SDRTV-to-HDRTV via Hierarchical Dynamic Context Feature Mapping

Gang He
Xidian University,
Kuaishou Technology
ghe@xidian.edu.cn

Kepeng Xu*
Xidian University
kepengxu11@gmail.com

Li Xu
Xidian University
cherylxu@stu.xidian.edu.cn

Chang Wu
Xidian University
Jackwu0630@outlook.com

Ming Sun
Kuaishou Technology
sunming03@kuaishou.com

Xing Wen
Kuaishou Technology
td.wenxing@gmail.com

Yu-Wing Tai
Kuaishou Technology
yuwing@kaist.ac.kr

* Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ABSTRACT

In this work, we address the task of SDR videos to HDR videos (SDRTV-to-HDRTV conversion). Previous approaches use global feature modulation for SDRTV-to-HDRTV conversion. Feature modulation scales and shifts the features in the original feature space, which has limited mapping capability. In addition, the global image mapping...
cannot restore detail in HDR frames due to the luminance differences in different regions of SDR frames. To resolve the appeal, we propose a two-stage solution. The first stage is a hierarchical Dynamic Context feature mapping (HDCFM) model. HDCFM learns the SDR frame to HDR frame mapping function via hierarchical feature modulation (HME and HM) module and a dynamic context feature transformation (DYCT) module. The HME estimates the feature modulation vector, HM is capable of hierarchical feature modulation, consisting of global feature modulation in series with local feature modulation, and is capable of adaptive mapping of local image features. The DYCT module constructs a feature transformation module in conjunction with the context, which is capable of adaptively generating a feature transformation matrix for feature mapping. Compared with simple feature scaling and shifting, the DYCT module can map features into a new feature space and thus has a more excellent feature mapping capability. In the second stage, we introduce a patch discriminator-based context generation model PDCG to obtain subjective quality enhancement of over-exposed regions. The proposed method can achieve state-of-the-art objective and subjective quality results. Specifically, HDCFM achieves a PSNR gain of 0.81 dB at about 100K parameters. The number of parameters is 1/14th of the previous state-of-the-art methods. The test code will be released on https://github.com/cooperlike/HDCFM.

1 INTRODUCTION

High Dynamic Range (HDR) video allows for a more realistic display and reproduction of the natural world. HDR video has a higher bit depth, wider color range, and higher brightness per pixel. Although HDR display device technology is now widely available and HDR video has many advantages, most video sources are still stored in Standard Dynamic Range (SDR) format. Therefore, converting many existing SDR videos to HDR videos can dramatically improve the user experience.

We present the process of SDR/HDR video content acquisition and playback in Fig.2. From the moment light enters the camera to the playback of the video image using the monitor, it goes through the following stages: 1) convert the light signal into a digital signal through CMOS; 2) reduce the high dynamic digital signal to a low dynamic signal by Tone Mapping [29]; 3) transfer the image color to the target color gamut by Gamut mapping; 4) convert linear signal to nonlinear signal by OETF [1, 33]; 5) quantize digital signal and arithmetic coding [35]; 6) encode and decode by codec [8, 40]; 7) convert the decoded nonlinear signal to a linear digital signal by EOTF; 8) convert the linear signal to an optical signal for playback. The main difference between SDR and HDR is using different EOTF (Electro-Optical Transfer Function) and OETF (Optical-Electro Transfer Function).

To distinguish SDR video to HDR video from the SDR image to HDR image task, we follow the definition of method [7] and define SDR video to HDR video as SDRTV-to-HDRTV conversion. LDR-to-HDR (LDR image to HDR image) refers to the conversion of SDR image to HDR image. HDR images can play on display devices by tone mapping. The previous approach [20, 22] united super-resolution with SDRTV-to-HDRTV conversion, and tried to build a pipeline from low-resolution SDR video to high-resolution HDR video. HDRTVNET [7] proposed a multi-stage scheme to complete SDRTV-to-HDRTV conversion by global tone mapping, local image enhancement, and image generation.

In the SDRTV-to-HDRTV conversion, the most crucial issue is to map the SDR features to the HDR feature domain, which is called feature mapping in this paper. The second issue is generating information on over-exposure areas that do not exist in SDR. An SDR video to HDR video conversion pipeline is constructed to address these two issues. The pipeline is divided into two parts, Hierarchical Dynamic Context Feature Mapping model (HDCFM) and a Patch discriminator-based Dynamic Context Generation network (PDCG). The first part obtains HDR frames with superior objective quality by feature mapping, and the second part accomplishes over-exposure area image enhancement.

Specifically, HDCFM contains the Hierarchical feature Modulation vector Estimation (HME) module, Hierarchical Modulation (HM) module, and Dynamic Context feature Transformation (DYCT) module. For HME, we construct a hierarchical modulation vector estimation module that captures the global and local image prior to estimating the global and local feature modulation vectors. For HM, the global and local modulation vectors estimated by HME are used to modulate the input features. Such feature modulation enables adaptive mapping of local images in different regions of different frames. For DYCT, we propose the joint context local feature transformation module to extract image context information.
and accomplish local feature transformation by dynamic convolution. HDCFM can complete feature mapping and obtain HDR frames based on the above structure. The proposed HDCFM has two advantages over the previous methods. The first is that the proposed HM and HME can perform spatially adaptive mapping using image local information. In addition, the proposed DYCT module models a more robust dynamic feature transformation: the ability to transform features directly to a new feature space instead of the previous simple feature scaling and shifting. This dramatically enhances the mapping performance of the model. Thus, a more complex mapping process can be modeled to map SDR frames to HDR frames better. For PDCG, a Patch GAN with an over-exposure mask is proposed, which can generate over-exposure region image information to obtain the higher subjective quality of HDR frames. The proposed HDCFM with a smaller number of parameters can obtain the best conversion performance. In order to compare with previous methods, we selected five evaluation metrics PSNR, SSIM, SR-SIM[20], ΔE_{ITT} and HDR-VDP3[30] to evaluate the proposed method.

In summary, our contributions include the following main points.

- We propose a hierarchical feature modulation module that can perform spatially adaptive feature modulation on image features; local feature modulation can improve the quality of HDR reconstructed frames.
- We propose a dynamic feature transformation method that can further improve the feature mapping capability of the model to obtain higher quality HDR converted frames.
- We analyzed the problem of over-exposure in the SDRTV-to-HDRTV conversion. Propose a Patch discriminator-based over-exposure image generation model that can obtain a higher subjective quality HDR frame.
- With about 100K parameters, the proposed method can obtain state-of-the-art results compared to previous methods.

2 RELATED WORK

Converting previous SDR videos to HDR videos is a valuable task. More and more researchers are focusing on this topic. There are several main methods for SDR to HDR conversion as follows. 1) Multi-exposure SDR images to single-frame HDR images. 2) single-frame SDR image to single-frame HDR image. 3) SDR video to HDR video. Our goal is to convert SDR video that already existed to HDR video.

LDR-to-HDR. The traditional method estimates the light source density, based on which the dynamic range is further expanded [2–4, 31]. Researchers have proposed a method based on deep convolutional neural network [27] to convert LDR images to HDR images directly. HCRCNN[10, 28, 37] propose method that can recover the over-exposure area of the image. [9, 25, 25, 34, 44] proposed methods can predict multi-exposure LDR image pairs by a single LDR image, then synthesize HDR images based on the predicted multi-exposure image pairs.

SDRTV-to-HDRTV conversion. The SDRTV-to-HDRTV conversion approach has only emerged in the last two years. [20] proposes a GAN-based architecture that jointly achieves super-resolution and SRTV to HDRTV. [22] proposes a hierarchical GAN architecture to accomplish super-resolution and SDRTV to HDRTV.

![Figure 3: The framework of the proposed method. Firstly, a Hierarchical Dynamic Context Feature Mapping model(HDCFM) Convert Standard Dynamic Range(SDR) frame X_S to High Dynamic Range(HDR) frame X_{HR}. Secondly, we propose a Patch discriminator-based Dynamic Context Generation model(PDCG).](image)

[7] proposed a method using global feature modulation, local enhancement, and over-exposure compensation, which achieved the best performance.

Dynamic Convolution. The vanilla convolutional layer learns the parameters of the convolutional kernel through big data during training. The parameters of the convolution kernel are fixed during the inference phase and do not change for different inputs; such a convolution kernel is also called a static convolution kernel. Dynamic convolution, meaning that the convolution kernel parameters are dynamically updated during the inference phase, allows the model to extract more complex features and build complex pattern recognition methods. Current researchers focus on how to construct dynamic convolution kernels [5, 46].

Context Convolution. In convolutional neural networks, context extraction aims to extract the correlation between the current location features and the global features, thus improving the modeling capability. To extract non-local information from images [43] proposes a non-local generic module to complement the long-range dependencies to capture the global context information. [6] finds that the global relevance information obtained from different locations during non-local modeling is almost the same. Therefore, a generic feature aggregation module is proposed to extract global feature information directly instead of global relevance for each feature element. Such a modeling approach dramatically reduces the computational cost and improves the feature extraction performance.

Patch GAN. The Generative Adversarial Networks (GAN) model [12, 13, 19, 32, 41, 42] has been widely used in the field of image generation in recent years, thanks to its unique architecture design. Patch GAN[17] improves the clarity of the generated images, enabling higher resolution images. In this paper, the Patch discriminator can solve the problem of low model performance caused by the low proportion of over-exposure region.

3 METHODOLOGY

3.1 Framework

SDR frames to HDR frames can be modeled as a feature mapping and feature complement process. For this purpose, we propose a two-stage solution. The first stage converts SDR frames X_S to HDR frames X_{HR} by a Hierarchical Dynamic Context Feature Mapping model HDCFM. The second stage uses a Patch discriminator-based Dynamic Context Generation model PDCG to complete the over-exposure enhancement and obtain HDR frames X_{HG} with higher resolution.
subjective quality. The framework of the whole scheme is shown in Fig.3 and Formula (1). The main challenge of the SDRTV-to-HDRTV conversion is that the data distribution of SDR frames differs significantly from that of HDR frames and that SDR frames store less information (there are over-exposure problems). $M_H$ is the over-exposure area mask calculated similarly as [38]. The specific motivation and architecture of each module will be described next.

$$X_{HR} = HDCFM(X_S)$$
$$X_{HG} = PDCG(X_{HR}, M_H)$$ (1)

Figure 5: The architecture of Hierarchical Modulation (HM). The global and local feature modulation tandem constitutes the HM.

3.2 HDCFM

In this part, to transform SDR frame $X_S$ to HDR frame $X_{HR}$, we construct HDCFM, which can achieve the instance-space adaptive feature mapping method in the image feature space. Specifically, HDCFM can perform different feature mapping for different input and pixels at different spatial locations in the input image. A locally adaptive image feature mapping model is constructed, which can obtain high quality HDR frames.

The critical point 1 of HDCFM is the hierarchical feature modulation vector estimation module HME and the hierarchical feature modulation module HM, which can accomplish hierarchical global and local feature modulation. Essentially, the SDR image features are scaled and shifted in the feature space to obtain features closer to HDR images. The critical point 2 of HDCFM lies in the dynamic feature transformation module DYCT of context features, which can accomplish dynamic feature transformation (matrix transformation). The structure of HDCFM is shown in Fig.4. The motivation and methodology of these two points will introduce next.

3.2.1 Motivation of Feature Mapping. The feature extracted from SDR video frame is in SDR feature space, and the feature extracted from HDR video frame is in HDR feature space, so SDRTV-to-HDRTV conversion can be modelled as a feature mapping. For the input SDR frame $X_S$, the low dynamic feature $F_S$ is first obtained by convolution, and then $F_S$ needs to be mapped to the high dynamic feature $F_H$, and finally $F_H$ is recovered to the image space. The feature mapping proposed in this paper consists of two parts, which are hierarchical feature modulation and local feature transformation.

3.2.2 Motivation of HM. During SDRTV-to-HDRTV conversion, pixels at different spatial locations need to be processed differently. For example, in one frame, there are both over-exposed and under-exposed areas, then different processing should be performed on the both under-exposed and over-exposed image area. To address this problem, we design HM composed of global feature modulation and local feature modulation. The global feature modulation can make macro adjustments to the image, and the local feature modulation can further complete the local fine-tuning.

3.2.3 Architecture of HM. To obtain spatially adaptive feature modulation vectors, our HDCFM constructs a joint global and local hierarchical modulation vector estimation module HME. HME can predict not only the global feature modulation vectors $V_{Ga}$, $V_{Gb}$, but also the local feature modulation vectors $V_{La}$, $V_{Lb}$; this gives the ability to perform different feature modulations on image features at different spatial locations. The structure of HM is shown in Fig.5. Then comes calculating the feature modulation vector by the HM module. Specifically, for the input $X_S$, $F_{DS}$ is obtained after five downsamples. $F_{DS}$ goes through the global downsample to obtain $V_{Ga}$ and $V_{Gb}$, $F_{DS}$ passes through the upsample to obtain $V_{La}$ and $V_{Lb}$; such a calculation process is shown in Fig.6. Next, the feature modulation of $F_S$ is performed using HM. The HM is divided into two steps. The first step uses the global modulation parameter $V_{Ga}$ to dot-multiply $F_S$, followed by adding the features after point multiplication using $V_{Gb}$. The second step uses the local modulation parameter $V_{La}$ to dot-multiply $F_S$, followed by $V_{Lb}$ to sum the dot-multiplied features.

3.2.4 Motivation of the DYCT Module. The core of DYCT (Dynamic Context Feature Transformation module) is feature transformation. Firstly, we introduce the difference between feature modulation and feature transformation. Feature modulation: The modulated feature layer is obtained by dot multiplying the feature layer $F_S$ by the modulation vector, a simple feature mapping that...
only scales and shifts the features in the original feature space. So it can model a relatively single mapping process. This modeling process is shown in Fig.5. **Feature transformation**: feature transformation (matrix multiplication) of the input features \( F_S \) can transform \( F_S \) from the original feature space to the new feature space. This modeling approach has a much more robust feature mapping capability. Therefore, we propose to build a feature mapping module based on feature transformation. It can model a more complex feature mapping process and thus able to obtain better HDR transformed frames.

### 3.2.5 Modeling Process of the DYCT Module

We begin with an introduction to the local feature transformation module, as shown in the left part of Fig.7, for the local feature \( F_P \) (with shape \((K, K, C_I)\)) of input \( F_S \), is flattened to a vector \( F_T \) of \((K \cdot K \cdot C_I, 1)\). At the same time, we need a conditional generation module \( C_T \) to predict the parameters \( K_T \) of the feature transformation (in shape \((C_O, K \cdot K \cdot C_I)\)). Next, \( K_T \) is subjected to a matrix multiplication operation with \( F_T \), which is called the linear feature transformation, and finally, the transformed feature \( O_{i,j} \) (with shape \((C_O, 1)\)) is obtained.

To further analyze the process of feature transformation, we can use the local feature transformation in convolution specific implementation form. As shown in the right Fig.7, for the input local feature \( F_P \), \( C_O \) convolution kernels \( K_T \) of shape \((K, K, C_I)\) are needed to convolve with \( F_P \). The output is \( O_{i,j} \), where \( K_T \) is generated online by the \( C_T \) module when inference is needed. \( C_I \) is the number of channels in the input feature layer. We continue to analyze the generation of \( K_T \). In a practical application, the resolution of the input image is \((H \cdot W)\), and the size of the input feature layer of the local feature transform is \((H, W, C_I)\). For each pixel, \( C_O \) convolution kernels are predicted, and the shape of each convolution kernel is \((K, K, C_I)\). The number of parameters for all convolution kernels is \((H \cdot W \cdot C_O \cdot K \cdot K \cdot C_I)\). The total number of parameters in the 4K image processing task is \(3.057 \times 10^{11} \), directly leading to memory out.

\[
K_S = SKP(F_S) \\
K_C = CKP(F_S) \\
F_{mid} = DDP(F_S, K_S, K_C) \\
F_O = CB(F_{mid}) \\
CKP = GAP \bullet Conv \\
SKP = Conv \bullet Conv \\
CB = SpatialConv \bullet ChannelConv
\]  

(2)

### 3.2.6 Architecture of the DDCN Module

This paragraph will introduce the specific architecture of the DDCN. To solve memory out, we borrow the idea of decoupled dynamic convolution kernels. The architecture of the whole DDCN module is shown in Fig.8. The specific process is as follows. Decompose the original \( C_O \) convolution kernel \( K_T \) into a combination of spatial convolution kernel \( K_S(K, K, H, W) \) and channel convolution kernel \( K_C(C, K, K) \). \( K_S \) and \( K_C \) are generated by \( SKP \) and \( CKP \) respectively, the computation process of \( SKP \) and \( CKP \) is defined in Formula(2). After obtaining \( K_S \) and \( K_C \), the output feature layer \( F_{mid} \) is obtained by convolving through the decoupling convolution method DDF proposed by [46]. During the convolution of dynamic filters, the convolution kernel weights are obtained in real-time by sample inference. This can enhance the feature mapping capability of the model. This is a convolutional implementation of the feature transformation.
Table 1: Quantitative comparisons. Red text indicates the best and blue text indicates the second. The result of the previous methods in the table are consistent with [7].

| Methods                | Params[| | PSNR↑ | SSIM↑ | SR-SIM↑ | ΔEITP ↓ | HDR-VDP3↑ |
|------------------------|-------|-------|-------|-------|---------|---------|-----------|
| HuoPhyEO[16]TVC       | -     | 25.90 | 0.9296| 0.9981| 38.06   | 7.893   |
| Kovaleski[24]SRCRAPI  | -     | 27.89 | 0.9273| 0.9809| 28.00   | 7.431   |
| ResNet[15]ECCV16      | 1.37M | 37.32 | 0.9720| 0.9950| 9.02    | 8.391   |
| Pixel2Pixel[18]CVPR17 | 11.38M| 25.80 | 0.8777| 0.9871| 44.25   | 7.136   |
| CycleGAN[47]ICCV17    | 11.38M| 21.33 | 0.8496| 0.9595| 77.74   | 6.941   |
| HDRNET[11]TOG         | 482K  | 35.73 | 0.9664| 0.9957| 11.52   | 8.462   |
| CSRNET[14]ECCV20      | 36K   | 35.04 | 0.9625| 0.9955| 14.28   | 8.400   |
| Ada-3DLUT[45]TPAMI    | 594K  | 36.22 | 0.9658| 0.9967| 10.89   | 8.423   |
| Deep SR-ITM[21]ICCV19 | 2.87M | 37.10 | 0.9686| 0.9950| 9.24    | 8.233   |
| JSI-GAN[23]AAAI20     | 1.06M | 37.01 | 0.9694| 0.9928| 9.36    | 8.169   |
| HDRTVNET[7]ICCV21     | 1.41M | 37.61 | 0.9726| 0.9967| 8.89    | 8.613   |
| HDCFM(Proposed)       | 100.63K | 38.42 | 0.9732| 0.9974| 7.83    | 8.5716  |

It is worth noting that only local feature transformations may lead to large differences in the transformed results for image contents that have the same color in different regions. Therefore, this paper uses a context module to count the overall feature context information and fine-tune the features. In this part, we choose the context information module CB proposed by [6], input \( f_{mid} \) to CB, and the output result is the final output of the DYCT module. The structure of SpatialCono and ChannelCono is similar to [6], using DDF convolution instead of vanilla convolution. This module has the advantage of being computationally simple, and the computation procedure is given in Formula(2).

$$X_{HG} = PDCG(X_{HR}, M_H) \bullet M_H + X_{HR} \bullet (1 - M_H)$$  \hspace{1cm} (3)

$$L_{HG} = \alpha L_1 + \beta L_p + \gamma L_{GAN}$$  \hspace{1cm} (4)

3.3 PDCG

Figure 9: The architecture of Patch discriminator Dynamic Context Generation model (PDCG).

3.3 PDCG

The objective quality of \( X_{HR} \) generated by HDCFM is very high, but \( X_{HR} \) still suffers from the problem of missing over-exposure information. The problem of loss of over-exposure information can cause large bright spots or even wrong colors in HDR reconstructed frames. To be able to generate over-exposure region image information, we propose the PDCG model, the structure of PDCG is shown in Fig. 9. The generation method of the over-exposure section is formally defined in Formula(3), and the mask of the over-exposure section is defined as \( M_H \). We input the preliminary results \( X_{HG} \) generated by HDCFM into the PDCG model. After three convolutions with a stride of 2, the resolution of the feature layer is reduced, and \( F_{d1}, F_{d2}, F_{d3} \) are obtained. \( F_{d3} \) is input into 16 blocks in series. Each block contains a DYCT module, a vanilla convolution, and a skip connection. Then perform an upsampling to obtain \( F_{u1} \), add \( F_{u1} \) and \( F_{d2} \), and continue to upsample twice to obtain the final \( X_{HG} \).

4 EXPERIMENT

4.1 Experiment Settings

Dataset. For a fair comparison with previous methods, we use the dataset used by [7] captured HDR and SDR versions of each video; each HDR video was HDR10 with BT.2020 color gamut. Frames from the videos were extracted using FFMPEG, cropping the images to 480x480 size. 117 pairs of unduplicated images were included in the test set, each at 4K in size.

Training Setup. In the model’s training process, we use L1 as the loss function to optimize the HDCFM model. The Adam optimizer is used, the initial learning rate is set to 0.0005, and the learning rate is set to 1/2 of the initial rate every 200000 iterations; the total number of iterations is set to 1000000.

Evaluation Setup. To verify the effectiveness of the proposed method, we evaluate the effectiveness of the proposed method on the evaluation index of PSNR, SSIM, SR-SIM, \( \Delta E_{ITP} \) and HDR-VDP3.

4.2 Experiment Results

Qualitative and Quantitative Results. We first compare our method in Table 1 to compare with other methods in PSNR, SSIM, SR-SIM, \( \Delta E_{ITP} \), HDR-VDP3. PSNR (Peak Signal to Noise Ratio) can
Figure 10: HDR reconstruction of frame results. The proposed method is able to reconstruct local color information in higher quality compared to previous state-of-the-art methods. The proposed method in the red box clearly achieves higher quality color conversion results.

Figure 11: Local results. The proposed HDCFM can perform adaptive feature mapping based on local information, thus effectively improving the quality of HDR reconstructed frames in overexposed regions.

measure the pixel value difference between images. SSIM (Structural Similarity Index Measure) evaluates the structural similarity of two images. SR-SIM is able to evaluate the image similarity of HDR standard images. \( \Delta E_{ITP} \) can evaluate the colour difference of HDRTV. HDR-VDP3 can assess image visual difference. Compared to HDR-VDP2, HDR-VDP3 supports the BT.2020 gamut. The test results were calculated on 117 images with a resolution of 2160x3840. We calculate HDR-VDP3 scores on linear HDR images. As Table.1, our method produces significantly better objective results, which indicates the ability of our network to accurately reconstruct HDR frames. To demonstrate the visual effect of the proposed method, we directly save the 16bit bit-depth image in PNG format, which is able to preserve all image information despite the fact that such a saving method will gray out the image. Another method is to convert 16bit to 8bit using tone mapping, but this conversion removes some overexposed areas, so the result of generating overexposed areas between different methods cannot be shown. The direct visualization method preserves all the details and the visualization results are displayed in Fig.1 and Fig.10. We also demonstrate in Fig.10 that our proposed method can construct adaptive feature mappings for localities, thus mitigating to some extent the quality degradation caused by overexposure. It can be seen that the previous method is unable to construct adaptive mapping for local images, and the HDCFM proposed in this paper is able to generate HDR frames with higher quality.

In addition, as shown in the lower right corner of Fig.11, HDCFM cannot perfectly recover the details of HDR frames when the input SDR image is overexposed due to tone mapping. Nevertheless, our results are still an improvement compared to previous state-of-the-art methods.

To further analyze the performance of the proposed method, we calculated the histogram of the generated HDR video frames. The pixel intensity distribution of the HDR video frames generated by the proposed method is closer to the ground truth, and the pixel intensity distribution is smoother. As shown in Table.1, the proposed model HDCFM outperforms the past method approach in all evaluation metrics. And the number of parameters of the HDCFM model is much smaller than that of the past method.

Ablation Study. We performed ablation experiments on the whole model to demonstrate the effectiveness of each module of the proposed method. Table 2 shows the performance of the SDRTV-to-HDRTV transformation after the addition of different modules. M0,
Figure 12: We analyze the histograms of HDR video frames generated by different methods. The left and right are the histograms of two different frames obtained by different methods. The histogram density is set to 72, and the proposed method can obtain smoother and more accurate histogram results compared to the previous methods.

Figure 13: Qualitative results of Patch discriminator-based Dynamic Context Generation model(PDCG). PDCG can generate over-exposure areas image with higher subjective quality, and the subjective quality of the sun and flame in the figure is improved.

| M0 | M1 | M2 | M3 | M4 | PSNR | SSIM | ΔEITP | SR-SIM |
|----|----|----|----|----|------|------|-------|--------|
| √  | ×  | ×  | ×  | ×  | 36.88 | 0.9655 | 9.78   | 0.9967 |
| √  | √  | ×  | ×  | ×  | 37.74 | 0.9705 | 8.85   | 0.9972 |
| √  | √  | √  | ×  | ×  | 38.20 | 0.9725 | 8.06   | 0.9972 |
| ×  | ×  | √  | √  | ×  | 38.26 | 0.9729 | 7.90   | 0.9973 |
| √  | √  | √  | √  | √  | 38.42 | 0.9732 | 7.83   | 0.9974 |

0.0074, respectively, and reduces ΔEITP by 1.88. After adding the combined local and global Hierarchical Local Feature Transform, the model can perform both global and local feature modulation, and the corresponding PSNR, SSIM is improved by 1.54 and 0.0077, respectively, and ΔEITP is reduced by 1.95.

The proposed PDCG module is capable of generating more realistic HDR reconstruction frames that are capable of generating subjective and comfortable images of over-exposure areas. In Fig.13 we show the comparison images of the results generated by the proposed method. In areas where the content is saturated and over-exposed (the sun part and the flame part), PDCG can address the existing artifacts and overexposure problems. Therefore, HDCFM uses the overexposed content for feature mapping, and the resulting HDR frame still has overexposure. PDCG can dynamically restore HDR frames in overexposed areas, so as to obtain HDR frames with higher subjective quality.

5 CONCLUSION

For standard dynamic range (SDR) to high dynamic range(HDR) video. Previous methods performed global conversion of HDR frames without taking into account local information and the quality of HDR frames in overexposed areas during conversion. In this paper, we proposed a two-stage SDRTV to HDRTV scheme to address these two problems. In the first stage, a feature mapping model is proposed. Proposed method can perform non-consistent mapping for image local information, and the proposed dynamic feature transformation module is able to simulate more complex feature mapping. The converted HDR frames have a higher objective quality. In the second stage, a patch discriminator and a context-based dynamic image generation model are constructed for overexposed areas. The patch discriminator can solve the problem that the model is difficult to train due to the low percentage of high-light areas. This model can improve the subjective quality of the reconstructed frames. Comprehensive experiments show that the proposed method achieves the best performance in both objective and subjective quality.

ACKNOWLEDGEMENT

This work was supported in part by the Innovation Fund of Xidian University (No. 5001-20109215456); and in part by the Fundamental Research Funds for the Central Universities.
