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

Cheng Guo[1,2], Leidong Fan[3,2], Ziyu Xue[4,1] and Xiuhua Jiang[2,1]
[1]State Key Laboratory of Media Convergence and Communication, Communication University of China
[2]Peng Cheng Laboratory [3]Peking University
[4]Academy of Broadcasting Science, National Radio and Television Administration
{guocheng,jiangxiuhua}@cuc.edu.cn, fanleidong@stu.pku.edu.cn, xueziyu@abs.ac.cn

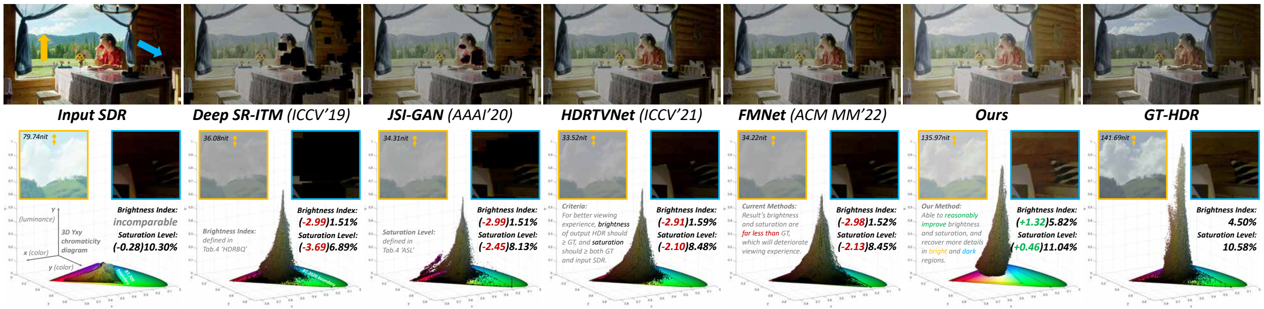

Figure 1. **Up-converting SDR content for HDR-WCG TV display**. To ensure better viewing experience [1], we use (1) *3D Yxy chromaticity diagram* (vertical axis *Y* for HDR/lumiannce, *xy* plane for WCG/color) to see how HDRTV's advance on HDR&WCG volume is recovered, and (2) *detailed visuals* (yellow and blue boxes) to assess method's recover ability. Note that HDR is dimer here in print version since its large luminance&color container is interpreted by small SDR capacity, and will appear normal if correctly visualized (top Fig.7). Still, from (3) '*Brightness Index*' & '*Saturation Level*' we know that result from **current methods** is more **dim and desaturated** than GT.

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 vision 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 understate 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 incentive 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.

	Task name	**From**	**To** single	**Methods**
①	HDR-style enhancement	single SDR	enhanced SDR	*e.g*. [23–26]
②	multi-exposure HDR imaging	multiple SDR	linear HDR	*e.g*. [27–32]
③	SI-HDR*	single SDR		*e.g*. [33–37]
④ (ours)	iTM* or up-convertion		HDRTV frame	[4–9] [38–53]

*: SI-HDR: Single-Image HDR reconstruction, iTM: inverse Tone-Mapping.

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

In our scope (④), methods designed their networks with distinct philosophy: Except *e.g*. semantic-requiring highlight 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.

Dataset (Usage)	**#pair**	**Resolution**	**HDR format**
KAIST [4] ([5,48,52])	39840	160×160	uint16 MATLAB .mat YUV
Zeng20 [6]	23229	UHD	H.265 main10 YUV
HDRTV1K [7] ([9,47,50])	1235		16bit .png RGB
HDRTV4K (ours new)	3878	HD&UHD	16bit lossless .tif RGB

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.

2.3. 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):

DM	***YouTube***	***Reinhard***	other DMs
Usage	[4, 5, 7–9, 40, 44, 46–50, 52]	[6, 38, 41–43]	***2446a*** [45] *etc*. [51, 53, 56]
Dataset	KAIST & HDRTV1K	Zeng20	

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 $\mathbf{x}$ the SDR, $\mathbf{y}$ the HDR. Previous methods [7, 47, 49] hypothesis that $\mathbf{x}$ and $\mathbf{y}$ are from single RAW file of same camera.

This applies to SDR-HDR synchronous production where $\mathbf{x}$ is also of high quality, thus up-conversion ($\mathbf{y} = f(\mathbf{x})$) is only supposed to follow specific style and recover less missing information. However, since our method is for existing SDR, we assume that SDR ($\mathbf{x}$) and it imaginary HDR counterpart ($\mathbf{y}$) 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 ($\mathbf{x} = d(\mathbf{y})$).

Some works [7, 35] suppose $d(\cdot)$ are introduced orderly, and assign hierarchical sub-networks to for $f(\cdot)$. 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 ($\mathbf{x}_h$), *low-luminance range* ($\mathbf{x}_l$) is more susceptible to noise *etc*., while the quality of *mid-luminance range* ($\mathbf{x}_m$) 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* ($\mathbf{x}$), and 2 branches respectively on $\mathbf{x}_l$ and $\mathbf{x}_h$ which are segmented by:

$$\mathbf{x}_l = max(0, \frac{t - \mathbf{x}}{t}),\ \mathbf{x}_h = max(0, \frac{\mathbf{x} - 1}{t} + 1) \qquad (1)$$

where $\mathbf{x} \in [0, 1]$. 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, $\mathbf{x}$, $\mathbf{x}_l$ and $\mathbf{x}_h$ require distinct DNN operation. First, as found by [7] *etc*. (§2.1), the majority of $f(\cdot)$ 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* ($\mathbf{x}$) as the **Global Mapping Trunk**. Similar to [7, 9, 47–50], we append a modulation branch (Fig.3 left) to involve $\mathbf{x}$'s global prior into DNN's decision.

On the other hand, $f(\cdot)$ still involves non-global restoration where local semantic should be aggregated. Specifically, DNN's mission in $\mathbf{x}_l$ is similar to denoising and low-light enhancement, while that in $\mathbf{x}_h$ 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 $\mathbf{x}_l$ and $\mathbf{x}_h$ 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:

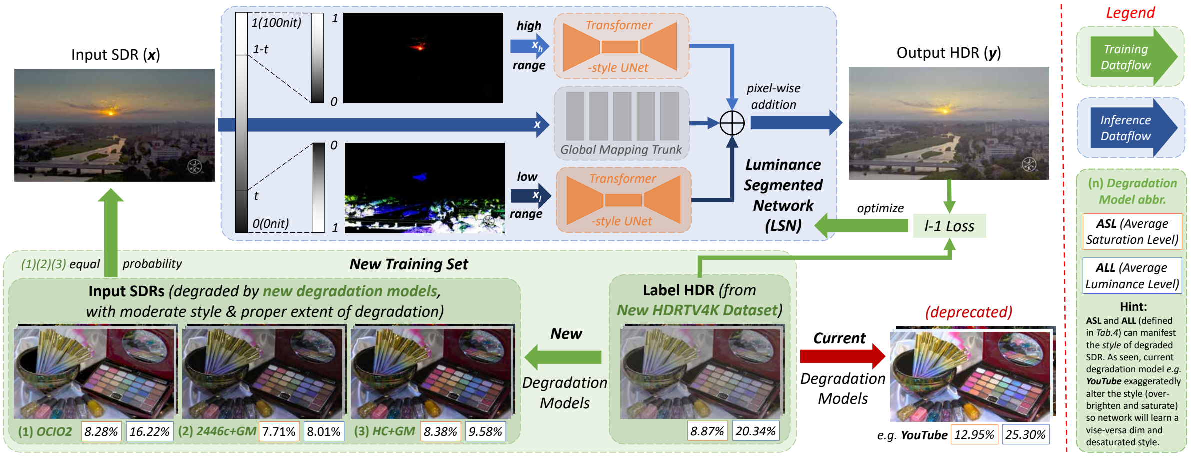
Input SDR (x)
1(100nit)
1-t
t
0(0nit)
1
0
high
x_h
range
low
x_l
range
x
Transformer-style UNet
Global Mapping Trunk
pixel-wise addition
Luminance Segmented Network (LSN)
Output HDR (y)
optimize
l-1 Loss
(1)(2)(3) equal probability
New Training Set
Input SDRs (degraded by new degradation models, with moderate style & proper extent of degradation)
(1) OCIO2 8.28% 16.22%
(2) 2446c+GM 7.71% 8.01%
(3) HC+GM 8.38% 9.58%
New Degradation Models
Label HDR (from New HDRTV4K Dataset)
8.87% 20.34%
Current Degradation Models
(deprecated)
e.g. YouTube 12.95% 25.30%
Legend
Training Dataflow
Inference Dataflow
(n) Degradation Model abbr.
ASL (Average Saturation Level)
ALL (Average Luminance Level)
Hint:
ASL and ALL (defined in Tab.4) can manifest the style of degraded SDR. As seen, current degradation model e.g. YouTube exaggeratedly alter the style (over-brighten and saturate) so network will learn a vise-versa dim and desaturated style.

Figure 2. Overview of this work. Our Luminance Segmented Network (LSN, §3.1) is designed based on novel problem fromulation, then supervisedly trained with label HDR from the proposed HDRTV4K dataset (§3.2), and input SDR degraded by novel degradation models (DMs, §3.3). The **major concerns** of our LSN, HDRTV4K dataset, and DMs are respectively: recovering dark&bright areas, improving the quality (Tab.5) and diversity (Fig.4) of label GT-HDR, and ensuring the LQ-SDR is with proper style and degradation (Tab.6 & Fig.5).

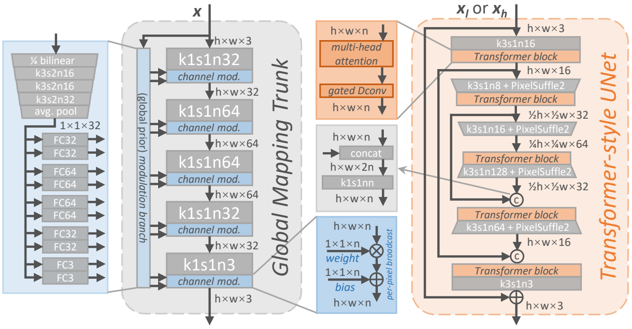

Figure 3. Detailed structure of LSN modules. 'k3s1n16' means convolution layer with 3×3 `kernel`, `stride`=1, `out nc`=16, and 'FC32' is fully-connected layer with `out nc`=32. We deploy only 1 Transformer block [17] at each en/decode level of UNet to lighten LSN (#param=325k) for its application on 4K resolution.

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.

Metrics on the *extent of HDR/WCG* ↓	
FHLP	**F**raction of **H**igh**L**ight **P**ixel: Spatial ratio of 'highlight' pixel *i.e.* whose normalized luminance $Y = 0.2627R + 0.6780G + 0.0593B > 0.01$ ($100nit$, SDR's peak luminance.)
EHL	**E**xtent of **H**igh**L**ight: Average pxiel (i) distance between the luminance of HDR and its clip-to-$100nit$ version[1]: $\frac{1}{n}\sum_{i=1}^{n}\sqrt{[Y_i - clip(Y_i)]^2},\ clip(x) = clamp(x, 0, 0.01)$
FWGP	**F**raction of **W**ide-**G**amut **P**ixel: Spatial ratio of WCG pixel *i.e.* whose $[x, y]$ coordinates fall inside BT.2020 but outside SDR's BT.709 gamut in Yxy chromaticity diagram.
EWG	**E**xtent of **W**ide-**G**amut [62]: Average pixel-distance between WCG-HDR and its gamut-hard-clipped [63] version[2]: $\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{S}_i - HC(\mathbf{S}_i)\|_2,\ \mathbf{S} = [X, Y, Z]^{\mathrm{T}}$
Metrics on *intra-frame diversity* ↓ (all variance-based)	
SI	**S**patial **I**nformation: Standard deviation over the pixels of Sobel-filtered frame, defined in Annex 6 of [64].
CF	**C**olor**F**ulness: Defined in [65].
stdL	standard deviation of **L**uminance, over all pixels of a frame.
Metrics on *overall-style* ↓	
ASL	**A**verage **S**aturation **L**evel: Normalized pixel-average length of HDR chrominance component $\mathbf{C} = [C_t, C_p]^{\mathrm{T}}$ [66]: $\frac{\sqrt{2}}{n}\sum_{i=1}^{n}\|\mathbf{C}_i\|_2,\ \mathbf{C} \in [-0.5, 0.5]$
ALL	**A**verage **L**uminance **L**evel: Pixel-average of Y in **FHLP**.
HDRBQ	**HDR B**rightness **Q**uantification [67], visual salience involved.

[1]: **EHL** is to compensate cases *e.g.* an all-$101nit$ HDR frame with 100% *FHLP* but less extent of highlight.

[2]: You can find the formulation of gamut hard-clipping ($HC(\cdot)$) at Eq.5.

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 ([68–72] *etc.*) with greater *extent of HDR/WCG*, higher *intra-frame diversity*

Metrics / Dataset	*Extent of HDR*		*Extent of WCG*		*Intra-frame diversity*			*Overall-style*		
	FHLP	**EHL**	**FWGP**	**EWG**	**SI**	**CF**	**stdL**	**ASL**	**ALL**	**HDRBQ**
KAIST [4]	1.5250	0.2025	5.4771	0.1104	1.9372	5.9485	0.9597	8.9087	17.2854	1.8597
Zeng20 [6]	0.0197	0.0012	0.4792	0.0034	0.1231	4.2048	0.3146	3.8061	6.0805	0.3781
HDRTV1K [7]	1.2843	0.1971	2.9089	0.1633	2.2378	11.0722	1.8006	10.9414	15.1626	2.7970
HDRTV4K (ours)	5.3083	0.9595	2.6369	0.5123	3.5508	10.5882	3.4837	9.8274	21.1996	5.1593

Table 5. Quality of label HDR frames, manifested in the frame-average of 10 metrics from Tab.4. Greater *extent of HDR/WCG* encourages network to produce more pixels in non-SDR volume, while *intra-frame diversity* is helpful for network's generalization ability. All numbers are in percentage (%), and we highlight those the most favorable to training. Their diversity is further demonstrated in Fig.4.

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.

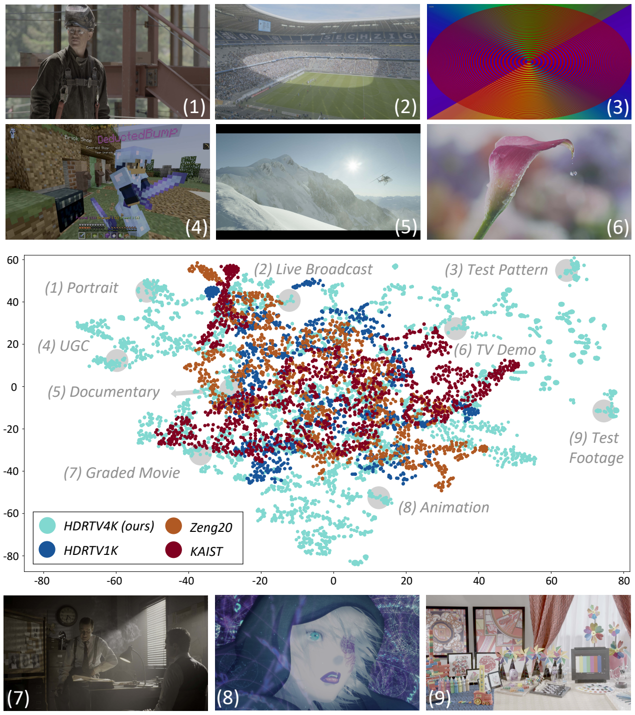

Figure 4. Diversity comparison: our HDRTV4K dataset *vs*. others. Here, each 2D-coordinate is the projection of single frame's 10-D vector (containing 10 metrics from Tab.5) using t-SNE [73] (`exg=3`, `prep=50`). We also depict thumbnail HDR frames from our dataset (1)-(9) with their corresponding 2D-coordinates highlighted with gray circle. As seen, our dataset provides wider frame distribution *i.e*. more diversified scenes for network to learn.

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 Tab.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.

SDR Degraded by		**FOEP**[1]	**ALL**	**ASL**[2]	**SI**	**CF**
Cur-rent DM	***2446a***	0.181	5.880	3.624	6.041	6.573
	Reinhard	1.264	21.887	7.442	14.147	12.568
	YouTube	5.439	28.219	14.641	18.545	25.225
Ours DM	***2446c+GM***	1.739	11.669	10.183	12.503	19.391
	HC+GM	4.252	14.062	10.377	15.090	20.146
	OCIO2	1.580	18.887	9.977	13.578	18.052
criteria↑: better when		kept[3]	samller	kept	-	-
Sourse HDR (Tab.5)[4]		5.308	21.200	9.827	3.551	10.588

[1] : **FOEP**: **F**raction of **O**ver-**e**xposed **P**ixels: Spatial ratio of pixels whose normalized lumiannce $Y = 1$.

[2] : For SDR, **ASL** is also calculated as Tab.4, but with $\mathbf{C} = [C_b, C_r]^{\mathrm{T}}$ [55] rather $[C_t, C_p]^{\mathrm{T}}$.

[3] : Assuming that HDR's highlight (>$100nit$) part should all be clipped to 1 (over-exposure) in SDR.

[4] : Gray means container discrepancy *i.e*. HDR/SDR's metric should be different even for well-degraded SDR.

Table 6. Given label HDR from HDRTV4K dataset, SDR's statistics alters when degraded by different DMs. Results are in percentage, we mark those unfavorable / neutral / beneficial for training based on the observation in §2.3 & 3.3. As seen, our DMs provide adequate over-exposure (**FOEP**) and no undue brightness (**ALL**) & saturation (**ASL**) change for network to learn. Example of different degraded SDR can be found in supplementary material.

This motivate us to utilize/exploit new DMs with proper extent of degradation and no deficiencies on style:

(1) ***OCIO2***: *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 $\mathbf{E}' = [R', G', B']^{\mathrm{T}}$ (we use superscript $'$ for non-linearity) is linearize to $\mathbf{E}$ by PQ EOTF, and then transferred to Yxy. Then Y is globally mapped to Y_{SDR}, while xy are kept and then recomposed with Y_{SDR} to form $\mathbf{E}_{SDR}$ and later $\mathbf{E}'_{SDR_{709}}$. Our modification lies in:

$$Y_{SDR} = clamp(TM_{2446c}(Y), 0, 100nit) \qquad (2)$$

where $TM_{2446c}(\cdot) \in [0, 118nit]$ 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.

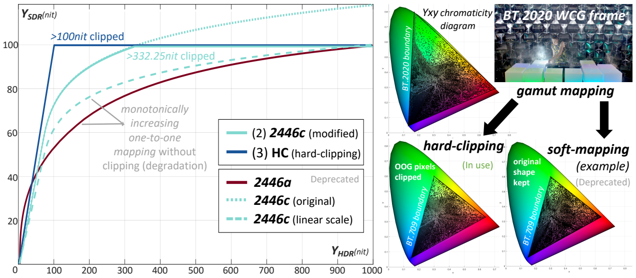

Figure 5. One of the key ideas of the proposed degradation models (DMs, §3.3): HDR's luminance (left) and color (right) volume should be properly clipped, so that network will learn corresponding restoration. Another idea is keeping a sensible brightness & color/saturation *style* during degradation, 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 luniannce part, we clip all >$100nit$ pixels to $100nit$:

$$\mathbf{E}_{SDR} = 100 \times clamp(\mathbf{E}, 0, 0.01),\ \mathbf{E} \in [0, 1] \tag{3}$$

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 $10000nit \rightarrow 100nit$).

So far, $\mathbf{E}_{SDR}$ 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:

$$\mathbf{E}_{SDR_{OOG}} = \mathbf{M}\mathbf{E}_{SDR},\ \mathbf{M} = \begin{bmatrix} 1.6605 & 0.5876 & -0.0728 \\ -0.1246 & 1.1329 & 0.0083 \\ -0.0182 & -0.1006 & 1.1187 \end{bmatrix} \tag{4}$$

where $\exists\ \mathbf{E}_{SDR_{OOG}} \notin [0, 1]$ *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):

$$\mathbf{E}_{SDR_{709}} = clamp(\mathbf{E}_{SDR_{OOG}}, 0, 1) \tag{5}$$

Then, $\mathbf{E}_{SDR_{709}}$ from both (2) & (3) is converted to SDR encoded value $\mathbf{E}'_{SDR_{709}}$ 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 $clamp(\cdot)$ 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 2×10^{-4} and decaying to half every 10^5 iters till 4×10^5.

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-pleasuring, from both brightness (***C1A***) & color appearance (***C1B***) *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.

	Visuals in Fig.1&6	**Metrics** in Tab.8	**subj. exp.** (§4.3)
C1A	-	**ALL, HDRBQ**	overall rating & selection of attribution
C1B		**ASL**	
C3	yellow&blue boxes	-	
C4	detailed visuals		
C2	*3D Yxy diagram*	**FHL(WG)P, EHL(WG)**	-

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

Competitors. 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

Method	*how network is trained*		(recovery rate %, GT is 100%) *how HDR/WCG volume is recovered*				(shift rate %, GT is 0%) *overall-style*			*conventional metrics*			
(network)	dataset (GT)	DM	**FHLP**	**EHL**	**FWGP**	**EWG**	**ASL**	**ALL**	**HDRBQ**	**PSNR**	**SSIM**	**△E**	**VDP3**
Input SDR	-		0	0	0	0	6.570	10.76	-	23.92dB	0.8861	44.97	6.571
Deep SR-ITM [4]	KAIST	***YouTube***	(14.03)0.2323	(9.70)0.372	(175.2)1.0964	(71.96)0.172	(-20.33)5.485	(-15.06)9.580	(-72.80)1.428	26.59dB	0.8115	32.54	6.917
JSI-GAN [5]			(12.33)0.2041	(3.55)0.136	(213.1)1.3334	(88.23)0.212	(-16.62)5.741	(-14.36)9.659	(-79.20)1.092	27.87dB	0.8420	30.23	7.452
SR-ITM-GAN [6]	Zeng20	***Reinhard***	(8.04)0.1332	(14.33)0.550	(00.00)0.0000	(00.00)0.000	(-7.04)6.400	(-30.37)7.853	(-57.12)2.251	28.04dB	0.8831	24.78	6.707
HDRTVNet [7]	HDRTV1K	***YouTube***	(18.59)0.3078	(16.29)0.625	(388.8)2.4334	(28.82)0.069	(-15.51)5.817	(-13.47)9.759	(-69.65)1.593	30.82dB	0.8812	27.58	8.120
KPN-MFI [8]	own		(1.17)0.0193	(0.02)0.001	(392.9)2.4592	(15.82)0.038	(-21.74)5.388	(-16.10)9.462	(-82.47)0.920	29.37dB	0.8746	27.47	7.785
FMNet [9]	HDRTV1K		(13.67)0.2264	(12.35)0.474	(396.5)2.4813	(91.29)0.220	(-16.91)5.770	(-13.48)9.758	(-71.24)1.510	30.91dB	0.8855	27.16	8.069
LSN **(ours)**	HDRTV4K	**ours**×3	(256.6)4.2509	(71.96)2.599	(109.8)0.6873	(150.0)0.361	(+9.25)7.522	(+81.12)20.42	(-27.04)3.829	24.47dB	0.8310	37.84	8.130
DaVinci	-		(54.96)0.9103	(43.16)1.655	(310.0)1.9399	(85.22)0.205	(-21.36)5.414	(-21.42)8.863	(-39.23)3.190	26.39dB	0.8918	35.47	8.528
Nuke			(112.1)1.8565	(24.70)0.384	(00.00)0.0000	(00.00)0.000	(-27.53)4.990	(+17.93)13.30	(-51.19)2.562	20.87dB	0.7273	64.29	7.479
HDR-GT (ref.)			(100.0)1.6562	(100.0)3.835	(100.0)0.6258	(100.0)0.240	(0.00)6.885	(0.00)11.28	(0.00)5.248	-	-	-	-
ablation studies ↓													
LSN **(ours)**	HDRTV4K	***YouTube***	(13.09)0.2168	(6.63)0.254	(401.3)2.5118	(59.81)0.144	(-14.68)5.874	(-13.46)9.760	(-76.89)1.213	30.15dB	0.8858	28.04	7.902
	HDRTV4K	***Reinhard***	(9.06)0.1501	(19.14)0.734	(00.00)0.0000	(00.00)0.000	(-5.74)6.490	(-25.65)8.387	(-52.84)2.475	27.70dB	0.8436	26.03	6.832
	Zeng20	**ours**×3	(109.6)1.8147	(9.00)0.345	(0.95)0.0593	(0.76)0.002	(+3.36)7.117	(+66.64)18.79	(-59.52)2.124	25.17dB	0.8179	28.58	7.874
	HDRTV1K	**ours**×3	(177.5)2.9405	(52.49)2.013	(279.5)1.7494	(70.43)0.169	(+6.56)7.337	(+52.30)17.18	(-34.83)3.420	24.30dB	0.8351	38.03	8.002

Table 8. Metrics. We use **fine-grained tailored metrics** (*column 3-9*, defined in Tab.4) to assess criteria ***C1&2*** (§4.1): The more significant metrics in *column 3-6* are, the better HDR&WCG volume is recovered (***C2***) 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 (***C1***) than GT. Based on this, we highlight those results unfavorable for viewing experience.

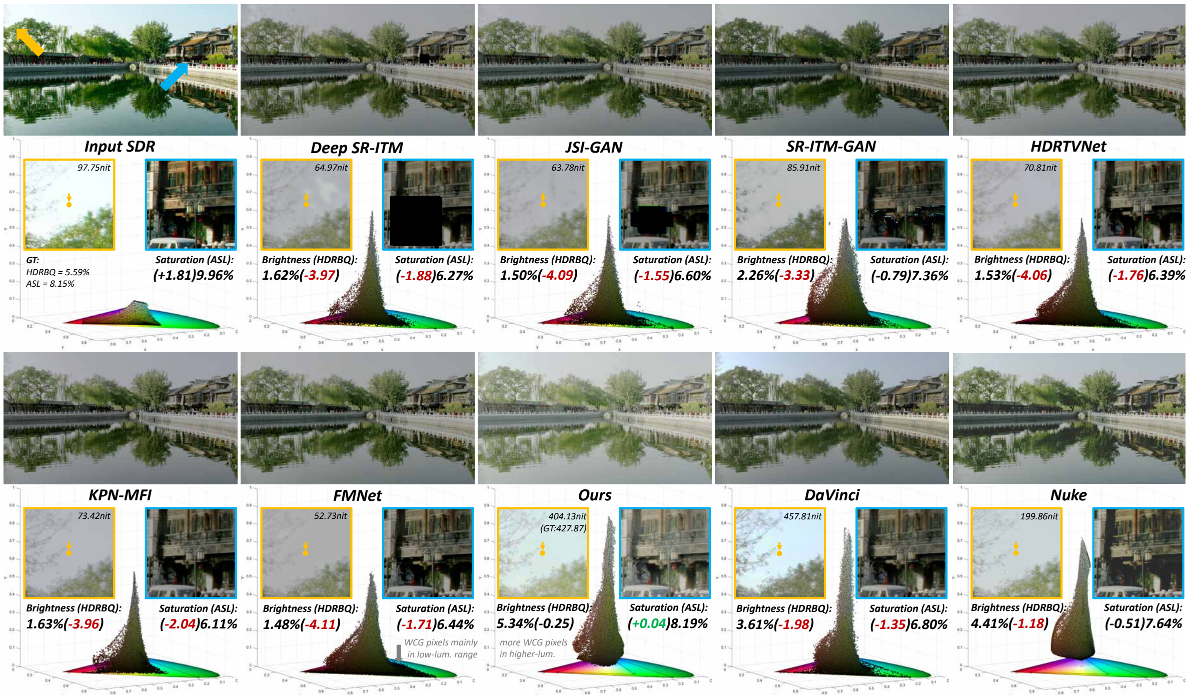

Figure 6. Visuals. We provide comparisons on *3D Yxy chromaticity diagram* to assess how HDR/WCG volume is recovered (***C2***), and detailed visuals on bright (yellow arrow, with luminance of same pixel indicated) and dark (blue arrow) areas to assess method's recover ability (***C3***). Note that conclusion '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 (***C1***). Comparison on more scenes will be provided in supplementary material.

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. [1]

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.

[1] We also provide conventional PSNR, SSIM, ΔE (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.

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(#clip) times.

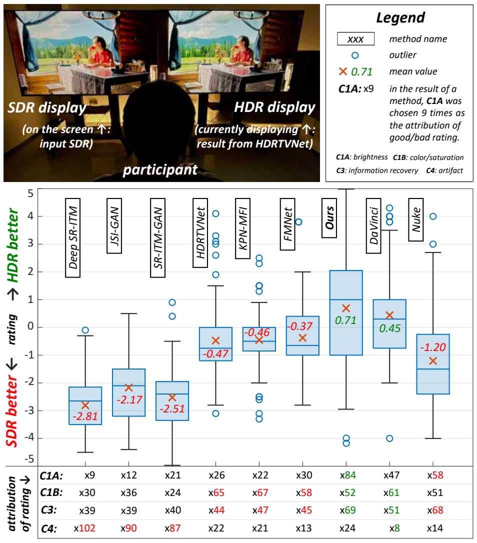

Figure 7. Environment and result of subjective experiment. 2 displays are caliberated differently to ensure both SDR and HDR are **correctly visualized**. Results are provided in quartile chart.

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.

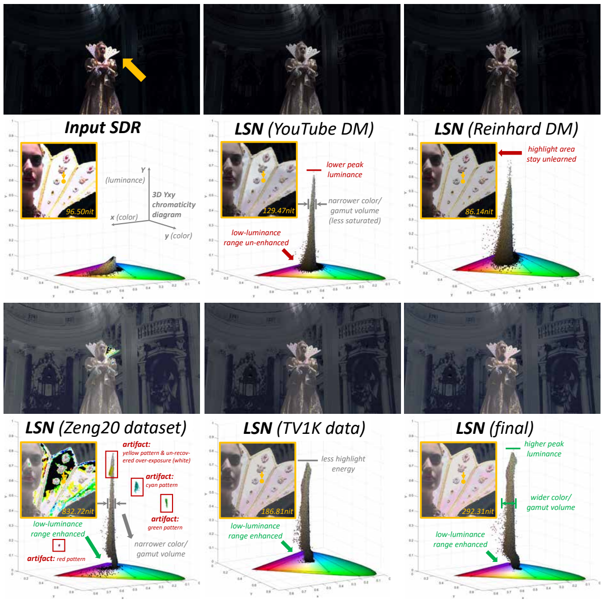

Figure 8. Result of ablation studies proves the importance of both high-quality label HDR and rational DMs. Specifically, absent of both label HDR's HDR/WCG volume (Zeng20) and DM's degradation (***Reinhard***) will impair LSN's recover ability, meanwhile ***YouTube*** DM's *style* will make our LSN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

[1] ITU, Geneva, Switzerland, *Report ITU-R BT.2381-0: Requirements for high dynamic range television (HDR-TV) systems*, 0 ed., 7 2015. 1, 6

[2] ITU, Geneva, Switzerland, *Recommendation ITU-R BT.2100-2: Image parameter values for high dynamic range television for use in production and international programme exchange*, 2 ed., 07 2018. 1, 2

[3] ITU, Geneva, Switzerland, *Recommendation ITU-R BT.2020-2: Parameter values for ultra-high definition television systems for production and international programme exchange*, 2 ed., 10 2015. 1, 2

[4] S. Y. Kim, J. Oh, and M. Kim, "Deep sr-itm: Joint learning of super-resolution and inverse tone-mapping for 4k uhd hdr applications," in *Proc. ICCV*, pp. 3116–3125, 2019. 2, 3, 5, 6, 7, 8

[5] S. Y. Kim, J. Oh, and M. Kim, "Jsi-gan: Gan-based joint super-resolution and inverse tone-mapping with pixel-wise task-specific filters for uhd hdr video," in *Proc. AAAI*, vol. 34, pp. 11287–11295, 2020. 2, 3, 6, 7, 8

[6] H. Zeng, X. Zhang, Z. Yu, and Y. Wang, "Sr-itm-gan: Learning 4k uhd hdr with a generative adversarial network," *IEEE Access*, vol. 8, pp. 182815–182827, 2020. 2, 3, 5, 6, 7, 8

[7] X. Chen, Z. Zhang, J. S. Ren, L. Tian, Y. Qiao, and C. Dong, "A new journey from sdrtv to hdrtv," in *Proc. ICCV*, pp. 4500–4509, 2021. 2, 3, 5, 6, 7, 8

[8] G. Cao, F. Zhou, H. Yan, A. Wang, and L. Fan, "Kpn-mfi: A kernel prediction network with multi-frame interaction for video inverse tone mapping," in *Proc. IJCAI*, pp. 806–812, 2022. 2, 3, 6, 7, 8

[9] G. Xu, Q. Hou, L. Zhang, and M.-M. Cheng, "Fmnet: Frequency-aware modulation network for sdr-to-hdr translation," in *Proc. ACMMM*, pp. 6425–6435, 2022. 2, 3, 6, 7, 8

[10] Z. Liang, "Cvpr2022-1st workshop on vision dataset understanding." `https://sites.google.com/view/vdu-cvpr22`, 2022. 2

[11] K. Zhang, J. Liang, L. Van Gool, and R. Timofte, "Designing a practical degradation model for deep blind image super-resolution," in *Proc. ICCV*, pp. 4791–4800, 2021. 2, 3

[12] V. Dewil, A. Barral, G. Facciolo, and P. Arias, "Self-supervision versus synthetic datasets: which is the lesser evil in the context of video denoising?," in *Proc. CVPR*, pp. 4900–4910, 2022. 2, 3

[13] J. Jiang, K. Zhang, and R. Timofte, "Towards flexible blind jpeg artifacts removal," in *Proc. ICCV*, pp. 4997–5006, 2021. 2, 3

[14] K. Jiang, Z. Wang, Z. Wang, C. Chen, P. Yi, T. Lu, and C.-W. Lin, "Degrade is upgrade: Learning degradation for low-light image enhancement," in *Proc. AAAI*, vol. 36, pp. 1078–1086, 2022. 2, 3

[15] Y. Zhou, Y. Song, and X. Du, "Modular degradation simulation and restoration for under-display camera," in *Proc. ACCV*, pp. 265–282, 2022. 2, 3

[16] C. Guo and X. Jiang, "Lhdr: Hdr reconstruction for legacy content using a lightweight dnn," in *Proc. ACCV*, pp. 3155–3171, 2022. 2, 3

[17] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proc. CVPR*, pp. 5728–5739, 2022. 2, 3, 4

[18] G. Eilertsen, S. Hajisharif, P. Hanji, A. Tsirikoglou, R. K. Mantiuk, and J. Unger, "How to cheat with metrics in single-image hdr reconstruction," in *Proc. ICCV*, pp. 3998–4007, 2021. 2, 3, 6, 7

[19] P. Hanji, R. Mantiuk, G. Eilertsen, S. Hajisharif, and J. Unger, "Comparison of single image hdr reconstruction methods—the caveats of quality assessment," in *Proc. SIGGRAPH*, pp. 1–8, 2022. 2, 6, 7, 8

[20] L. Wang and K.-J. Yoon, "Deep learning for hdr imaging: State-of-the-art and future trends," *IEEE Trans. PAMI*, 2021. 2

[21] P. Debevec, "Image-based lighting," in *ACM SIGGRAPH 2006 Courses*, pp. 4–es, 2006. 2

[22] E. Reinhard, W. Heidrich, P. Debevec, *et al.*, *High dynamic range imaging: acquisition, display, and image-based lighting*. Morgan Kaufmann, 2010. 2

[23] M. Gharbi, J. Chen, J. T. Barron, S. W. Hasinoff, and F. Durand, "Deep bilateral learning for real-time image enhancement," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–12, 2017. 2

[24] X. Yang, K. Xu, Y. Song, Q. Zhang, X. Wei, and R. W. Lau, "Image correction via deep reciprocating hdr transformation," in *Proc. CVPR*, pp. 1798–1807, 2018. 2

[25] Z. Zheng, W. Ren, X. Cao, T. Wang, and X. Jia, "Ultra-high-definition image hdr reconstruction via collaborative bilateral learning," in *Proc. ICCV*, pp. 4449–4458, 2021. 2

[26] B. Mildenhall, P. Hedman, R. Martin-Brualla, P. P. Srinivasan, and J. T. Barron, "Nerf in the dark: High dynamic range view synthesis from noisy raw images," in *Proc. CVPR*, pp. 16190–16199, 2022. 2

[27] N. K. Kalantari, R. Ramamoorthi, *et al.*, "Deep high dynamic range imaging of dynamic scenes.," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 144–1, 2017. 2

[28] S. Wu, J. Xu, Y.-W. Tai, and C.-K. Tang, "Deep high dynamic range imaging with large foreground motions," in *Proc. ECCV*, pp. 117–132, 2018. 2

[29] Q. Yan, D. Gong, Q. Shi, A. v. d. Hengel, C. Shen, I. Reid, and Y. Zhang, "Attention-guided network for ghost-free high dynamic range imaging," in *Proc. CVPR*, pp. 1751–1760, 2019. 2

[30] G. Chen, C. Chen, S. Guo, Z. Liang, K.-Y. K. Wong, and L. Zhang, "Hdr video reconstruction: A coarse-to-fine network and a real-world benchmark dataset," in *Proc. ICCV*, pp. 2502–2511, 2021. 2

[31] Y. Niu, J. Wu, W. Liu, W. Guo, and R. W. Lau, "Hdr-gan: Hdr image reconstruction from multi-exposed ldr images with large motions," *IEEE Transactions on Image Processing*, vol. 30, pp. 3885–3896, 2021. 2

[32] E. Pérez-Pellitero *et al.*, “Ntire 2022 challenge on high dynamic range imaging: Methods and results,” in *Proc. CVPR*, pp. 1009–1023, 2022. 2

[33] G. Eilertsen, J. Kronander, *et al.*, “Hdr image reconstruction from a single exposure using deep cnns,” *ACM Trans. Graph.*, vol. 36, no. 6, pp. 1–15, 2017. 2, 3

[34] D. Marnerides, T. Bashford-Rogers, *et al.*, “Expandnet: A deep convolutional neural network for high dynamic range expansion from low dynamic range content,” *Comput. Graph. Forum*, vol. 37, no. 2, pp. 37–49, 2018. 2

[35] Y.-L. Liu, W.-S. Lai, Y.-S. Chen, Y.-L. Kao, M.-H. Yang, Y.-Y. Chuang, and J.-B. Huang, “Single-image hdr reconstruction by learning to reverse the camera pipeline,” in *Proc. CVPR*, pp. 1651–1660, 2020. 2, 3

[36] M. S. Santos, T. I. Ren, and N. K. Kalantari, “Single image hdr reconstruction using a cnn with masked features and perceptual loss,” *ACM Trans. Graph.*, vol. 39, no. 4, pp. 80–1, 2020. 2

[37] X. Chen, Y. Liu, *et al.*, “Hdrunet: Single image hdr reconstruction with denoising and dequantization,” in *Proc. CVPR*, pp. 354–363, 2021. 2

[38] S. Ning, H. Xu, L. Song, R. Xie, and W. Zhang, “Learning an inverse tone mapping network with a generative adversarial regularizer,” in *Proc. ICASSP*, pp. 1383–1387, 2018. 2, 3

[39] K. Hirao, Z. Cheng, M. Takeuchi, and J. Katto, “Convolutional neural network based inverse tone mapping for high dynamic range display using lucore,” in *2019 IEEE International Conference on Consumer Electronics (ICCE)*, pp. 1–2, 2019. 2

[40] S. Y. Kim, D.-E. Kim, and M. Kim, “Itm-cnn: Learning the inverse tone mapping from low dynamic range video to high dynamic range displays using convolutional neural networks,” in *Proc. ACCV*, pp. 395–409, 2018. 2, 3

[41] Y. Xu, S. Ning, R. Xie, and L. Song, “Gan based multi-exposure inverse tone mapping,” in *Proc. ICIP*, pp. 4365–4369, 2019. 2, 3

[42] Y. Xu, L. Song, R. Xie, and W. Zhang, “Deep video inverse tone mapping,” in *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, pp. 142–147, 2019. 2, 3

[43] D.-E. Kim and M. Kim, “Learning-based low-complexity reverse tone mapping with linear mapping,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 2, pp. 400–414, 2019. 2, 3, 8

[44] J. Zou, K. Mei, and S. Sun, “Multi-scale video inverse tone mapping with deformable alignment,” in *2020 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, pp. 9–12, 2020. 2, 3

[45] T. Chen and P. Shi, “An inverse tone mapping algorithm based on multi-scale dual-branch network,” in *2021 International Conference on Culture-oriented Science & Technology (ICCST)*, pp. 187–191, 2021. 2, 3

[46] G. Xu, Y. Yang, J. Xu, L. Wang, X.-T. Zhen, and M.-M. Cheng, “Joint super-resolution and inverse tone-mapping: A feature decomposition aggregation network and a new benchmark,” 2022. 2, 3

[47] G. He, K. Xu, L. Xu, C. Wu, M. Sun, X. Wen, and Y.-W. Tai, “Sdrtv-to-hdrtv via hierarchical dynamic context feature mapping,” in *Proc. ACMMM*, pp. 2890–2898, 2022. 2, 3

[48] G. He, S. Long, L. Xu, C. Wu, J. Zhou, M. Sun, X. Wen, and Y. Dai, “Global priors guided modulation network for joint super-resolution and inverse tone-mapping,” 2022. 2, 3

[49] K. Xu, L. Xu, G. He, C. Wu, Z. Ma, M. Sun, and Y.-W. Tai, “Sdrtv-to-hdrtv conversion via spatial-temporal feature fusion,” 2022. 2, 3

[50] T. Shao, D. Zhai, J. Jiang, and X. Liu, “Hybrid conditional deep inverse tone mapping,” in *Proc. ACMMM*, pp. 1016–1024, 2022. 2, 3

[51] A. Mustafa, P. Hanji, and R. K. Mantiuk, “Distilling style from image pairs for global forward and inverse tone mapping,” in *The 19th ACM SIGGRAPH European Conference on Visual Media Production (CVMP)*, pp. 1–10, 2022. 2, 3, 4

[52] M. Yao, D. He, X. Li, Z. Pan, and Z. Xiong, “Bidirectional translation between uhd-hdr and hd-sdr videos,” *IEEE Transactions on Multimedia*, 2023. 2, 3

[53] R. Tang, F. Meng, and L. Bai, “Zoned mapping network from sdr video to hdr video,” in *2022 IEEE 5th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*, vol. 5, pp. 1466–1472, 2022. 2, 3

[54] SMPTE, NY, USA, *ST 2084:2014 - SMPTE Standard - High Dynamic Range Electro-Optical Transfer Function of Mastering Reference Displays*, 2014. 2

[55] ITU, Geneva, Switzerland, *Recommendation ITU-R BT.709-6: Parameter values for the HDTV standards for production and international programme exchange*, 6 ed., 6 2015. 2, 5

[56] Z. Cheng, T. Wang, Y. Li, F. Song, C. Chen, and Z. Xiong, “Towards real-world hdrtv reconstruction: A data synthesis-based approach,” in *Proc. ECCV*, pp. 199–216, 2022. 3

[57] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda, “Photographic tone reproduction for digital images,” in *Proc. 29th SIGGRAPH*, pp. 267–276, 2002. 3

[58] ITU, Geneva, Switzerland, *Report ITU-R BT.2446-1: Methods for conversion of high dynamic range content to standard dynamic range content and vice-versa*, 1 ed., 3 2021. 3, 5

[59] P. Mohammadi, M. T. Pourazad, and P. Nasiopoulos, “A perception-based inverse tone mapping operator for high dynamic range video applications,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 5, pp. 1711–1723, 2020. 3, 8

[60] D. Marnerides, T. Bashford-Rogers, and K. Debattista, “Deep hdr hallucination for inverse tone mapping,” *Sensors*, vol. 21, no. 12, p. 4032, 2021. 3

[61] Y. Zhang and T. Aydın, “Deep hdr estimation with generative detail reconstruction,” in *Computer Graphics Forum*, vol. 40, pp. 179–190, 2021. 3

[62] L. Bai, Y. Yang, and G. Fu, “Analysis of high dynamic range and wide color gamut of uhdtv,” in *2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, vol. 5, pp. 1750–1755, 2021. 4

[63] ITU, Geneva, Switzerland, *Report ITU-R BT.2407-0: Colour gamut conversion from Recommendation ITU-R BT.2020 to Recommendation ITU-R BT.709*, 0 ed., 10 2017. 4, 6

[64] ITU, Geneva, Switzerland, *Recommendation ITU-R BT.500-14: Methodologies for the subjective assessment of the quality of television images*, 14 ed., 10 2019. 4

[65] D. Hasler and S. E. Suesstrunk, "Measuring colorfulness in natural images," in *Human vision and electronic imaging VIII*, vol. 5007, pp. 87–95, 2003. 4

[66] Dolby, "What is ictcp? - introduction." https://professional.dolby.com/siteassets/pdfs/ictcp_dolbywhitepaper_v071.pdf. 4

[67] S. Ploumis, R. Boitard, and P. Nasiopoulos, "Image brightness quantification for hdr," in *2020 28th European Signal Processing Conference (EUSIPCO)*, pp. 640–644, 2021. 4

[68] Arri, "Camera sample footage & reference image." https://www.arri.com/en/learn-help/learn-help-camera-system/camera-sample-footage-reference-image, 2022. 4

[69] Netflix, "Netflix open content." https://opencontent.netflix.com/, 2020. 4

[70] J. Froehlich, S. Grandinetti, B. Eberhardt, S. Walter, A. Schilling, and H. Brendel, "Creating cinematic wide gamut hdr-video for the evaluation of tone mapping operators and hdr-displays," in *Digital photography X*, vol. 9023, pp. 279–288, 2014. 4

[71] Y. Wang, S. Inguva, and B. Adsumilli, "Youtube ugc dataset for video compression research," in *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1–5, 2019. 4

[72] Z. Shang, J. P. Ebenezer, A. C. Bovik, Y. Wu, H. Wei, and S. Sethuraman, "Subjective assessment of high dynamic range videos under different ambient conditions," in *Proc. ICIP*, pp. 786–790, 2022. 4

[73] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008. 5

[74] D. Walker, C. Payne, P. Hodoul, and M. Dolan, "Color management with opencolorio v2," in *ACM SIGGRAPH 2021 Courses*, pp. 1–226, 2021. 5

[75] J. Morovič, *Color gamut mapping*. John Wiley & Sons, 2008. 6

[76] H. Le, T. Jeong, A. Abdelhamed, H. J. Shin, and M. S. Brown, "Gamutnet: Restoring wide-gamut colors for camera-captured images," in *Color and Imaging Conference*, vol. 2021, pp. 7–12, 2021. 6

[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

[78] ITU, Geneva, Switzerland, *Report ITU-R BT.2245-10: HDTV and UHDTV including HDR-TV test materials for assessment of picture quality*, 10 ed., 9 2022. 6

[79] ITU, Geneva, Switzerland, *Recommendation ITU-R BT.2124-0: Objective metric for the assessment of the potential visibility of colour differences in television*, 0 ed., 1 2019. 7

[80] K. Wolski, D. Giunchi, *et al.*, "Dataset and metrics for predicting local visible differences," *ACM Trans. Graph.*, vol. 37, no. 5, pp. 1–14, 2018. 7

[81] C. Bist, R. Cozot, G. Madec, and X. Ducloux, "Tone expansion using lighting style aesthetics," *Computers & Graphics*, vol. 62, pp. 77–86, 2017. 8

[82] G. Luzardo, J. Aelterman, H. Luong, S. Rousseaux, D. Ochoa, and W. Philips, "Fully-automatic inverse tone mapping algorithm based on dynamic mid-level tone mapping," *APSIPA Transactions on Signal and Information Processing*, vol. 9, p. e7, 2020. 8

[83] A. Stojkovic, J. Aelterman, H. Luong, H. Van Parys, and W. Philips, "Highlights analysis system (hans) for low dynamic range to high dynamic range conversion of cinematic low dynamic range content," *IEEE Access*, vol. 9, pp. 43938–43969, 2021. 8