Deep Video Inverse Tone Mapping Based on Temporal Clues

Yuyao Ye[1],Ning Zhang[2], Yang Zhao[3], Hongbin Cao[1,4], Ronggang Wang[1]
[1]School of Electronic and Computer Engineering, Peking University [2]Baidu Netdisk
[3]School of Computer and Information, Hefei University of Technology [4]Bytedance Inc.
yeyuyao@pku.edu.cn zhangning34@baidu.com
yzhao@hfut.edu.cn caohongbin.181@bytedance.com rgwang@pkusz.edu.cn

Abstract

Inverse tone mapping (ITM) aims to reconstruct high dynamic range (HDR) radiance from low dynamic range (LDR) content. Although many deep image ITM methods can generate impressive results, the field of video ITM is still to be explored. Processing video sequences by image ITM methods may cause temporal inconsistency. Besides, they aren't able to exploit the potentially useful information in the temporal domain. In this paper, we analyze the process of video filming, and then propose a Global Sample and Local Propagate strategy to better find and utilize temporal clues. To better realize the proposed strategy, we design a two-stage pipeline which includes modules named Incremental Clue Aggregation Module and Feature and Clue Propagation Module. They can align and fuse frames effectively under the condition of brightness changes and propagate features and temporal clues to all frames efficiently. Our temporal clues based video ITM method can recover realistic and temporal consistent results with high fidelity in over-exposed regions. Qualitative and quantitative experiments on public datasets show that the proposed method has significant advantages over existing methods. The code is available at https://github.com/ye3why/VITM-TC/.

1. Introduction

Recently, high dynamic range (HDR) technology has elicited considerable interest due to its capability to provide more vivid visual experiences. But because of the lack of HDR content, it's highly demanded to convert existing low dynamic range (LDR) content to HDR. High dynamic range imaging (HDRI) methods use fusion algorithms to combine multi-exposure sequences and remove the ghost artifacts caused by misalignment. However, in most situations the multiple-exposure sequences are unavailable(e.g., single LDR images or videos on the Internet). Therefore, inverse tone mapping (ITM) methods are designed to estimate HDR radiance directly from single LDR content.

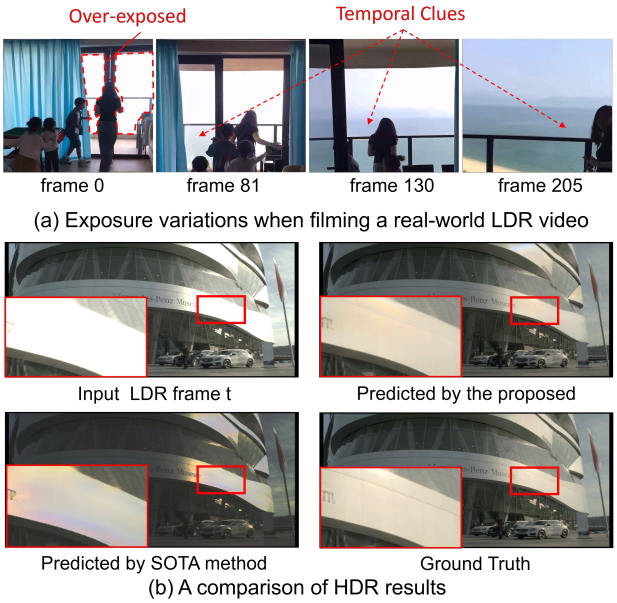

Figure 1. (a) A real-world LDR video on the Internet, the dynamic range of frame 0 is too wide so that details outside the window are lost. Fortunately, we can find clues along time axis. The textures from distant frames 81, 130, and 205 can be utilized to recover over-exposed regions. (b) An LDR video with the corresponding HDR ground truth from [9]. Compared to the SOTA image ITM method [22], the proposed method can use temporal clues to reconstruct more realistic textures for frame t.

With the development of deep learning, a growing number of image ITM methods have been proposed, where [7] [22] [26] [32] [38] learn an end-to-end model to recover HDR from LDR straightforwardly and [8] [14] [18] [19] [20] simulate the generation process of HDR images and estimate the multi-exposure stack. Compared to the highly sought-after image ITM, the field of video ITM has received little attention. Although The deep image ITM methods can generate impressive results in restoring lost textures, directly using it to process LDR videos may en-

counter unexpected issues. On the one hand, these methods can only estimate HDR radiance from spatial information of current frame, ignoring any potentially valid information in the temporal domain. On the other hand, it may also lead to temporal inconsistencies such as flicker. Xu et al. [37] propose the first deep video ITM method. However, they simply employ 3D convolution on a sequence of adjacent frames and still fail to completely handle flicker. The SDRTV-to-HDRTV conversion task [15] [36] [36] has attracted much attention, which aims to convert the standard dynamic range video into HDR for displaying on advanced televisions. However, most of them still focus on a single frame and assume there is no severe over-exposure problem in input content. In this paper, by analyzing the principles of camera imaging, we believe that temporal information can be investigated to reconstruct HDR videos with higher quality.Upon that, we propose a novel video ITM pipeline that can efficiently generate temporal consistent and realistic HDR videos.

The dynamic range in the real world is so wide that even the most advanced camera today can only record a part of that brightness range. In order to capture the most important information from a scene, cameras usually choose an appropriate exposure value based on the lighting conditions, the subject and so on. As shown in Fig. 1(a), when shooting the dim indoor scene, the exposure value is relatively high, resulting in a severe over-exposure outside the window. As the scene changes, the exposure value changes accordingly, and the details outside the window are revealed. Thus, it is possible that areas that are over-exposed in current frame are exposed normally in other frames, providing valid information for recovering lost textures in the current one. This phenomenon is named as temporal clues in this paper. The most straightforward ideas to leverage temporal clues are using temporal models, such as sliding windows [34], 3D convolution [37] or recurrent model [31]. However, such a exposure changing process tends to be designed to be very slow and smooth in order to ensure filmed video not flicker, which makes it difficult to find useful temporal clues. A short sequence of adjacent frames may not cover valid information while a long one may lead to a significant increase in computational complexity and also makes it more difficult to find and utilize temporal clues from long-term information.

In this paper, we propose a novel video ITM pipeline which can effectively and efficiently find and leverage temporal clues from a long sequence to recover natural textures in over-exposed areas with high fidelity. Specifically, we propose a Global Sampling and Local Propagation strategy, which firstly sample some reference frames from the whole input sequence with a large stride, and extract useful temporal clues from these reference frames to recover key frames. And then propagate information in key frames to their neighbors. In order to get meaningful temporal clues more effectively, we propose an Incremental Clue Aggregation Module which can fuse temporal clues with target key frames in a incremental way. In this module, we design a flow correction convolution to align frames with different luminance under ITM circumstances and we also utilize deformable convolution[5] and swin transformer block[23] to align local and global features respectively. After key frames are fused with temporal features, a BiRNN-like module, namely Feature and Clue Propagation Module is used to propagate features of these key frames to their adjacent frames while ensuring temporal consistency. Fig. 1(b) shows results from the state-of-the-art image ITM method [22] and the proposed method. [22] can not recover lost details by only using spatial information. On the contrary, with the temporal clues from the entire sequence, the proposed method can utilize more information and generate impressive results.In addition, due to the lack of HDR video datasets, we design a novel method to synthesize HDR video dataset based on available HDR images and LDR video datasets. Experimental results demonstrate that the model trained with this synthetic dataset can achieve good performance on several publicly available real-world HDR video testsets.

In summary, this paper has the following main contributions:

(1) we analyze the temporal clues in LDR videos and propose a novel video ITM pipeline with the Global Sampling and Local Propagation strategy which can exploit temporal clues effectively and recover over-exposed areas with high fidelity efficiently.

(2) We design a two-stage pipeline, it includes an Incremental Clue Aggregation Module to align and fuse frames with brightness changes under ITM problem. And it includes an efficient Feature and Clue Propagation Module which propagate features of the key frames to their neighbors and generate temporally consistent results.

(3) we propose a novel dataset synthesis method to obtain HDR video training dataset only using available HDR images and LDR video dataset.

(4) Experiments demonstrate that the proposed method outperforms the state-of-the-art methods both on quantitative and visual quality.

2. Related Work

Multi-exposure HDR reconstruction HDRI technology fuses the stack of multi-exposure images into the HDR radiance. Recently, many deep-learning-based methods have been proposed to generate ghost-free HDR images or videos from multi-exposure sequences. Chen et al. [4] propose a coarse-to-fine framework for HDR video reconstruction from alternating exposures based on optical flow and deformable alignment. In contrast, we focus on reconstructing

an HDR video from a single LDR video.

Single HDR image reconstruction There are many deep-learning-based image ITM methods. The direct learning method aims to generate the HDR radiance from an end-to-end model. Eilertsen et al. [7] focus on restoring the lost information in the saturated image areas by an end-to-end network. Liu et al. [22] utilize the LDR image formation pipeline to reverse it and reconstruct the HDR image step by step. Zheng et al. [38] propose an ultra-high-definition HDR reconstruction method via a collaborative learning manner. The stack-based methods change the exposure of the input image and estimate the multi-exposure stack. Endo et al. [8] estimate LDR images with different exposures by 3D convolution. Lee et al. [19] use the generative adversarial network (GAN) to recursively generate the exposure-changed image. Kim et al. [14] propose a differentiable HDR synthesis layer that forms the end-to-end stack-based ITM network.

HDR video reconstruction Rempel et al. [29] propose the video ITM method by contrast stretching and saturated brightness enhancement. However, the lost details of the over-exposed areas are not recovered. Recently, SDRTV-HDRTV conversion has attracted much attention. Kim et al. [15] joint handle the super-resolution and inverse tone mapping by a single model. Xu et al. [36] propose a frequency-aware modulation network that can reduce the structural distortions in the translated low-frequency regions. However, these methods usually deal with a single frame of video and assume there are no severely over-exposed regions. Xu et al. [37] use a 3D convolutional neural network to perform the video ITM, which simply processes a batch of adjacent frames, where the exposure settings are relatively close and little available information can be found to help the recovery. Different from the above methods, we generate the linear luminance HDR video based on temporal clues.

3. Method

3.1. Overview

Given an input LDR video $\{L_t, t = 1...T\}$, our goal is to reconstruct the corresponding HDR video $\{H_t, t = 1...T\}$. As with [7][32], we focus on the reconstruction of the over-exposed regions, which is the most difficult part during HDR conversion. Specifically, we fuse the linearized LDR frame $f^{-1}(L_t)$ and the output of the proposed method Y_t by a soft over-exposed mask α_t to get the reconstructed result $\hat{H}_t$:

$$\hat{H}_t = (1 - \alpha_t) f^{-1}(L_t) + \alpha_t Y_t, \tag{1}$$

where f^{-1} is the inverse camera curve which transforms the LDR frame into the linear domain, and Y_t is the output of the proposed method at time step t. The soft over-exposed mask α_t is calculated as in [7][32].

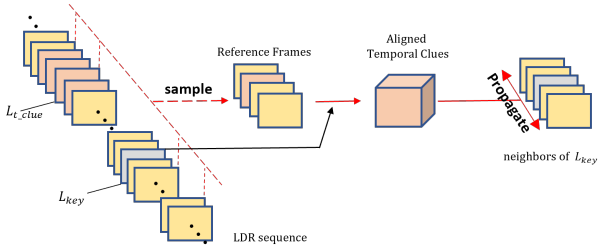

Figure 2. **Global Sampling and Local Propagation Strategy** for temporal clues. Orange frames such as L_{t_clue} contains useful temporal clues for recovering over-exposed regions in target key frame L_{key} and usually L_{t_clue} is far away from L_{key}. We sample key frames and reference frames from the whole LDR sequence, and aligns these references with each L_{key}. And then propagate aligned temporal clues to the neighbors of L_{key}.

Because of the video filming process, it's difficult to make full use of temporal information. Instead of reconstructing all of the frames directly, we use an effective and efficient recovery strategy called Global Sampling and Local Propagation strategy. Firstly, we only focus on the recovery of some keyframes with the help of global reference frames and then propagate the restored keyframes locally to their neighbor frames. To better find useful temporal clues from global-sampled reference frames and reconstruct keyframes, we design the Incremental Clue Aggregation Module with flow correction convolution which can align and fuse frames with brightness changes under the circumstances of ITM problem . After that, we utilize the proposed Feature and Clue Propagation Module which transfers the information from reconstructed keyframes to their neighbors and generates temporally consistent results. We will describe each component in detail in the following.

3.2. Strategy of Temporal Clues

During video filming, as the exposure value changes, objects that are over-exposed in current frame may become normally exposed in following frames. This phenomenon makes it possible to recover realistic textures for over-exposed regions by using valid information provided by other frames, which are the temporal clues. However, such a exposure changing process tends to be designed to be very slow and smooth in order to ensure filmed video not flicker, which makes it difficult to find and utilize temporal clues. Current ITM strategies can not fully utilize temporal clues. Image ITM methods like [7][14][22] don't utilize temporal clues. Because of temporal clues often in distant frames, the sliding window based strategy in [34] which only look at a short sequence of adjacent frames may not reach valid temporal clues. Although the RNN-based strategy like in [2] can cover long-term information in a sequence while it

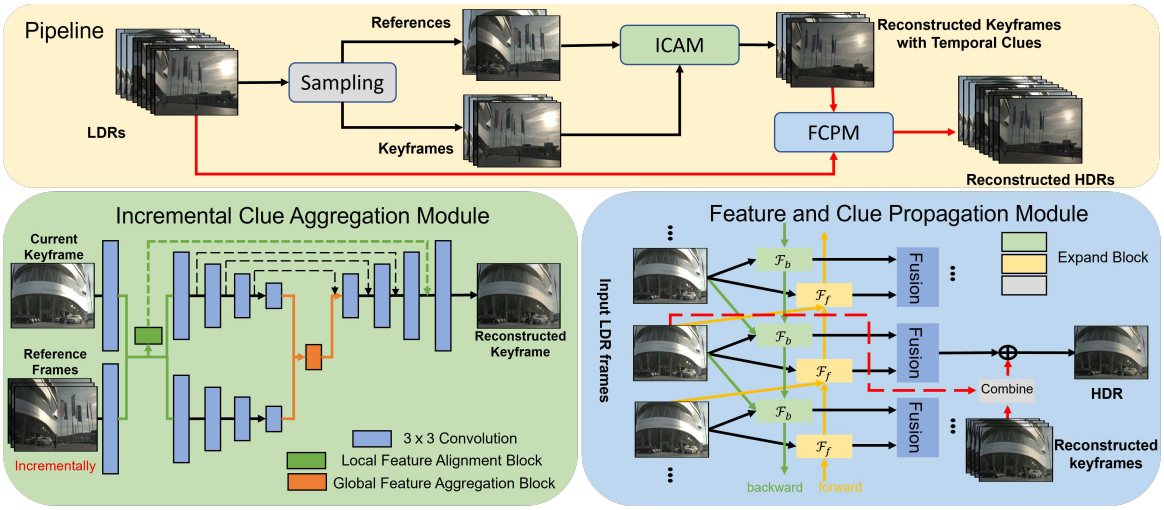

Figure 3. The proposed **pipeline** can be divided into two stages (above). In the first stage, we sample reference frames and key frames from input LDR sequence. And then extract and align temporal clues to reconstruct keyframes by ICAM (lower left). In the second stage, we utilize the restored keyframes with aligned temporal clues to propagate information to their neighbors to generate the full sequence by FCPM (lower right). The dashed lines of ICAM indicate skip connections and the reference frames are utilized incrementally to perform clue aggregation.

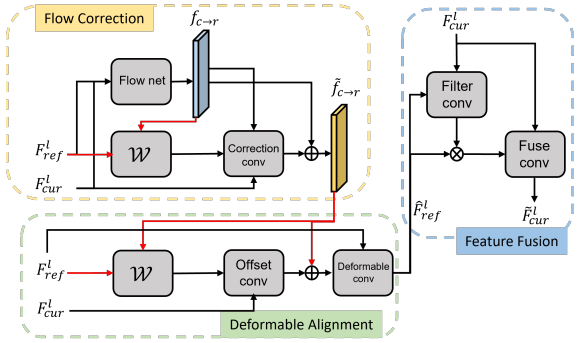

Figure 4. The details of the Local Feature Alignment Block.

may lead to a significant increase in computational complexity and when propagates information, useful temporal clues may fade gradually and errors also be accumulated.

To address these problems, we propose a Global Sampling and Local Propagation strategy as Fig. 2 shows. Specifically, we uniformly sample T_{key} keyframes from a input video, and split the entire input sequence into several equal-length groups of pictures (GOP). A key frame is the center frame of a GOP. We reconstruct these key frames with the help of temporal clues. To get valid temporal clues we sample T_{ref} reference frames with large stride globally from the input LDRs. And then we align these reference frames with each key frame to obtain aligned temporal clues. After that, we propagate aligned temporal clues of each key frame locally among its GOP to generate HDR results.Note that choosing the best matching references for temporal clues may produce better results, but it will lead to very cumbersome and inefficient frame-by-frame searching calculation, thus reducing the practical application value of the algorithm. Due to the smooth changes of exposure value, a preset stride to sample reference frames is sufficient for exploiting temporal clues and this makes a good balance between performance and complexity. There is some overlap among GOPs, so each input LDR frame will be taken care of by several reconstructed key frames, which not only ensure consistency locally among GOP but also conveys consistency globally across the whole input sequence.

3.3. Incremental Clue Aggregation Module

Note that the number of reference frames corresponding to a key frame is not fixed. For example, if there is no camera motion or light condition change in the video, there will be no temporal clue. To adapt to different situations, we propose the ICAM which can borrow useful temporal clues from global reference frames in an incremental way. As shown in Fig. 3, ICAM is designed based on 5-level U-Net[30]. However, a simple U-net cannot meet the requirements of extracting and aligning temporal clues. So we design the local and global feature alignment blocks for aligning reference frames with current key frame spatially.

3.3.1 Local Feature Alignment Block

Flow Correction Convolution To align temporal clues with over-exposed regions in current key frame, at first, we calculate the optical flow between them by the pre-trained

optical flow estimation network GMA [13]. But typical optical flow methods like GMA estimate optical flows based on the assumption that the observed brightness of any object point is constant over timeline. Unfortunately, for ITM task, there may be huge differences in the luminance among frames. What's worse, it is hard to find corresponding pixels for over-exposed regions due to the lack of textures in these regions, which will result in an inaccurate optical flow. To overcome this limitation, we design a flow-correction block which takes as input the optical flow estimated by pre-trained optical flow net $\mathcal{F}$ and the corresponding features extracted by encoders and predict a corrected optical flow. Specifically, at first we use the pre-trained GMA [13] as $\mathcal{F}$ to predict the initial optical flow from the current key frame L_{cur} to reference ones L_{ref}:

$$f_{c\to r} = \mathcal{F}(L_{cur}, L_{ref}). \tag{2}$$

Then we use the estimated optical flow $f_{c\to r}$ to warp the local feature F^l_{ref} of L_{ref} extracted by the first encoder block of ICAM to be aligned with F^l_{cur}. We design a correction convolutional block $\mathcal{C}$ which consists of several 3 $\times$ 3 convolutional layers to refine the initial optical flow:

$$\tilde{f}_{c\to r} = \mathcal{C}(F^l_{cur}, F^l_{ref}, \mathcal{W}(F^l_{ref}, f_{c\to r})) + f_{c\to r}, \tag{3}$$

where $\mathcal{W}$ denotes the warp operation. To train the correction convolutional block with supervision, we generate the corresponding LDR frames from the ground truth HDR frames by the global tone mapping operator as in [35], which can preserve details in highlight regions and ensure lighting consistency. Then we use pre-trained GMA [13] to estimate the ground truth optical flow f_{gt} for them and calculate the flow correction loss:

$$\mathcal{L}_{flow} = \left| \tilde{f} - f_{gt} \right|. \tag{4}$$

Flow Guided Deformable Convolution Due to possible large motions between the key frame and long-distance references, the corrected optical flow isn't enough to handle all of this complicated situation. So we further incorporate the deformable convolutional block $\mathcal{D}$ [5] to get more accurately aligned features. We also employ $\tilde{f}_{c\to r}$ to guide the deformable convolutional block as in [3]:

$$\hat{F}^l_{ref} = \mathcal{D}(F^l_{cur}, F^l_{ref}, \mathcal{W}(F^l_{ref}, \tilde{f}_{c\to r}), \tilde{f}_{c\to r}), \tag{5}$$

where $\hat{F}^l_{ref}$ is the final aligned feature.

Feature Fusion Convolution Before fusing $\hat{F}^l_{ref}$ with F^l_{cur}, we calculate a filter mask M^l_{ref} for $\hat{F}^l_{ref}$ by two 3 $\times$ 3 convolutional layers and adopt Sigmoid to determine if the feature is valuable or not for the current key frame. Finally, we fuse the above features by the fusion convolutional block $\mathcal{F}_u$ which consists of three dense blocks [11] to get the aligned local feature $\tilde{F}^l_{cur}$:

$$\tilde{F}^l_{cur} = \mathcal{F}_u(F^l_{cur}, \hat{F}^l_{ref} \cdot M^l_{ref}). \tag{6}$$

3.3.2 Global Feature Aggregation Block

As we know, global spatial information is also important for ITM task. Therefore, on the basis of aligned local spatial features, we introduce a global feature aggregation block to further utilize global spatial information in the smallest resolution of the Unet. In terms of implementation, we refer to the transformer block used in [23] [21] which can capture long-distant dependencies by the self-attention mechanism, which is helpful to model high-level semantic features and aggregate textures for over-exposed regions. We extract global features $\hat{F}^g_{ref}$ from the aligned $\hat{F}^l_{ref}$ by four 3 $\times$ 3 convolutional layers with stride 2. Then we concatenate these features with the global features F^g_{cur} of the current key frame in the channel dimension, and utilize four swin transformer blocks $\mathcal{S}$ to get aggregated global features $\tilde{F}^g_{cur}$ of the current key frame:

$$\tilde{F}^g_{cur} = \mathcal{S}(F^g_{cur}, \hat{F}^g_{ref}), \tag{7}$$

After alignment locally and globally in space, there are four decoder blocks with skip connections to obtain temporal clues and reconstruct the HDR key frame. The details of each proposed component can be found in the supplementary material. Overall, the proposed ICAM takes the current keyframe L_{cur} and reference frames $L_{ref_j, j=1...T_{clues}}$ as input and obtain temporal clues and predict HDR key frames X_{cur}. Note that the Local Feature Align block and the Global Feature Aggregation block are used repeatly if there is more than one reference frame. We train ICAM with losses between X_{cur} and H_{cur}, that's the $L1$ pixel loss $\mathcal{L}_{pix}$ and the perceptual loss $\mathcal{L}_{per}$ by the VGG-16 [33] pre-trained on ImageNet [6]. Therefore, the total training loss of ICAM is:

$$\mathcal{L}_{ICAM} = \mathcal{L}_{pix} + \lambda_{per}\mathcal{L}_{per} + \lambda_{flow}\mathcal{L}_{flow}, \tag{8}$$

where λ_{per} and λ_{flow} are set to 0.1 and 0.05 separately.

3.4. Feature and Clue Propagation Module

To convert the whole input LDR sequence to HDR sequence, the most straightforward idea is to regard all input frames as key frames and process them one by one using above proposed ICAM. However, This straightforward way has two drawbacks. First, this way only considers global temporal clues but neglects local adjacent frames, which may ignores valuable information and cause the temporal inconsistency problem. Second, ICAM involves complicated feature manipulations, processing all frames by it introduces much computational cost. Considering that the exposure usually changes smoothly, the reconstruction of neighbor frames in a GOP should be nearly consistent. Therefore, we propose the Feature and Clue propagation module which can broadcast recontructed textures and aggregated temporal clues in HDR keyframes to their neighbor frames. Our method can release the burden of repeat

computation of temporal clues and keep consistency along time axis. Specifically, for each frame L_t of input LDR video, we regard it in five nearest GOPs. Upon that, five according HDR keyframes $X_{ngb_j}, j = 0...5$ reconstructed by ICAM are warped to L_t using the optical flow estimated by the proposed flow correction block. And a combination masks M_{ngb_j} is predicted by a combination convolution block, which extracts the local, dilated, and global features and fuses them into the output features like ExpandNet[26]. Then we generate the merged result $\hat{X}_t$:

$$\hat{X}_t = \sum_{j=0}^{4} M_{ngb_j} \cdot \mathcal{W}(X_{ngb_j}, f_{t \to ngb_j}). \tag{9}$$

In this way, the restored textures and temporal clues in X_{ngb_j} can be propagated into preliminarily reconstructed result $\hat{X}_t$. And thus reducing computational cost compared with directly processing L_t by ICAM. Furthermore, to utilize the local information which may be ignored by ICAM, we perform the bidirectional features and clues propagation among the input frames like BasicVSR [2]. To further adapt ITM task, we use the corrected optical-flow and devise the basic block as the combination block:

$$h_t^b = \mathcal{F}_b(\hat{X}_t, \hat{X}_{t+1}, \mathcal{W}(h_{i+1}^b, f_{t \to t+1})), \tag{10}$$

$$h_t^f = \mathcal{F}_f(\hat{X}_t, \hat{X}_{t-1}, \mathcal{W}(h_{i-1}^f, f_{t \to t-1})), \tag{11}$$

where h_t^b and h_t^f denote the corresponding features of the backward propagation block $\mathcal{F}_b$ and forward propagation block $\mathcal{F}_f$ in time step t. Then the forward and backward features are concatenated and fused by two 3×3 convolutional layers to predict the residuals, which are added into the preliminary image $\hat{X}_t$ to get the final reconstructed results:

$$Y_t = \mathcal{F}_u(h_t^f, h_t^b) + \hat{X}_t. \tag{12}$$

The training loss of FCPM also contains pixel reconstruction loss and perceptual loss. Besides, the generative adversarial loss has been proven to improve perceptual quality by forcing the distribution of generated results closer to that of ground truth. Therefore, we also incorporate the 3D patchGAN [12] as the discriminator to distinguish the predicted sequences by FCPM from the corresponding ground truth HDR sequences. We adopt the least-square GAN [25] as the adversarial loss $\mathcal{L}_{GAN}$. Therefore, the total training loss of FCPM is:

$$\mathcal{L}_{FCPM} = \mathcal{L}_{pix} + \lambda_{per}\mathcal{L}_{per} + \lambda_{gan}\mathcal{L}_{GAN}, \tag{13}$$

where λ_{per} and λ_{gan} are set to 0.1 and 0.05 separately.

3.5. Dataset Synthesis

Since there are few publicly available HDR video datasets, we propose an HDR video data synthesis method that can utilize existing HDR image and LDR video datasets. Firstly, we convert the pixel values of existing LDR image datasets such as REDS [27] L_{reds} to the linearized domain by inverse camera response curve mapping. Then the linearized image is multiplied by a randomly sampled value T to simulate the exposure duration T. Finally, we clip the pixel to [0,1] and use the camera response curve to map them back to the pixel domain as the input of the network I_{reds} by the following equation:

$$O_{reds} = f(clip(f^{-1}(L_{reds}) \cdot T)), \tag{14}$$

where f denotes the camera response curve and we use the gamma function with 1/2.2 here. However compared to the real HDR, the fake HDR data generated using LDR images still has a large gap in dynamic range and information magnitude. Besides datasets from LDR videos, we use HDR image datasets like SICE dataset [1] to simulate HDR videos. As specified before, one of the main reasons that cause the exposure setting to change is camera motions. Therefore, we perform a random perspective transformation on the HDR images to simulate camera motions, which we use as the HDR video clip and follow Eq. 14 to generate the input LDR frames except for the inverse camera curve mapping.

We utilize the video characteristics of LDR videos and the HDR characteristics of HDR images to generate HDR video datasets. Experimental results show that networks trained by our dataset can achieve good performance on real-world HDR video testing datasets. Examples of the synthetic datasets can be found in the supplementary material.

4. Experiments

4.1. Implementation details

Dataset The HDR image dataset used to generate synthetic data is SICE [1], which contains 589 HDR images. And we use the "sharp" training dataset of REDS [27] as the LDR video dataset. As for the testing dataset, at first, we evaluate the performance of the synthetic data generated from the validation dataset of REDS [27]. Then we use three public real-world HDR video datasets: HDM-HDRv [9], LiU-HDRv [17], and MPI-HDRv [10]. Because there is no corresponding LDR version for these datasets, we generate the LDR videos from them by simulating the camera imaging pipeline. The details of the datasets can be found in the supplementary material.

Experiment setup The implementation environment is PyTorch 1.9 version and the Adam optimizer is applied to train

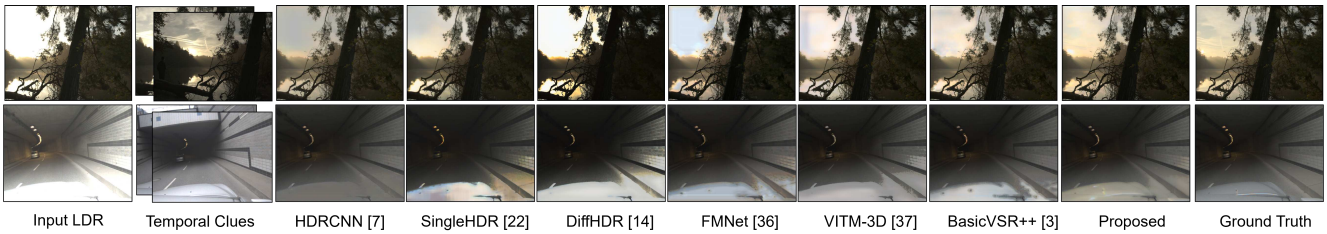

Figure 5. Visual comparisons on the frame of "fishing_longshot" sequence from HDM-HDRv [9] (above) and "sb-tunnel-exr" sequence from MPI-HDRv [10] (below). (**Zoom in for a better view**).

Table 1. Quantitative comparison on HDR videos with existing methods. The scores here are HDR-VDP-3/HDR-VQM, where a higher score of HDR-VDP-3 and a lower score of HDR-VQM mean better. Red text indicates the best and blue text indicates the second best result, respectively.

	REDS-val [27]	HDM-HDR [9]	LiU-HDR [17]	MPI-HDRv [10]
HDRCNN [7]	6.774/0.452	5.822/0.467	7.884/0.545	6.733/0.053
Diff HDRI [14]	6.820/0.502	5.703/0.514	7.751/0.594	6.501/0.061
Single HDRI [22]	7.059/0.492	6.346/0.481	8.143/0.569	7.225/0.058
FMNet [36]	6.895/0.483	5.847/0.496	7.974/0.572	6.847/0.057
Deep VITM [37]	7.106/0.458	5.992/0.445	8.036/0.553	7.104/0.049
Bascivsr++ [3]	7.254/0.443	6.131/0.427	8.265/0.547	7.119/0.052
Proposed	**7.891/0.398**	**6.754/0.356**	**8.591/0.522**	**7.970/0.037**

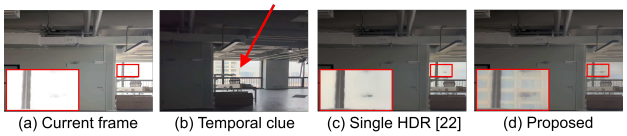

Figure 6. Visual comparison on the frame of a real-world LDR video shot by iPhone 13.

the model with a learning rate of 0.0002. We first resize the training pairs in the training dataset to 512 × 512 and augment them by randomly cropping to 384 × 384. The training images are randomly flipped and rotated. The frames of testing datasets are resized to 512 × 512 for evaluation. For each video sequence, we sample keyframes uniformly with the stride of six and find six temporal clues for each keyframe in the forward and backward direction respectively with the stride of 15.

Evaluation metrics We evaluated the estimated HDR in terms of HDR-VDP-3 [24] which is a commonly used metric to measure the quality of single HDR image reconstruction, and HDR-VQM [28] which is designed for evaluating the quality of HDR videos.

4.2. Comparisons on the predicted HDR videos

The proposed method is compared with three image ITM methods (HDRCNN [7], Differentiable HDR [19], and Single HDR [22]), one SDRTV-HDRTV conversion method FMNet [36], one video ITM method Deep VITM [37], and one video restoration method BasicVSR++ [3]. For fair comparisons, we re-train these models with the same training dataset. (For the first four methods, only a single frame is taken as input, and for the last two , we feed them with frames as the proposed methods and only reconstruct the over-exposed regions too.)

Visual comparisons. Fig. 5 show the results of these ITM methods on two LDR images with severely over-exposed regions. The single-frame-based methods (HDRCNN [7], Differentiable HDR [19], Single HDR [22], and FMNet [36]) can not restore textures from the rare available temporal information. Deep VITM [37] performs 3D convolutions directly and fails to model the correlation between frames. BasicVSR++ [3], due to the lack of explicit exploit, temporal clues fade gradually when propagated. On the contrary, based on full use of temporal clues, the proposed method can generate impressive results with high fidelity. Fig. 6 shows an example of a real-world LDR video shot by iPhone 13, where compared to the SOTA image ITM method [22], the proposed method can reconstruct realistic and natural textures with the help of temporal clues in reference frame. All of the HDR frames are tone mapped by the [16] for display on LDR devices.

Quantitative comparisons. Table 1 shows the average scores of HDR-VDP-3 and HDR-VQM on the REDS-val [27], HDM-HDRv [9], LiU-HDRv [17], and MPI-HDRv [10] datasets. The proposed method performs favorably against the state-of-the-art methods on all four datasets.

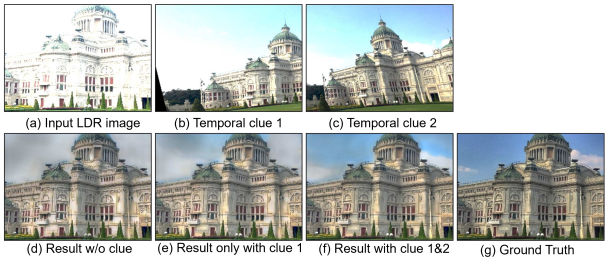

Figure 7. Visual comparison on reconstructed result with different temporal clues.

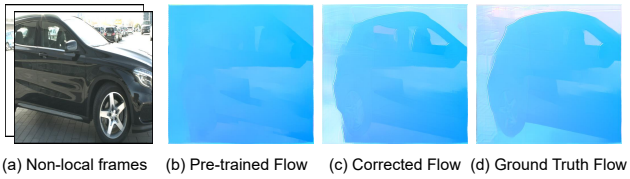

Figure 8. Visual comparison between the pre-trained and corrected optical flow.

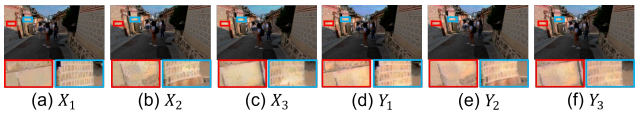

Figure 9. Visual comparison between the reconstructed adjacent frames. (a)-(c) all generated only by ICAM. ((d)-(f)) generated by the proposed pipeline with FCPM and reconstructed key frames.

Table 2. Ablation study of aggregation components. HDR-VQM scores of (a)pre-trained flow, (b)corrected flow, (c)flow-guided Deformable Convolution and (d)global aggregation.

w/o align	(a)	(b)	(c)	(d)
0.418	0.396	0.380	0.367	**0.356**

Table 3. Quantitative ablation study of the propagation methods.

w/o FCPM	Warp	Forward	Backward	FCPM
0.383	0.376	0.369	0.366	**0.356**

4.3. Ablation studies

Incremental Clue Aggregation. We validate the effectiveness of the incremental aggregation. As shown in Fig. 7, with more temporal clues, ICAM can learn the desired textures from the clues in an incremental way and generates increasingly realistic and natural details.

Components of the ICAM. We evaluate the effectiveness of the modules in Incremental Clue Aggregation Module. The results of HDR-VQM scores on HDM-HDR dataset are shown in Table 2. The proposed flow correction convolution and according flow-guided deformable convolution both contribute to generate more accurate results. Meanwhile, the global feature aggregation can also imporve the performance. Fig. 8 shows optical flows of two frames with large motion and different exposure.

Flow Correction. The pre-trained flow is estimated by pre-trained GMA [13], corrected flow is estimated by the flow correction block, and the ground truth flow is generate by pre-trained GMA [13] for the tone mapped images with consistent luminance and no over-exposed regions.

Components of the FCPM. We conduct experiment to verify the effect of different feature propagation methods on the results. Specifically, we first obtain preliminary results by warping and combining neighboring reconstructed HDR keyframes. Then we perform forward, backward, and bidirectional feature propagation respectively to fine-tune the preliminary reconstructed frames using inter-frame information, and the HDR-VQM scores on the HDM-HDR dataset are shown in Table 3.

Temporal consistency. Fig. 9 shows the comparison between the results of all generated by ICAM and generated by the proposed pipeline. only using ICAM generates temporal inconsistent textures while the proposed pipeline can avoid this problem and reconstruct more realistic result.

Running Time. We also compare the average running time processing a 512 $\times$ 512 frame on Tesla V100 GPU of only using ICAM for entire sthe equence (1236ms) and the proposed pipeline (521ms).

5. Conclusions

In this paper, we analyze the characteristics of LDR videos, and propose a novel global sampling and local propagation strategy to fully exploit Temporal Clues. In order to better server the proposed strategy, we design the Incremental Clue Aggregation Module and Feature and Clue Propagation Module. These modules can make full use of temporal clues to recovery details with high fidelity meanwhile ensure temporal consistency. In addition, we devise an HDR video dataset synthesis method to train our method. Experimental results show that the proposed video ITM method outperforms the state-of-the-art methods in both quantitative and qualitative evaluations.

Acknowledgements. This work is financially supported for Outstanding Talents Training Fund in Shenzhen, Shenzhen Science and Technology Program-Shenzhen Cultivation of Excellent Scientific and Technological Innovation Talents project(Grant No. RCJC20200714114435057) , Shenzhen Science and Technology Program-Shenzhen Hong Kong joint funding project (Grant No. SGDX20211123144400001), National Natural Science Foundation of China U21B2012 and 62272142, R24115SG MIGU-PKU META VISION TECHNOLOGY INNOVATION LAB.

References

[1] Jianrui Cai, Shuhang Gu, and Lei Zhang. Learning a deep single image contrast enhancer from multi-exposure images. *IEEE Transactions on Image Processing*, 27(4):2049–2062, 2018. 6

[2] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Basicvsr: The search for essential components in video super-resolution and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4947–4956, 2021. 3, 6

[3] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5972–5981, 2022. 5, 7

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

[5] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 2, 5

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[7] Gabriel Eilertsen, Joel Kronander, Gyorgy Denes, Rafał K Mantiuk, and Jonas Unger. Hdr image reconstruction from a single exposure using deep cnns. *ACM transactions on graphics (TOG)*, 36(6):1–15, 2017. 1, 3, 7

[8] Yuki Endo, Yoshihiro Kanamori, and Jun Mitani. Deep reverse tone mapping. *ACM Trans. Graph.*, 36(6):177–1, 2017. 1, 3

[9] Jan Froehlich, Stefan Grandinetti, Bernd Eberhardt, Simon Walter, Andreas Schilling, and Harald Brendel. Creating cinematic wide gamut hdr-video for the evaluation of tone mapping operators and hdr-displays. In *Digital photography X*, pages 279–288. SPIE, 2014. 1, 6, 7

[10] Vlastimil Havran, Miloslaw Smyk, Grzegorz Krawczyk, Karol Myszkowski, and Hans-Peter Seidel. Interactive system for dynamic scene lighting using captured video environment maps. In *Rendering Techniques*, pages 31–42, 2005. 6, 7

[11] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 5

[12] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 6

[13] Shihao Jiang, Dylan Campbell, Yao Lu, Hongdong Li, and Richard Hartley. Learning to estimate hidden motions with global motion aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9772–9781, 2021. 5, 8

[14] Jung Hee Kim, Siyeong Lee, and Suk-Ju Kang. End-to-end differentiable learning to hdr image synthesis for multi-exposure images. *arXiv preprint arXiv:2006.15833*, 2020. 1, 3, 7

[15] Soo Ye Kim, Jihyong Oh, and Munchurl Kim. Deep sr-itm: Joint learning of super-resolution and inverse tone-mapping for 4k uhd hdr applications. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3116–3125, 2019. 2, 3

[16] Chris Kiser, Erik Reinhard, Mike Tocci, and Nora Tocci. Real time automated tone mapping system for hdr video. In *IEEE International Conference on Image Processing*, pages 2749–2752. IEEE Orlando, FL, 2012. 7

[17] Joel Kronander, Stefan Gustavson, Gerhard Bonnet, Anders Ynnerman, and Jonas Unger. A unified framework for multi-sensor hdr video reconstruction. *Signal Processing: Image Communication*, 29(2):203–215, 2014. 6, 7

[18] Siyeong Lee, Gwon Hwan An, and Suk-Ju Kang. Deep chain hdri: Reconstructing a high dynamic range image from a single low dynamic range image. *IEEE Access*, 6:49913–49924, 2018. 1

[19] Siyeong Lee, Gwon Hwan An, and Suk-Ju Kang. Deep recursive hdri: Inverse tone mapping using generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 596–611, 2018. 1, 3, 7

[20] Siyeong Lee, So Yeon Jo, Gwon Hwan An, and Suk-Ju Kang. Learning to generate multi-exposure stacks with cycle consistency for high dynamic range imaging. *IEEE Transactions on Multimedia*, 2020. 1

[21] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1833–1844, 2021. 5

[22] Yu-Lun Liu, Wei-Sheng Lai, Yu-Sheng Chen, Yi-Lung Kao, Ming-Hsuan Yang, Yung-Yu Chuang, and Jia-Bin Huang. Single-image hdr reconstruction by learning to reverse the camera pipeline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1651–1660, 2020. 1, 2, 3, 7

[23] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 2, 5

[24] Rafał Mantiuk, Kil Joong Kim, Allan G Rempel, and Wolfgang Heidrich. Hdr-vdp-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Transactions on graphics (TOG)*, 30(4):1–14, 2011. 7

[25] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE Interna-*

tional Conference on Computer Vision, pages 2794–2802, 2017. 6

[26] Demetris Marnerides, Thomas Bashford-Rogers, Jonathan Hatchett, and Kurt Debattista. Expandnet: A deep convolutional neural network for high dynamic range expansion from low dynamic range content. In *Computer Graphics Forum*, pages 37–49. Wiley Online Library, 2018. 1, 6

[27] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In *CVPR Workshops*, 2019. 6, 7

[28] Manish Narwaria, Matthieu Perreira Da Silva, and Patrick Le Callet. Hdr-vqm: An objective quality measure for high dynamic range video. *Signal Processing: Image Communication*, 35:46–60, 2015. 7

[29] Allan G Rempel, Matthew Trentacoste, Helge Seetzen, H David Young, Wolfgang Heidrich, Lorne Whitehead, and Greg Ward. Ldr2hdr: on-the-fly reverse tone mapping of legacy video and photographs. *ACM transactions on graphics (TOG)*, 26(3):39–es, 2007. 3

[30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 4

[31] Mehdi SM Sajjadi, Raviteja Vemulapalli, and Matthew Brown. Frame-recurrent video super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6626–6634, 2018. 2

[32] Marcel Santana Santos, Tsang Ing Ren, and Nima Khademi Kalantari. Single image hdr reconstruction using a cnn with masked features and perceptual loss. *ACM Transactions on Graphics (TOG)*, 39(4):80–1, 2020. 1, 3

[33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[34] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2, 3

[35] Shangzhe Wu, Jiarui Xu, Yu-Wing Tai, and Chi-Keung Tang. Deep high dynamic range imaging with large foreground motions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 117–132, 2018. 5

[36] Gang Xu, Qibin Hou, Le Zhang, and Ming-Ming Cheng. Fmnet: Frequency-aware modulation network for sdr-to-hdr translation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6425–6435, 2022. 2, 3, 7

[37] Yucheng Xu, Li Song, Rong Xie, and Wenjun Zhang. Deep video inverse tone mapping. In *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, pages 142–147. IEEE, 2019. 2, 3, 7

[38] Zhuoran Zheng, Wenqi Ren, Xiaochun Cao, Tao Wang, and Xiuyi Jia. Ultra-high-definition image hdr reconstruction via collaborative bilateral learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4449–4458, 2021. 1, 3