Learning a Practical SDR-to-HDRTV Up-conversion using New Dataset and Degradation Models

Cheng Guo\textsuperscript{1,2}, Leidong Fan\textsuperscript{3,2}, Ziyu Xue\textsuperscript{4,1} and Xiuhua Jiang\textsuperscript{2,1}

\textsuperscript{1}State Key Laboratory of Media Convergence and Communication, Communication University of China
\textsuperscript{2}Peng Cheng Laboratory \textsuperscript{3}Peking University
\textsuperscript{4}Academy of Broadcasting Science, National Radio and Television Administration

\{guocheng,jiangxiuhua\}@cuc.edu.cn, fanleidong@stu.pku.edu.cn, xueziyu@abs.ac.cn

Abstract

In media industry, the demand of SDR-to-HDRTV up-conversion arises when users possess HDR-WCG (high dynamic range-wide color gamut) TVs while most off-the-shelf footage is still in SDR (standard dynamic range). The research community has started tackling this low-level vision task by learning-based approaches. When applied to real SDR, yet, current methods tend to produce dim and desaturated result, making nearly no improvement on viewing experience. Different from other network-oriented methods, we attribute such deficiency to training set (HDR-SDR pair). Consequently, we propose new HDRTV dataset (dubbed HDRTV4K) and new HDR-to-SDR degradation models. Then, it’s used to train a luminance-segmented network (LSN) consisting of a global mapping trunk, and two Transformer branches on bright and dark luminance range. We also update assessment criteria by tailored metrics and subjective experiment. Finally, ablation studies are conducted to prove the effectiveness. Our work is available at: https://github.com/AndreGuo/HDRTVDM.

1. Introduction

The dynamic range of image is defined as the maximum recorded luminance to the minimum. Larger luminance container endows high dynamic range (HDR) a better expressiveness of scene. In media and film industry, the superiority of HDR is further boosted by advanced electro-optical transfer function (EOTF) e.g. PQ/HLG [2], and wide color-gamut (WCG) RGB primaries e.g. BT.2020 [3].

While WCG-HDR displays are becoming more readily available in consumer market, most commercial footage is still in standard dynamic range (SDR) since WCG-HDR version is yet scarce due to exorbitant production workflow. Hence, there raise the demand of converting vast existing SDR content for HDRTV service. Such SDR may carry irreproducible scenes, but more likely, imperfections brought by old imaging system and transmission. This indicates that SDR-to-HDRTV up-conversion is an ill-posed low-level vi-
sion task, and research community has therefore begun involving learning-based methods ([4–9] etc.).

Yet, versatile networks they use (§2.1), we find current methods’ result dim and desaturated when feeding real SDR images (Fig.1), conflicting with the perceptual motive of SDR-to-HDRTV up-conversion. As reported by CVPR22-1st Workshop on Vision Dataset Understanding [10], most methods are network-oriented and underestimate the impact of training set. For restoration-like low-level vision, there are 2 ingredients of a training set: the quality of label GT itself, and the GT-to-LQ degradation model (DM) i.e. what the network learns to restore. Such neglect is getting remedied in other low-level vision tasks [11–16], but still pervasive in learning-based SDR-to-HDRTV up-conversion.

Not serendipitously, we find dataset the reason why current methods underperform. We exploit several HDRTV-tailored metrics (Tab.4) to assess current training set:(1) by measuring label HDR’s extent of HDR/WCG etc. (Tab.5), we notice that its quality and diversity are inadequate to incentivize the network to produce appealing result, (2) via the statistics of degraded SDR, we find current HDR-to-SDR DMs’ tendency to exaggeratedly alter the saturation and brightness (see Tab.6) thus network will learn a SDR-to-HDR deterioration. Hence, we propose HDRTV4K dataset (§3.2) consisting of high-quality and diversified (Fig.4) HDRTV frames as label. Then exploit 3 new HDRTV-to-SDR DMs (§3.3) avoiding above insufficiency, meanwhile possessing appropriate degradation capability (Tab.6) so the network can learn reasonable restoration ability.

Afterwards, we formulate the task as the combination of global mapping on the full luminance range and recovery of low/high luminance range. Correspondingly, we propose Luminance Segmented Network (LSN, §3.1) where a global trunk and two Transformer-style UNet [17] branches are assigned to respectively execute divergent operations required in different segmented luminance ranges (areas).

Lastly, as found by [18,19], conventional distance-based metrics well-performed in solely-reconstruction task (e.g. denoising) fail for perceptual-motivated HDR reconstruction, we therefore update the assessment criteria with fine-grained metrics (§4.2) and subjective experiment (§4.3) etc.

Our contributions are three-fold: (1) Emphasizing & verifying the impact of dataset on SDR-to-HDRTV task, which has long been understated. (2) Exploiting novel HDRTV dataset and HDR-to-SDR degradation models for network to learn. (3) Introducing new problem formulation, and accordingly proposing novel luminance segmented network.

2. Related Works

2.1. Nomenclature and our scope

Plenty of HDR-related learning-based methods [20] have been proposed, yet, they markedly differ in intended application. As in Tab.1, ‘linear HDR’ means scene-referred HDR images dedicating to record the linear radiance for graphics application e.g. image-based lighting [21,22], and ‘HDRTV’ stands for our display-referred WCG-HDR [2] format. Clearly, (1) synthesize HDR view on SDR display, (2) emphasize dealing with misalignment between multiple SDRs and is designed for new imaging pipeline, while (3) and (4) are all oriented for existing SDR, but for different application. Since many works are proposed before community’s clarity on above nomenclature, we classify a method as SDR-to-HDRTV up-conversion (4), our task) if its target HDR is claimed in PQ/HLG EOTF and BT.2020 WCG.

Table 1. Various learning-based HDR-related tasks. Green/cyan/magenta terms are respectively from [18/our/7] nomenclature.

<table>
<thead>
<tr>
<th>Task name</th>
<th>From</th>
<th>To single</th>
<th>Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDR-style enhancement</td>
<td>single SDR</td>
<td>enhanced SDR</td>
<td>e.g. [23–26]</td>
</tr>
<tr>
<td>multi-exposure HDR imaging</td>
<td>multiple SDR</td>
<td>linear HDR</td>
<td>e.g. [27–32]</td>
</tr>
<tr>
<td>3</td>
<td>S-HDR*</td>
<td>single SDR</td>
<td>HDRTV frame</td>
</tr>
<tr>
<td>(ours)</td>
<td>TSM* or up-conversion</td>
<td>[38–51]</td>
<td></td>
</tr>
</tbody>
</table>

*: S-HDR: Single-Image HDR reconstruction, TSM: Tone Mapping

In our scope (4), methods designed their networks with distinct philosophy: Except e.g. semantic-require high-light recovery, main process of our task belongs to global mapping resembling image retouching/enhancement. Following this, [7] formulates the task as 3 steps and use 1 × 1 convolution for global part, [43] learn the mapping between small sorted image patches, while [51] conducts explicitly-defined per-pixel mapping using the learned latent vector. Feature modulation is also popular to involve global prior information. Compared with [7], they change: prior’s type [47,49], modulation tensor’s shape [48–50], modulation mechanism [9] and modulation’s position [50].

To coordinate consecutive video frames, [42] applies 3D-convolution with extra temporal-D, [8] take 3 frames as input and deploy multi-frame interaction module at UNet bottleneck. Also, for alignment, [44,49] use deformable convolution whose offset map is predicted by different modules.

Due to the resolution discrepancy between SDRTV and HDRTV, some methods [4–6,46,48,52] jointly conduct super-resolution and up-conversion. Also, some methods are assisted by non-learning pre-processing [53] or bypass [43,45], while [41] is trained for badly-exposed SDR.

2.2. HDRTV dataset

Diversified networks they use, there’re currently only 3 open HDRTV training set (Tab.2). All datasets (including ours) contain HDRTV frames in (D65 white) BT.2020 [3] RGB primaries (gamut), PQ [54] EOTF and 1000nit peak luminance, and SDR counterpart in BT.709 [55] gamut, gamma EOTF and 100nit peak luminance.
Table 2. Status of different HDRTV dataset: training set part. Our HDR frames sized both HD (1920×1080) and UHD (3840×2160) are manually chosen from >220 different videos clips, and encapsulated in lossless (LZW or deflate) TIFF. Others are respectively extracted from only 7 [4]/18 [7] TV demos and 1 graded movie [6], their quality and diversity are quantified later in Tab.5 & Fig.4.

<table>
<thead>
<tr>
<th>Dataset (Usage)</th>
<th>#pair</th>
<th>Resolution</th>
<th>HDR format</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDRTV1K / [4,5,7–9,40,44,46–50,52]</td>
<td>1250</td>
<td>UHD</td>
<td>16bit yuv 420</td>
</tr>
<tr>
<td>HDRTV4K (ours new)</td>
<td>3878</td>
<td>HD/1920</td>
<td>16bit yuv 420</td>
</tr>
</tbody>
</table>

Table 3. Current HDRTV-to-SDR degradation models (DMs). ‘Dataset’ means SDR there is degraded from HDR using that DM. Yet, few discussion is made in SDR-to-HDRTV (Tab.3):

<table>
<thead>
<tr>
<th>Usage</th>
<th>Youtube</th>
<th>Reinhard</th>
<th>other DMs</th>
</tr>
</thead>
<tbody>
<tr>
<td>KAIST &amp; HDRTV</td>
<td>[16,28,38,41–43,51]/[53,56]</td>
<td>Zeng20</td>
<td>2446a [58]</td>
</tr>
</tbody>
</table>

3.2. HDRTV-to-SDR degradation model

Most methods follow the common practice to degrade label HDR(GT) to input SDR(LQ). Degradation model (DM) matters since it determined what network can learn, and is relatively paid more attention even in SI-HDR [16, 33, 35]. Yet, few discussion is made in SDR-to-HDRTV (Tab.3):

<table>
<thead>
<tr>
<th>DM</th>
<th>Youtube</th>
<th>Reinhard</th>
<th>other DMs</th>
</tr>
</thead>
<tbody>
<tr>
<td>lossless (LZW or deflate) TIFF</td>
<td>16bit yuv 420</td>
<td>16bit yuv 420</td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Current HDRTV-to-SDR degradation models (DMs). ‘Dataset’ means SDR there is degraded from HDR using that DM.

YouTube stands for the default conversion YouTube applied to BT.2020/PQ1000 HDR content to produce its SDR-applicable version, Reinhard/2446a means tone-mapping HDR to SDR using Reinhard TMO [57]/BT.2446 [58]Method A. [51]/[53]/[56] respectively degrade HDR to SDR by grading/Habel TMO/another learned network.

In other restoration tasks [11–16], DMs are designed to have proper extent and diversity of degradation so network can learn appropriate restore capability and good generalization. Accordingly, we argue that current DMs are not favorable for training. Specifically, the motive of YouTube is to synthesize HDR view for user possessing only SDR display, it tends to enhance/increase the brightness and saturation so network will, vise-versa, learn to deteriorate/decline them. Also, tone-mapping e.g. Reinhard and 2446a dedicate to preserve as much information from HDR, they have monotonically increasing mapping curves (Fig.5) without highlight clipping, therefore the trained network is not likely to recover much information in over-exposed areas. Above observations are later proven in Tab 6, Fig.1&6 etc.

3. Proposed Method

The overview of our method is illustrated in Fig.2.

3.1. Network structure

Problem formulation helps the researcher to clarify what degradation the network should restore. Given x the SDR, y the HDR. Previous methods [7, 47, 49] hypothesis that x and y are from single RAW file of camera.

This applies to SDR-HDR synchronous production where x is also of high quality, thus up-conversion (y = f(x)) is only supposed to follow specific style and recover less missing information. However, since our method is for existing SDR, we assume that SDR (x) and imaginary HDR counterpart (y) were simultaneously shot by different camera and later pipeline, similar to [18]. The imperfection of SDR pipeline makes existing SDR susceptible to limited latitude, narrow gamut and other degradations (x = d(y)).

Some worked [7, 35] suppose d(·) are introduced orderly, and assign hierarchical sub-networks to for f(·). Since cascaded networks are bulky for real application, we instead formulate that specific degradation is more visible in different luminance range. Concretely, over-exposure occurs in high-luminance range in SDR (x_h), low-luminance range (x_l) is more susceptible to noise etc., while the quality of mid-luminance range (x_m) is relatively higher. Segmented luminance ranges are treated separately by some traditional up-conversion operators ([59] etc.), and here we introduce this idea to our deep neural network (DNN)—LSN:

**Luminance Segmented Network (LSN)** consist of an trunk on full-luminance range (x), and 2 branches respectively on x_l and x_h which are segmented by:

\[
x_l = \max(0, \frac{t - x}{t}), \quad x_h = \max(0, \frac{x - 1}{t} + 1)
\]

where x ∈ [0, 1]. That is, luminance range lower/higher than threshold t (empirically set to 0.05) is linearly mapped to more significant value [0, 1], as in top-right Fig.2.

After segmentation, x, x_l and x_h require distinct DNN operation. First, as found by [7] etc. (§2.1), the majority of f(·) belongs to global (pixel-independent) operation similar to image enhancement/retouching. Therefore, we assign 5 cascaded receptive-field-free 1 × 1 convolution layers on full luminance range (x) as the **Global Mapping Trunk**. Similar to [7, 9, 47–50], we append a modulation branch (Fig.3 left) to involve x’s global prior into DNN’s decision.

On the other hand, f(·) still involves non-global restoration where local semantic should be aggregated. Specifically, DNN’s mission in x_l is similar to denoising and low-light enhancement, while that in x_h resembles image inpainting (claimed by [60,61]) where lost content is hallucinated. These low-level vision tasks require larger receptive-filed, we hence choose Transformer blocks [17] specializing long-distance dependency, and arranged them as encoder-decoder with skip-connections (UNet). **Transformer-style UNet branches** on x_l and x_h share the same structure but different parameters. The detail of LSN is depicted in Fig.3.

3.2. HDRTV4K dataset

After designing LSN, we need better training set to unleash its capability. We start with the quality of label HDR:
**Motivation.** In §1, we argue the quality of label HDR in current datasets. Here, it is assessed from 3 aspects by 10 metrics in Tab.4, based on following considerations:

First, greater **extent of HDR/WCG** stands more probability for network to learn pixel in advanced WCG-HDR volume beyond SDR's capability. Meanwhile, higher **intra-frame diversity** means better generalization the network can learn within small batch-size, i.e. bigger SI/CF/stdL indicate more diversified high-frequency patterns/richer color volume/greater luminance contrast for network to learn.

Also, style distillation [51] has become a major contributor to method's performance in similar task e.g. image enhancement/retouching, we therefore append 3 metrics quantifying the **overall-style** of label HDR. Note that network's learned style will be, substantially, affected more by degradation model (DM) which will be discussed later.

**Table 4.** The quality and diversity of label HDR is measured from 3 aspects by 10 metrics above (both in positive correlation), results are in Tab.5. See supplementary material for full illustration.

**Our work.** Statistics in Tab.5 confirms the deficiency of current datasets, i.e. lower **intra-frame diversity**, and most importantly, less **extent of HDR/WCG** preventing their network from producing true HDR-WCG volume. To this end, we propose a new dataset HDRTV4K consisting of 3878 BT.2020/PQ1000 (Tab.2) HDR frames with higher quality.

Specifically, these frames are manually extracted and aligned from various open content (e.g. [68–72] etc.) with greater **extent of HDR/WCG**, higher **intra-frame diversity**...
and reasonable style (Tab. 5). Note that we re-grade some original footage using DaVinci Resolve to add some perturbation on the diversity. The thumbnails of our dataset and its frame distribution comparison are visualized in Fig.4.

3.3. New degradation models

After obtaining label HDR with higher quality, we focus on degradation model (DM) which determines the restoration network will learn. As claimed in §2.3, current DMs fail for undue saturation & brightness change and meager over-exposure degradation, which are verified by ALL, ASL, FOEP in Table 6. That is, YouTube/2446a tend to increase/decline both ASL & ALL so network will learn to simultaneously de/over-saturate and unduly-darken/brighten. Also, 2446a and Rienhard provide small FOEP which is adverse to network’s over-exposure hallucination ability.

This motivate us to utilize/exploit new DMs with proper extent of degradation and no deficiencies on style:
(1) **OC102**: OpenColorIO v2 [74], commercial method from BT.2020/PQ1000 HDR container to BT.709/gamma SDR, implemented by 3D look-up table (LUT) here.
(2) **2446c+GM**: Modified Method C tone-mapping from Rec.2446 [58], followed by Gamut Mapping. Here, each HDR encoded value $E_i^\prime = [R_i^\prime, G_i^\prime, B_i^\prime]^T$ (we use superscript $\prime$ for non-linearity) is linearize to $E$ by PQ EOTF, and then transferred to $Y_{x,y}$. Then $Y$ is globally mapped to $Y_{SDR}$, while $x,y$ are kept and then recomposed with $Y_{SDR}$ to form $E_{SDR}^\prime$ and later $E_{SDR}$. Our modification lies in:

$$Y_{SDR} = \text{clamp}(TM_{2446c}(Y), 0, 100 \text{nit})$$

where $TM_{2446c}(\cdot) \in [0, 180 \text{nit}]$ is the original global mapping curve in [58] (Fig.5 dashed line). That is, we clip the original output range to 100nit to produce more FOEP, rather than linear-scale which share same shortcoming with DM 2446a. Gamut mapping will be introduced in Eq.4&5.
Figure 5. One of the key ideas of the proposed degradation models (DMs, §3.3): HDR’s luminance (left) and color (right) volume should be properly clipped, so that network will learn corresponding restoration. Another idea is keeping a sensible brightness & color/saturation style during degradation, please refer to Tab.6.

(3) **HC+GM:** Hard-Clipping and Gamut Mapping, a container conversion with SDR part unchanged and all luminance/color in HDR/WCG volume hard-clipped. For luminance part, we clip all >100nit pixels to 100nit:

$$E_{SDR} = 100 \times \text{clamp}(E, 0, 0.01), \ E \in [0, 1]$$  \hspace{1cm} (3)

where 0.01 correspond to 100nit in normalized linear HDR E (PQ container), and ×100 is to adapt the container discrepancy (nominal peak 1 means 10000nit → 100nit).

So far, $E_{SDR}$ is still in BT.2020 gamut. Therefore, we append Gamut Mapping (+GM): First, Color Space Transform (CST) from BT.2020 RGB primaries to BT.709:

$$E_{SDR_{OGG}} = ME_{SDR}, \ M = \begin{bmatrix} 1.6605 & 0.5876 & -0.0728 \\ -0.1246 & 1.1329 & 0.0083 \\ -0.0182 & -0.1006 & 1.1187 \end{bmatrix}$$  \hspace{1cm} (4)

where $\exists E_{SDR_{OGG}} \notin [0, 1]$ i.e. out-of-gamut (OOG) pixels will fall outside valid range after CST to a narrower volume. Dealing OOG is the key of gamut mapping, instead of soft-mapping [63, 75] (used by [76]’s DM) preserving as much WCG information to narrow gamut, we use hard-clipping which clips all OOG pixels to BT.709 boundary (Fig.5):

$$E_{SDR_{OOG}} = \text{clamp}(E_{SDR_{OGG}}, 0, 1)$$  \hspace{1cm} (5)

Then, $E_{SDR_{OOG}}$ from both (2) & (3) is converted to SDR encoded value $E_{SDR_{RGB}}$ by BT.1886 [77] OETF (approximate gamma2.22) and 8bit JPEG compression with QF=80.

Our philosophy lies in that most explicitly-defined operations can be streamlined to a $\text{clamp}(-)$ function: Only when luminance and color volume is clipped, multiple-to-one mapping occurs in DM, the trained network can learn corresponding recovery ability. Also, **ALL** & **ASL** in Tab.6 show that these DMs produce SDR with reasonable style, ensuring that network will not learn style deterioration.

4. Experiments

**Training detail.** For each HDR frame randomly resized to [0.25x,1x], 6 patches sized 600 × 600 are cropped from random position. Then, each of 6×3878 HDR patches is degraded to SDR by 1 of the 3 proposed DMs in equal probability, and again stored in JPEG with QF=75.

All LSN parameters are Kaiming initialized, optimized by $l_1$ loss and AdaM with learning rate starting from $2 \times 10^{-4}$ and decaying to half every $10^5$ iters till $4 \times 10^5$.

4.1. Criteria and configuration

**Criteria.** Since the destination of SDR-to-HDRTV up-conversion is human perception, to provide a ‘step-change’ improvement [1] of HDRTV than SDRTV, methods should be assessed based on if: C1: result is visually-pleasing, from both brightness (**CIA**) & color appearance (**CIB**) etc., C2: HDRTV’s advance on HDR/WCG volume is properly recovered and utilized, C3: bright and dark areas are recovered or at least enhanced and C4: artifacts are avoided. Yet, as found by [18,19], current distance-based metrics e.g. PSNR fails, since the main contributor of an appealing score lies in the accuracy of learned numerical distribution, rather C1-4. Hence, we use new assessment criteria in Tab.7.

<table>
<thead>
<tr>
<th>Visuals in Fig.1&amp;6</th>
<th>Metrics in Tab.8</th>
<th>subj. exp. (4.3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIA</td>
<td>ALL, HDRBQ</td>
<td>overall rating &amp; selection of attribution</td>
</tr>
<tr>
<td>CIB</td>
<td>ASL</td>
<td></td>
</tr>
<tr>
<td>C3 yellow&amp;blue boxes</td>
<td>&quot;</td>
<td></td>
</tr>
<tr>
<td>C4 detailed visuals</td>
<td>&quot;</td>
<td></td>
</tr>
<tr>
<td>C1 3D Yxy diagram</td>
<td>FHL/WG, FHL/EHL/WG</td>
<td></td>
</tr>
</tbody>
</table>

Table 7. How each criteria C1-4 is assessed: via detailed visuals, fine-grained tailored metrics and subjective experiment.

**Comparitors.** As in Tab.8, our method is compared with 6 learning-based methods [4-9], and 2 commercial software DaVinci and Nuke. Results from joint-SR methods [4-6] are **bilinear** downscaled before comparison. Note that learning-based methods are not re-trained with our dataset since we treat dataset as a crucial attribution to performance.

**Test set.** We select 12 4K SDR videos with 10s duration and 50fps. 3 of 12 are degraded from HDR-GT counterpart by **YouTube** DM (§2.3), the rest 9 are **graded** from HDR sequence No. 2,13,15,20,21,22,42,44,55 in [78]. Experiment fairness is guaranteed since both DMs (SDR-HDR relationship) in test set are ‘unseen’ by our network (this even gives a bonus to methods trained with **YouTube** DM [4,5,7-9]).

4.2. Result

Results from all competitors are shown in Tab.8 & Fig.6.

**Metrics.** From Tab.8 we know that methods trained with **YouTube** DM all tend to produce lower satiation (**ASL**) and brightness (**ALL&HDRBQ**), and their **ASL** is lower even than input SDR. Also, SR-ITM-GAN [6] recovers least HDR&WCG volume since its label HDR (from Zeng20 dataset) is of least **extent of HDR/WCG** (Tab.5). Note that current methods produce adequate **FWGP**, but from 3D Yxy chromaticity diagram in Fig.6 we known that their WCG
pixels mainly exists in low luminance range, which means they are of little help to viewing experience. Our method is able to recover adequate HDR/WCG volume, meanwhile reasonably enhance the brightness and saturation.  

1We also provide conventional PSNR, SSIM, $\Delta E$ (color difference [79]) and VDP3 (HDR-VDP-3 [80]), but they mostly represent output’s closer value with GT (for example, result both dimmer (e.g. Deep SR-ITM [4]) and more vivid (ours) than GT will have a similar low score). thus are helpless for our assessment. This phenomenon was first found by [18, 19], and will be further explained in supplementary material.

**Table 8. Metrics.** We use fine-grained tailored metrics (column 3-9, defined in Tab.4) to assess criteria $C_1$&$C_2$ (§4.1): The more significant metrics in column 3-6 are, the better HDR&WCG volume is recovered ($C_2$) by specific method. Also, column 8-10 closer with GT stands for similar brightness&color appearance. Note that we allow column 8-10 slightly bigger than GT, which means result HDR is reasonably more visual-pleasuring ($C_1$) than GT. Based on this, we highlight those results unfavorable for viewing experience.

**Figure 6. Visuals.** We provide comparisons on 3D $X_y$ chromaticity diagram to assess how HDR/WCG volume is recovered ($C_2$), and detailed visuals on bright (yellow arrow, defined with luminance of same pixel indicated) and dark (blue arrow) areas to assess method’s recover ability ($C_3$). Note that conclusion ‘SDR is more vivid than HDR’ could not be drawn because HDR will appear dimmer than SDR in print version (explained in Fig.1). We hence turn to ASL & HDRBQ (column 8-10) and later subjective experiment to see if HDR is indeed dimmer ($C_1$). Comparison on more scenes will be provided in supplementary material.

**Visuals.** Methods’ recover ability is illustrated by Fig.6: Current ones underperform in both bright (yellow) and dark (blue) areas. Specifically, methods trained with YouTube DM produce dim (HDRBQ) and desaturated (ASL) result, and even get lower luminance than input SDR at same position (arrow in yellow box). This confirms the finding in §2.3 that network will learn to darken and desaturate if DM tend to brighten and over-saturate. Also, our method recovers more detail in bright and dark areas with a better style.
4.3. Subjective experiment

Currently, few subjective studies [19, 43, 59, 81–83] are designed for SDR-to-HDR procedure (rather between different HDR). Similar to [19], we judge if output HDR is better than origin SDR. Specifically, as in Fig. 7, we use 2 side-by-side SONY KD85X9000H display supporting both SDR and HDRTV, one is calibrated to 100nit/BT.709 SDR and another PQ1000nit/BT.2020 HDR. Each (input)SDR-(output)HDR pair is displayed twice with 3s gray level interval and following 10s for rating: each participant is asked to continuously rate from -5(SDR much better) to 0(same) to 5(HDR much better), meanwhile select at least 1 attribution (bottom Fig. 7) of his/her rating. Such process is repeated 9(#participant)×9(#competitor)×12(#evaluation) times.

Result shows that only 2 methods (ours & DaVinci) are recognized better than input SDR. For methods [4–6], their main ‘attribution’ (bottom Fig. 7) is artifact (C4, see Fig. 6 blue box). For artifact-free ‘YouTube-DM’ methods [7–9], they are rated slightly worse mainly for lower saturation (C1B) and incapability in information recovery, consisting with Tab.8 and Fig. 6. Form ‘attribution’ we also notice that our method got better score mainly for viewing experience (C1) and information recovery (C3). Also, our ‘bad cases’ lies in recovered large dark area with intrinsic noise etc. amplified and uneliminated. This is why we got more checks on C4(×24) than commercial methods DaVinci and Nuke.

4.4. Ablation studies

On DM. When DM is changed to YouTube, Tab.8 witnesses a significant decline on FHLP, EHL, ASL, ALL and HDRBQ, while Fig. 8 confirms a result similar to those ‘YouTube-DM’ methods, i.e. worse viewing experience and less recover ability. Also, when using Reinhard DM which contains no clipping, result’s highlight area stay unlearned.

On dataset. Here, we use original DMs, but label HDR from other dataset. In Fig. 8 yellow box, our DMs encourage the network to output higher luminance, but since Zeng20 is of least extent of HDR i.e. these highlight do not exist in label HDR, our LSN will not ‘recognize’ them and thus produce artifact. Since this dataset is also of least extent of WCG, FWGP&EWG in Tab.8 drop obviously. When using slightly-inferior HDRTV1K as label, difference is relatively less significant. Yet, in both cases, ASL&ALL are similar since DM i.e. network’s style tendency is unaltered.

Figure 7. Environment and result of subjective experiment. 2 displays are calibrated differently to ensure both SDR and HDR are correctly visualized. Results are provided in quartile chart.

Figure 8. Result of ablation studies proves the importance of both high-quality label HDR and rational DMs. Specifically, absent of both label HDR’s HDR/WCG volume (Zeng20) and DM’s degradation (Reinhard) will impair LSN’s recover ability, meanwhile YouTube DM’s style will make our LSN commonplace as others.

5. Conclusion

There are 2 types of low level vision: ‘sole-restoration’ whose destination is only clean or GT e.g. denoising, and ‘perceptual-motivated’ aiming at better viewing experience e.g. image enhancement/retouching. SDR-to-HDRTV up-conversion belongs to both. Yet, current methods only realize (it belongs to) the former and neglect the latter, leading their concentration only on network mechanism.

To this end, our response is two-fold: (1) focusing on the impact of training set, and ameliorating its quality by proposing new dataset and DMs, (2) involving novel assessment criteria based on the ‘perceptual’ principal.
References


22239


[77] ITU, Geneva, Switzerland, Recommendation ITU-R BT.1886-0: Reference electro-optical transfer function for flat panel displays used in HDTV studio production, 0 ed., 3 2011. 6


