HDR image reconstruction literature for its robustness and simplicity.

3.3.2. Perceptual Loss

Perceptual loss targets both pixel accuracy and visual distinctiveness of generated images. It utilizes feature maps extracted from pre-trained layers of the VGG-19 network to compute perceptual differences:

$$\mathcal{L}_{perc} = \sum_l \|\phi_l(\hat{I}) - \phi_l(I)\|_2^2 \quad (13)$$

Here, ϕ_l denotes the feature map extracted from the l^{th} layer of the VGG network. This loss encourages the generated images to maintain semantic and structural consistency.

3.3.3. Total Variation Loss

Total variation loss aims to reduce sharp transitions and undesired artifacts in the reconstructed image, resulting in smoother outputs. It is particularly effective in minimizing checkerboard patterns and artificial noise in HDR reconstructions:

$$\begin{aligned} \mathcal{L}_{tv} = \sum_{i,j} & \left(|\hat{I}_{i+1,j} - \hat{I}_{i,j}|^2 \right. \\ & \left. + |- \hat{I}_{i,j}|^2 \right) \end{aligned} \quad (14)$$

This formulation minimizes brightness discontinuities between neighboring pixels in both horizontal and vertical directions.

3.3.4. Representation Loss

Representation loss is defined to ensure consistency between the latent features extracted by the HDR Encoding Network (\mathcal{N}_1) from two inputs captured at different exposure levels. It penalizes deviations between the encoded representations using the ℓ_2 norm:

$$\mathcal{L}_{rep} = \|\hat{X}_1 - \hat{X}_2\|_2^2 \quad (15)$$

where $\hat{X}_1 = \mathcal{N}_1(I_1)$ and $\hat{X}_2 = \mathcal{N}_1(I_2)$. This constraint ensures semantic coherence and structural alignment between feature representations from the HDR Encoding module.

4. Experimental Results

This section presents the experimental results of the proposed HDR-AttNet model, trained with different attention mechanism variants. The training configurations, quantitative metric comparisons, qualitative visual analyses, and ablation studies are thoroughly examined.

4.1. Training Configuration and Computational Environment

The HDR-AttNet model, designed for high dynamic range (HDR) scene reconstruction, was trained on the DrTMO dataset [1]. The dataset was synthesized from 1,043 collected HDR images with nine exposure values, resulting in 46,935 LDR images for training and 6,210 LDR images for testing, each with a resolution of 512×512 pixels. During training, random pairs of differently exposed images were sampled from the dataset. The training process was configured to run for a total of 200,000 steps, spanning approximately 6 epochs, with each epoch consisting of around 31,500 steps on average. The training durations for each model variant were measured under the same hardware configuration (NVIDIA Tesla V100-SXM2 GPU, 16 GB). The results showed notable differences in time efficiency among the attention mechanisms. The Self-Attention variant required the longest training time at approximately 28.5 hours, followed by Squeeze-and-Excitation 24.7 hours, Bottleneck Attention 21.3 hours, Spatial Attention 19.5 hours, and Channel Attention 18.2 hours. These outcomes reflect the computational complexity of each mechanism, with Self-Attention demanding more resources due to its global context modeling capabilities. The entire training pipeline was implemented using the PyTorch deep learning framework. The Adam optimizer was employed for model optimization, and four distinct loss functions were incorporated with specific

weights. The training configuration and selected hyperparameters are summarized in Table 1.

Table 1. Training Configuration and Hyperparameter Settings

Parameter	Value
Training Dataset	DrTMO
Total Training Steps	200,000
Total Epochs	6
Steps per Epoch	~31,500
Batch Size	8
Optimizer	Adam
Learning Rate	0.0001
Loss Weights	$\lambda_1=100.0, \lambda_2=1.0, \lambda_3=0.1, \lambda_4=0.00001$
Hardware	NVIDIA Tesla V100-SXM2 (16 GB)
Framework	PyTorch

Each sub-network in the proposed model (HDR Encoding Net, Up-Exposure Net, Down-Exposure Net) adopts a 7-level U-Net-like encoder-decoder architecture. At each level, two convolutional layers (3×3 kernels) are applied, followed by batch normalization and ReLU (or Leaky ReLU) activation. The input features are progressively increased from 16 to 256 channels (HDR Encoding Net) and from 32 to 512 channels (Up/Down-Exposure Nets). Sub-pixel convolution is used for upsampling, which is more efficient than traditional deconvolution. The total number of layers per sub-network is 28, excluding skip connections and normalization layers. Although the exact GFLOPs were not calculated, the architecture is designed to be lightweight and suitable for real-time inference.

4.2. Performance Evaluation and Metric Comparisons

To objectively evaluate the quality of the generated HDR scenes, several metrics were employed that assess both structural and perceptual fidelity. In this study, three primary metrics were used to analyze the accuracy of model outputs: PSNR (Peak Signal-to-Noise Ratio), SSIM (Structural Similarity Index), and LPIPS (Learned Perceptual Image Patch Similarity) [20-22]. Each metric focuses on evaluating different aspects of the HDR results generated from a single LDR input image. PSNR quantifies the pixel-level difference between the predicted and reference HDR images, offering a measure of structural fidelity [20]. This metric is particularly indicative of brightness accuracy and overall structural consistency. In contrast, SSIM evaluates visual coherence by simultaneously considering luminance, contrast, and structural components in a manner more aligned with human visual perception [21]. On the other hand, LPIPS is computed using deep feature representations learned by neural networks, providing a perceptual similarity assessment that better reflects human visual judgment. This metric plays a crucial role in evaluating visual realism and the preservation of scene details [22].

4.2.1. PSNR-Based Performance Evaluation

PSNR (Peak Signal-to-Noise Ratio) is a widely used objective metric that measures the pixel-level fidelity between a reconstructed image and its reference counterpart. It is derived from the Mean Squared Error (MSE) and expressed in decibels (dB). A higher PSNR value indicates better image quality with

less distortion.

The PSNR score between a reference image I and a reconstructed image \hat{I} is defined as:

$$PSNR(I, \hat{I}) = 10 \cdot \log_{10} \left(\frac{(MAX_I)^2}{\frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n (I(i, j) - \hat{I}(i, j))^2} \right) \quad (16)$$

where:

- $I(i, j)$ is the pixel value at position (i, j) in the reference HDR image.
- $\hat{I}(i, j)$ is the corresponding pixel in the reconstructed image.
- m and n are the image dimensions (height and width).
- MAX_I is the maximum possible pixel value (e.g., 1.0 for normalized images or 255 for 8-bit images).

The denominator represents the Mean Squared Error (MSE) between I and \hat{I} . As PSNR increases, the reconstructed image is considered to have higher visual accuracy and lower pixel-wise distortion relative to reference. The average PSNR performance of each model variant enhanced with different attention mechanisms is illustrated in Figure 2. Higher PSNR values indicate that the reconstructed HDR images are structurally more similar to the reference scenes and contain less distortion. This reflects the model's effectiveness in brightness preservation and noise reduction.

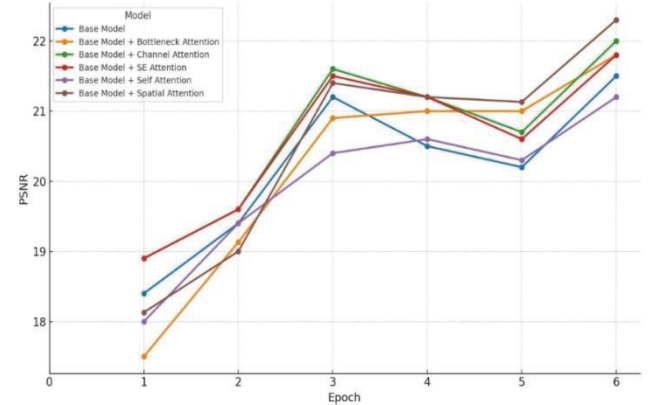


Figure 2. Comparison of average PSNR performance across model variants.

4.2.2. SSIM-Based Structural Consistency Analysis

SSIM (Structural Similarity Index Measure) is an advanced image quality metric that measures perceptual similarity by comparing the structural information between a reference and a reconstructed image. Unlike PSNR, which only considers pixel-wise differences, SSIM evaluates luminance, contrast, and structural components, providing a perceptually relevant assessment of visual fidelity.

The SSIM score between a reference image I and a reconstructed image \hat{I} is defined as:

$$SSIM(I, \hat{I}) = \frac{(2\mu_I\mu_{\hat{I}} + C_1)(2\sigma_{II} + C_2)}{(\mu_I^2 + \mu_{\hat{I}}^2 + C_1)(\sigma_I^2 + \sigma_{\hat{I}}^2 + C_2)} \quad (17)$$

where:

- $\mu_I, \mu_{\hat{I}}$: mean intensities of the reference and reconstructed images.
- $\sigma_I^2, \sigma_{\hat{I}}^2$: variances of the respective images.
- σ_{II} : covariance between I and \hat{I} .
- C_1 and C_2 : constants to stabilize the division in case of weak denominators.

SSIM values range from 0 to 1, where a value closer to 1 indicates higher structural similarity and better perceptual alignment with the reference image.

The structural consistency of each model variant within the scene was evaluated using average SSIM scores, and the results are presented in Figure 3. The model configured with the spatial attention mechanism demonstrated a clear advantage in preserving local details and achieved the highest performance in this metric.

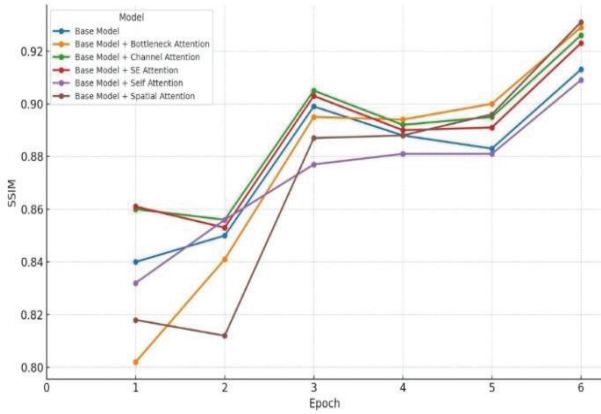


Figure 3. Comparative distribution of average SSIM scores across model variants.

4.2.3. LPIPS-Based Perceptual Similarity Evaluation

LPIPS (Learned Perceptual Image Patch Similarity) is a perceptual similarity metric that compares high-level feature representations extracted from pre-trained deep neural networks. Unlike PSNR and SSIM, which evaluate pixel-level differences, LPIPS operates in the feature space and reflects how similar images appear to the human visual system. It is considered more robust for evaluating perceptual quality in deep learning-based image generation tasks.

The LPIPS score between a reference image I and a reconstructed image \hat{I} is defined as:

$$LPIPS(I, \hat{I}) = \sum_l \frac{1}{H_l W_l} \sum_{h=1}^{H_l} \sum_{w=1}^{W_l} \|w_l \odot (\phi_l(I)_{hw} - \phi_l(\hat{I})_{hw})\|_2^2 \quad (18)$$

where:

- $\phi_l(\cdot)$: feature maps at layer l from a pre-trained network (e.g., VGG).
- H_l, W_l : spatial dimensions of the feature map at layer l .
- w_l : learned linear weights that calibrate channel

importance.

- \odot : element-wise multiplication.

A lower LPIPS score indicates that the reconstructed image is perceptually more similar to the reference, aligning more closely with human subjective evaluations.

The LPIPS results, which measure perceptual similarity, assess the alignment of model outputs with the human visual system. Lower LPIPS values indicate that the scenes are reconstructed in a more natural and visually pleasing manner. The corresponding comparisons are presented in Figure 4.

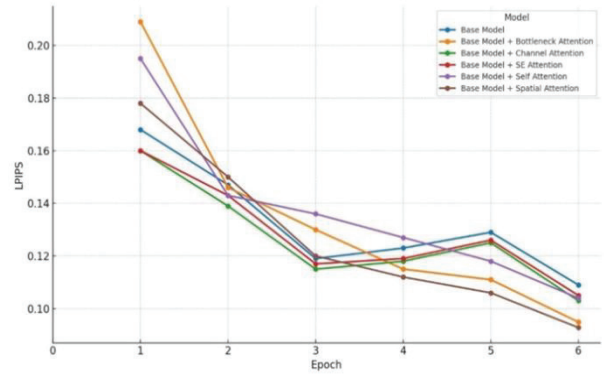


Figure 4. Comparative graph of average LPIPS scores across model variants.

The results show that integrating attention mechanisms significantly improves model performance across all metrics. Among them, the spatial attention module achieved the highest scores in PSNR, SSIM, and LPIPS. This indicates the model's ability to produce HDR scenes that are both structurally accurate and perceptually faithful. PSNR confirms high reconstruction quality, and its alignment with LPIPS shows that this also reflects improved visual realism. To evaluate the effectiveness of the proposed architecture, HDR-AttNet was compared both qualitatively and quantitatively with several state-of-the-art HDR reconstruction methods.

The comparative analysis indicates that HDR-AttNet consistently delivers superior performance in terms of visual fidelity and structural consistency. Unlike traditional approaches (e.g., HDRCNN, HDR-DANet, HDRUNet, etc.), the integration of attention mechanisms within HDR-AttNet enables more precise feature selection and improved contextual modeling, which leads to better preservation of fine textures and perceptual details. As a result, the model produces more balanced and clearer HDR reconstructions, particularly in challenging regions affected by overexposure or deep shadows. These findings highlight the strength of the proposed attention-driven design and confirm its advantages over baseline methods in high dynamic range image synthesis tasks.

4.3. Ablation Studies

In this section, the impact of the five different attention mechanisms integrated into the proposed HDR-AttNet architecture is comparatively analyzed. Each attention mechanism was individually embedded into the base model to create distinct model variants. These variants were evaluated

using three objective metrics: PSNR, SSIM, and LPIPS. The results of the comparative analysis are presented in Table 3.

Table 3. Comparison of model variants based on PSNR, SSIM, and LPIPS performance metrics

Model	PSNR (\uparrow)	SSIM (\uparrow)	LPIPS (\downarrow)
Base Model	21.50	0.9130	0.1090
Base Model + Spatial Attention	22.30	0.9310	0.0928
Base Model + Channel Attention	<u>22.00</u>	0.9260	0.1030
Base Model + Bottleneck Attention	21.80	<u>0.9290</u>	<u>0.0949</u>
Base Model + SE Attention	21.80	0.9230	0.1050
Base Model + Self-Attention	21.20	0.9090	0.1040

Quantitative results across datasets and methods for PSNR, SSIM, and LPIPS. The best scores are bolded, and the second best scores are underlined.

The results indicate that the spatial attention mechanism delivers the highest performance in terms of both structural similarity (SSIM) and perceptual quality (LPIPS). In particular, the low LPIPS score demonstrates that this module preserves visual consistency more effectively. PSNR values further support this observation, showing that the spatial attention variant achieves the closest reconstruction to the reference HDR scenes. Although other attention mechanisms contribute positively in specific aspects, the spatial attention module consistently provides superior results across all evaluation metrics.

4.4. Visual Comparison

In addition to the metric-based evaluation, the visual quality of HDR images produced by each model variant was also examined. Figure 5 presents the HDR outputs produced by different attention-based model variants alongside their corresponding LDR inputs. Visual enhancements can be clearly observed across various regions of the scenes. For instance, in the first row, the Spatial Attention variant (column c) preserves sharper foliage textures and improves shadow detail, offering more natural transitions compared to the baseline and other variants. In contrast, the Self-Attention model (column g) in the third row appears to underperform in reconstructing cloud structures, resulting in a flatter and less dynamic sky. In the fourth row, the Channel Attention variant (column d) maintains well-defined road edges and smooth luminance gradients, contributing to better color stability. Similarly, in the fifth row, the SE-integrated model (column f) achieves balanced illumination over wooden textures while avoiding overexposed highlights and preserving fine detail.

These visual findings align with the overall qualitative and quantitative performance trends observed in earlier sections, reinforcing the impact of attention mechanisms on perceptual quality. While channel attention demonstrates strong performance in color balance, the self-attention module delivers superior results in maintaining global scene consistency. The preservation of visual details is valuable not only from a numerical metrics perspective but also in terms of

human perceptual quality, underlining the model’s potential for practical deployment.

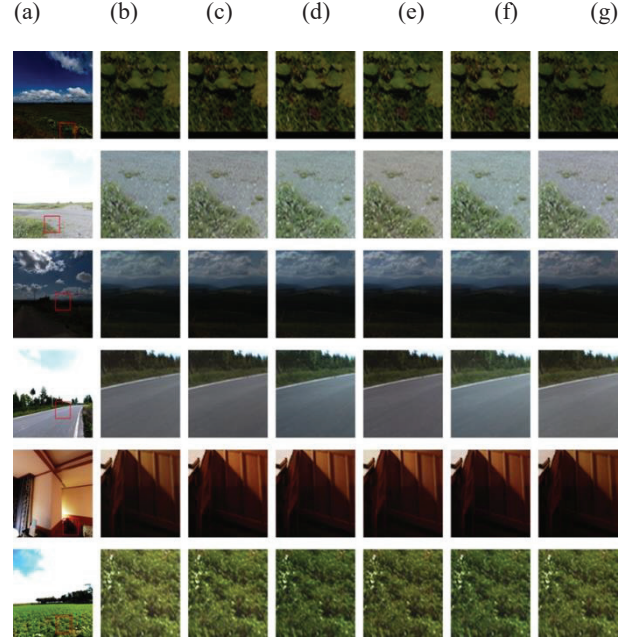


Figure 5. Comparative visualization of HDR outputs from different model variants, highlighting visual quality differences. Each column corresponds to: (a) Original LDR input, (b) Base Model, (c) Base + Spatial Attention, (d) Base + Channel Attention, (e) Base + Bottleneck Attention, (f) Base + SE Attention, and (g) Base + Self Attention. The first row displays the full-sized images with selected regions marked by red rectangles, while the second row presents the cropped regions corresponding to those selections.

5. Discussion

In this section, the performance of the HDR-AttNet architecture under five different attention configurations is comprehensively discussed based on both quantitative metrics and qualitative visual analysis. Each attention mechanism contributed to the overall model performance in distinct ways.

The Spatial Attention variant demonstrated the most consistent and superior performance across all evaluation metrics. It achieved the highest SSIM (0.9310), PSNR (22.30 dB), and the lowest LPIPS (0.0928), indicating its strong ability to preserve edges, textures, and fine details. Its localized focus helped reinforce salient regions without compromising overall scene structure.

The Channel Attention module exhibited strong performance in terms of color consistency and brightness calibration. Especially in scenes with high color saturation, it maintained perceptual clarity by re-weighting feature channels based on semantic importance.

The Squeeze-and-Excitation (SE) mechanism improved the model’s channel-wise feature recalibration, resulting in stable luminance mapping under diverse lighting conditions. It maintained above-average performance across all metrics.

The Bottleneck Attention variant contributed by reducing redundant feature dimensions, which improved computational efficiency and reduced noise. However, its gains in fine-

grained detail and visual sharpness were relatively limited compared to spatial or channel-based approaches.

Although the Self-Attention mechanism theoretically excels at modeling long-range dependencies, its performance in this context was comparatively weaker. This may be due to the relatively shallow depth of the encoder-decoder architecture, which limits the scope for capturing global relationships. The wide context focus of self-attention may weaken essential local structures needed for quality HDR reconstruction. As a result, it underperformed in SSIM (0.9090) and LPIPS (0.1040) metrics, highlighting the importance of local attention in tasks requiring fine spatial fidelity.

5.1. Limitations and Future Directions

While the HDR-AttNet architecture demonstrates high-quality HDR reconstruction through the integration of various attention mechanisms, there remain certain areas for potential enhancement. Since the model was trained exclusively on synthetic HDR data (DrTMO), its generalizability to real-world scenarios may be limited. Future work could benefit from incorporating datasets that include real HDR scenes for improved robustness. Additionally, training five separate attention-based variants independently has increased computational cost. This may require careful consideration in terms of scalability. However, this overhead can potentially be reduced by employing hybrid or lightweight attention modules. Lastly, increasing the depth of the encoder-decoder structure and integrating Transformer-based components may offer opportunities to enhance global context modeling in future architectural designs.

6. Conclusion

In this study, we introduced HDR-AttNet, a novel multi-branch encoder-decoder architecture tailored for single-image HDR reconstruction. The proposed model integrates three parallel subnetworks HDR Encoding Net, Up-Exposure Net, and Down-Exposure Net which collectively emulate the inverse exposure process to recover lost information due to overexposure or underexposure. In addition, five distinct attention mechanisms were independently incorporated to investigate their influence on visual fidelity and structural consistency. This modular integration allowed for a detailed comparative analysis and revealed that spatial attention, in particular, yielded the most visually coherent and perceptually faithful results. The outcomes of this work underscore the utility of attention mechanisms beyond classification or detection tasks, highlighting their relevance in complex image reconstruction pipelines. The attention-driven design adopted in HDR-AttNet contributes to enhanced detail preservation, improved scene balance, and more realistic HDR synthesis. Overall, the results validate our method and offer insights for designing future HDR systems that balance quality, efficiency, and modularity.

7. References

- [1] P.-H. Le, Q. Le, R. Nguyen, and B.-S. Hua, "Single-image HDR reconstruction by multi-exposure generation," in Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV), 2023, pp. 4063–4072.
- [2] Z. Liu, Y. Wang, B. Zeng, and S. Liu, "Ghost-free high dynamic range imaging with context-aware transformer," in Proc. Eur. Conf. Comput. Vis. (ECCV), Cham, Switzerland: Springer, Oct. 2022, pp. 344–360.
- [3] Y.-L. Liu, W.-S. Lai, Y.-S. Chen, Y.-L. Kao, M.-H. Yang, Y.-Y. Chuang, and J.-B. Huang, "Single-image HDR reconstruction by learning to reverse the camera pipeline," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2020, pp. 1651–1660.
- [4] X. Chen, Y. Liu, Z. Zhang, Y. Qiao, and C. Dong, "HDRUNet: Single image HDR reconstruction with denoising and dequantization," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2021, pp. 354–363.
- [5] K.-Y. Chen and J.-J. Leou, "Low light image enhancement using autoencoder-based deep neural networks," in Image Processing, Computer Vision, and Pattern Recognition and Information and Knowledge Engineering, Las Vegas, NV, USA: Springer, 2025, pp. 73–84. doi: 10.1007/978-3-031-85933-5_6.
- [6] A. Sharif, S. M. Naqvi, M. Biswas, and S. Kim, "A two-stage deep network for high dynamic range image reconstruction," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2021, pp. 550–559.
- [7] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2021, pp. 6881–6890.
- [8] Q. Yan, T. Hu, G. Chen, W. Dong, and Y. Zhang, "Boosting HDR image reconstruction via semantic knowledge transfer," arXiv preprint arXiv:2503.15361, 2025.
- [9] J. Ma and H. Zhang, "HDR-DANet: Single HDR image reconstruction via dual attention," *Multimedia Syst.*, vol. 31, no. 1, pp. 1–12, 2025.
- [10] S. Y. Kim, J. Oh, and M. Kim, "JSI-GAN: GAN-based joint super-resolution and inverse tone-mapping with pixel-wise task-specific filters for UHD HDR video," in Proc. AAAI Conf. Artif. Intell., vol. 34, no. 7, Apr. 2020, pp. 11287–11295.
- [11] Y. I. Park, J. W. Song, and S. J. Kang, "65 - 3: Invited paper: Deep learning - based image enhancement for HDR imaging," in *SID Symp. Dig. Tech. Papers*, vol. 53, no. 1, Jun. 2022, pp. 865–868.
- [12] A. de Santana Correia and E. L. Collobini, "Attention, please! A survey of neural attention models in deep learning," *Artif. Intell. Rev.*, vol. 55, no. 8, pp. 6037–6124, 2022.
- [13] J. Shen and T. Wu, "Learning spatially-adaptive squeeze-excitation networks for few-shot image synthesis," in Proc. IEEE Int. Conf. Image Process. (ICIP), Oct. 2023, pp. 2855–2859.
- [14] Z. Zhang, Z. Xu, X. Gu, and J. Xiong, "Cross-CBAM: A lightweight network for scene segmentation," arXiv preprint arXiv:2306.02306, 2023.

- [15] S. Mekruksavanich and A. Jitpattanakul, “Hybrid convolution neural network with channel attention mechanism for sensor-based human activity recognition,” *Sci. Rep.*, vol. 13, no. 1, p. 12067, 2023.
- [16] A. Srinivas, T.-Y. Lin, N. Parmar, J. Shlens, P. Abbeel, and A. Vaswani, “Bottleneck transformers for visual recognition,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 16519–16529.
- [17] H. Y. Lin, Y. R. Lin, W. C. Lin, and C. C. Chang, “Reconstructing high dynamic range image from a single low dynamic range image using histogram learning,” *Appl. Sci.*, vol. 14, no. 21, p. 9847, 2024.
- [18] H. Zhang, K. Zu, J. Lu, Y. Zou, and D. Meng, “EPSANet: An efficient pyramid squeeze attention block on convolutional neural network,” in *Proc. Asian Conf. Comput. Vis. (ACCV)*, 2022, pp. 1161–1177.
- [19] B. Zhang, S. Gu, B. Zhang, J. Bao, D. Chen, F. Wen, et al., “StyleSwin: Transformer-based GAN for high-resolution image generation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 11304–11314.
- [20] Y. Al Najjar, “Comparative analysis of image quality assessment metrics: MSE, PSNR, SSIM and FSIM,” *Int. J. Sci. Res. (IJSR)*, vol. 13, no. 3, pp. 110–114, 2024.
- [21] M. Azimi, “PU21: A novel perceptually uniform encoding for adapting existing quality metrics for HDR,” in *Proc. Picture Coding Symp. (PCS)*, Jun. 2021, pp. 1–5.
- [22] A. Ghildyal and F. Liu, “Shift-tolerant perceptual similarity metric,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Cham, Switzerland: Springer, Oct. 2022, pp. 91–107.

Özgeçmişler



Cevher Renk received his B.Sc. degree in Computer Engineering from Harran University in 2020 and has been pursuing an M.Sc. degree in the same department since 2022. Before his academic career, he gained nearly four years of professional experience as a full-stack developer in the private sector, where he effectively used .NET Core and Spring Boot for back-end development, and React.js and Next.js for front-end development. His professional contributions included e-commerce platforms, CRM solutions, enterprise systems, and AI-based applications.

Since 2024, he has been working as a Research Assistant at Harran University, where he focuses on image processing, deep learning, and artificial intelligence-based methods for computer vision. His research interests include high dynamic range (HDR) image reconstruction and attention-enhanced autoencoder architectures. Throughout both his industry and academic career, he has demonstrated strong teamwork skills, a passion for learning, and a research-driven mindset.



Serdar Çiftçi received the B.Sc. degree in computer engineering from Selçuk University in 2007 and the M.Sc. and Ph.D. degrees in computer engineering from Middle East Technical University in 2011 and 2017, respectively. He is an Assistant Professor with the Department of Computer Engineering, Harran University, Türkiye. During his Ph.D., he was a Visiting Researcher at the École Polytechnique Fédérale de Lausanne (EPFL), Switzerland, in 2015, and later at the Rochester Institute of Technology (RIT), USA, in 2022. From 2018 to 2021, he was the Department Chair of the Department of Computer Engineering, Harran University. His research interests focus on computer vision and deep learning. He is a member of the Chamber of Computer Engineers of Turkey and served as the Vice President of its Executive Board from 2012 to 2014. His doctoral work was recognized with the Thesis of the Year Award by the Parlar Foundation in 2017.