# A SINGLE CONVOLUTIONAL NEURAL NETWORK FOR JOINT SUPER-RESOLUTION, GAMUT EXTENSION, AND INVERSE TONE-MAPPING

Wenyao Gan, Hensheng Zhang, Li Chen, Rong Xie, Li Song✉

Institute of Image Communication and Network Engineering
Shanghai Jiao Tong University, China

## ABSTRACT

With rapid developments of display technology in recent years, Ultra-high definition (UHD) high dynamic range (HDR) displays have emerged in consumer markets. However, due to the lack of UHD HDR video contents, it is necessary to convert legacy high definition (HD) videos with standard dynamic range (SDR) to their UHD HDR versions. In this paper, we first introduce a workflow to down-convert existing UHD HDR videos to their HD SDR versions and then propose a joint super-resolution, gamut extension, and inverse tone-mapping network (JSGIN), which directly learns the up-conversion from the HD SDR videos to their UHD HDR versions. Our JSGIN can enhance visual experience by reconstructing lost information and achieves better subjective visual quality with fewer artifacts than recent state-of-the-art methods.

## INTRODUCTION

Display technology has developed fast in recent years, Ultra-high definition (UHD) higher dynamic range (HDR) displays have become available for consumers. Nevertheless, because of the shortage of UHD HDR video contents, it is required to up-convert legacy high definition (HD) standard dynamic range (SDR) videos to UHD HDR videos. Compared with the current HD SDR television systems ‘(1)’, UHD television systems ‘(2)’ provide higher spatial resolution and wider color gamut, and HDR television systems ‘(3)’ provide a higher dynamic range.

Super-resolution (SR) methods up-scale low-resolution images to high-resolution images. Recent convolutional neural network (CNN) based methods have achieved considerable improvements over conventional SR methods. SRCNN ‘Dong et al (4)’ was the first CNN-based SR method. Then, the CNN architecture was improved by various methods such as sub-pixel convolution ‘Shi et al (5)’ and modified residual blocks ‘Lim et al (6)’.

Gamut extension (GE) algorithms extend colors from a source gamut to a wider destination gamut. Linear color space conversion cannot restore color information outside the source gamut. Conventional GE algorithms attempt to make full use of the wider destination gamut. Recently, ‘TAKEUCHI et al (7)’ proposed a CNN-based GE algorithm that achieves significant gains against conventional GE algorithms.

Inverse tone-mapping (ITM) methods expand SDR images to HDR images. Compared with conventional ITM methods that only focus on mapping the dynamic range, CNN-based ITM methods can restore the lost details in highlights and shadows. ‘Eilertsen et al (8)’ introduced a deep learning system to reconstruct an HDR image from a single exposed SDR image.

UHD HDR videos can be reconstructed from HD SDR videos by cascading SR, GE, and ITM methods. However, the errors from the previous conversion may accumulate, which leads to less accurate results and more overall complexity compared with the joint learning of SR, GE, and ITM. A multi-purpose CNN structure ‘Kim and Kim (9)’ was first proposed to perform the joint learning task of SR, GE, and ITM to directly up-convert HD SDR videos to UHD HDR videos. Then, Deep SR-ITM ‘Kim et al (10)’ was proposed to achieve better results than ‘(9)’ by introducing input decomposition methods and modulation blocks.

ResNet ‘He et al (11)’ introduces local residual learning to ease the difficulty of training of deep CNNs. Global residual learning in SR was first adopted by VDSR ‘Kim et al (12)’ to facilitate training convergence for a deep CNN. Both local residual learning and global residual learning are adopted in our method.

In this paper, we first introduce a workflow to down-convert the existing UHD HDR videos to their HD SDR versions. Then, we propose a single CNN to jointly learn SR, GE, and ITM, which can directly up-convert HD SDR videos to their UHD HDR versions. Compared to recent state-of-the-art methods ‘(9) (10)’, UHD HDR videos generated by our method provide a better visual experience.

## METHODOLOGY

To train our network, both UHD HDR videos and their HD SDR versions are required. In our paper, UHD HDR videos collected by ‘(10)’ are used as ground truth. Their resolution is 4K ($3840 \times 2160$), bit depth is 10, and opto-electronic transfer function (OETF) is Perceptual Quantization (PQ). Different from ‘(10)’ where the automatic conversion process of YouTube is used to convert HDR videos to their SDR versions, we introduce a workflow to down-convert the UHD HDR videos to their HD SDR versions.

### Down-conversion For Creating Our Dataset

Figure 1 shows the workflow of down-conversion from UHD HDR videos to their HD SDR versions. In the 1st step, digitally represented luminance and color-difference signals $[D'_{Y,2020}, D'_{CB,2020}, D'_{CR,2020}]$ in the bit-depth of 10 bits are inverse-quantized to normalized luminance and color-difference signals $[E'_{Y,2020}, E'_{CB,2020}, E'_{CR,2020}]$ according to Recommendation ITU-R BT.2020 ‘(2)’ as follows:

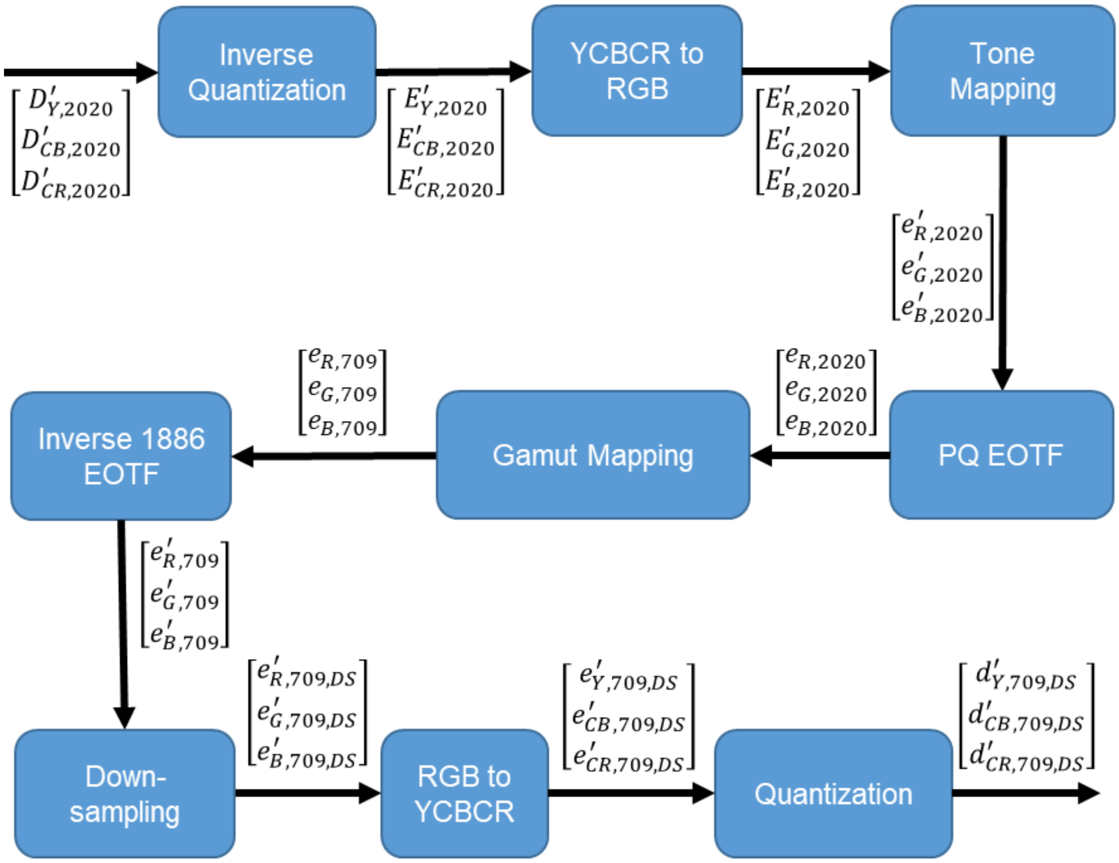
$$E'_{Y,2020} = (D'_{Y,2020}/4 - 16)/219,$$

$$E'_{CB,2020} = (D'_{CB,2020}/4 - 128)/224,$$

$$E'_{CR,2020} = (D'_{CR,2020}/4 - 128)/224.$$

In the 2nd step, luminance and color-difference signals $[E'_{Y,2020}, E'_{CB,2020}, E'_{CR,2020}]$ are converted to RGB color signals $[E'_{R,2020}, E'_{G,2020}, E'_{B,2020}]$ according to Recommendation ITU-R BT.2020 ‘(2)’ as follows:

$$\begin{bmatrix} E'_{R,2020} \\ E'_{G,2020} \\ E'_{B,2020} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1.4746 \\ 1 & -0.1646 & -0.5714 \\ 1 & 1.8814 & 0 \end{bmatrix} \begin{bmatrix} E'_{Y,2020} \\ E'_{CB,2020} \\ E'_{CR,2020} \end{bmatrix}.$$



$[D'_{Y,2020}, D'_{CB,2020}, D'_{CR,2020}]$
Inverse Quantization
$[E'_{Y,2020}, E'_{CB,2020}, E'_{CR,2020}]$
YCBCR to RGB
$[E'_{R,2020}, E'_{G,2020}, E'_{B,2020}]$
Tone Mapping
$[e'_{R,2020}, e'_{G,2020}, e'_{B,2020}]$
PQ EOTF
$[e_{R,2020}, e_{G,2020}, e_{B,2020}]$
Gamut Mapping
$[e_{R,709}, e_{G,709}, e_{B,709}]$
Inverse 1886 EOTF
$[e'_{R,709}, e'_{G,709}, e'_{B,709}]$
Down-sampling
$[e'_{R,709,DS}, e'_{G,709,DS}, e'_{B,709,DS}]$
RGB to YCBCR
$[e'_{Y,709,DS}, e'_{CB,709,DS}, e'_{CR,709,DS}]$
Quantization
$[d'_{Y,709,DS}, d'_{CB,709,DS}, d'_{CR,709,DS}]$

Figure 1 - Flow chart of down-conversion.

In the 3rd step, we tone map the HDR RGB color signals $[E'_{R,2020}, E'_{G,2020}, E'_{B,2020}]$ to SDR RGB color signals $[e'_{R,2020}, e'_{G,2020}, e'_{B,2020}]$ by electrical-electrical transfer function ($EETF$) specified in Recommendation ITU-R BT.2390 '(13)' as follows:

$$e'_{R,2020} = EETF\left(E'_{R,2020}\right),$$

$$e'_{G,2020} = EETF\left(E'_{G,2020}\right),$$

$$e'_{B,2020} = EETF\left(E'_{B,2020}\right).$$

In the 4th step, non-linearly represented RGB color signals $[e'_{R,2020}, e'_{G,2020}, e'_{B,2020}]$ are converted to linearly represented RGB color signals $[e_{R,2020}, e_{G,2020}, e_{B,2020}]$ by PQ electro-optical transfer function ($EOTF_{PQ}$) specified in Recommendation ITU-R BT.2100 '(3)' as follows:

$$e_{R,2020} = EOTF_{PQ}\left(e'_{R,2020}\right),$$

$$e_{G,2020} = EOTF_{PQ}\left(e'_{G,2020}\right),$$

$$e_{B,2020} = EOTF_{PQ}\left(e'_{B,2020}\right).$$

In the 5th step, BT.2020 RGB color signals [$e_{R,2020}$, $e_{G,2020}$, $e_{B,2020}$] are converted to BT.709 RGB color signals [$e_{R,709}$, $e_{G,709}$, $e_{B,709}$] according to Recommendation ITU-R BT.709 ‘(1)’ and Recommendation ITU-R BT.2020 ‘(2)’ as follows:

$$\begin{bmatrix} e_{R,709} \\ e_{G,709} \\ e_{B,709} \end{bmatrix} = \begin{bmatrix} 3.2405 & -1.5371 & -0.4985 \\ -0.9693 & 1.8760 & 0.0416 \\ 0.0556 & -0.2040 & 1.0572 \end{bmatrix} \begin{bmatrix} 0.6370 & 0.1446 & 0.1689 \\ 0.2627 & 0.6780 & 0.0593 \\ 0 & 0.0281 & 1.0610 \end{bmatrix} \begin{bmatrix} e_{R,2020} \\ e_{G,2020} \\ e_{B,2020} \end{bmatrix}.$$

In the 6th step, linearly represented RGB color signals [$e_{R,709}$, $e_{G,709}$, $e_{B,709}$] are converted to non-linearly represented RGB color signals [$e'_{R,709}$, $e'_{G,709}$, $e'_{B,709}$] by the inverse of electro-optical transfer function ($EOTF_{1886}^{-1}$) specified in Recommendation ITU-R BT.1886 ‘(14)’ as follows:

$$e'_{R,709} = EOTF_{1886}^{-1}\left(e_{R,709}\right),$$

$$e'_{G,709} = EOTF_{1886}^{-1}\left(e_{G,709}\right),$$

$$e'_{B,709} = EOTF_{1886}^{-1}\left(e_{B,709}\right).$$

In the 7th step, the image is bicubic down-sampled by a factor of 0.5. The resolution of the down-sampled image is $1920 \times 1080$, which is compliant with Recommendation ITU-R BT.709 ‘(1)’. [$e'_{R,709,DS}$, $e'_{G,709,DS}$, $e'_{B,709,DS}$] represent RGB color signals after down-sampling.

In the 8th step, RGB color signals [$e'_{R,709,DS}$, $e'_{G,709,DS}$, $e'_{B,709,DS}$] are converted to luminance and color-difference signals [$e'_{Y,709,DS}$, $e'_{CB,709,DS}$, $e'_{CR,709,DS}$] according to Recommendation ITU-R BT.709 ‘(1)’ as follows:

$$\begin{bmatrix} e'_{Y,709,DS} \\ e'_{CB,709,DS} \\ e'_{CR,709,DS} \end{bmatrix} = \begin{bmatrix} 0.2126 & -0.1146 & 0.5 \\ 0.7152 & -0.3854 & -0.4542 \\ 0.0722 & 0.5 & -0.0458 \end{bmatrix} \begin{bmatrix} e'_{R,709,DS} \\ e'_{G,709,DS} \\ e'_{B,709,DS} \end{bmatrix}.$$

In the 9th step, normalized luminance and color-difference signals [$e'_{Y,709,DS}$, $e'_{CB,709,DS}$, $e'_{CR,709,DS}$] are quantized to digitally represented luminance and color-difference signals [$d'_{Y,709,DS}$, $d'_{CB,709,DS}$, $d'_{CR,709,DS}$] in the bit-depth of 8 bits according to Recommendation ITU-R BT.709 ‘(1)’ as follows:

$$d'_{Y,709,DS} = round\left(219 \times e'_{Y,709,DS} + 16\right),$$

$$d'_{CB,709,DS} = round\left(219 \times e'_{CB,709,DS} + 128\right),$$

$$d'_{CR,709,DS} = round\left(219 \times e'_{CR,709,DS} + 128\right),$$

where the $round$ operator returns 0 for fractional parts below 0.5 and 1 for fractional parts above or equal to 0.5.

## Up-conversion

As shown in Figure 2, a joint SR, GE, and ITM network (JSGIN) is proposed to directly learn the up-conversion from HD SDR videos to their UHD HDR versions. Our JSGIN is composed of five parts: shallow feature extraction, deep feature extraction, up-scaling, global skip connection, and reconstruction.

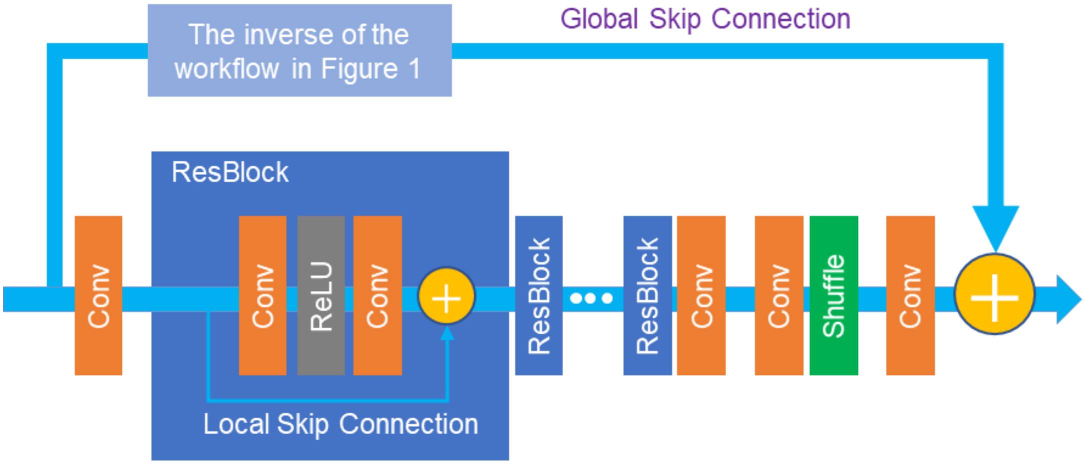


Figure 2 - Network architecture of our joint SR, GE, and ITM network (JSGIN).

Low-level features are extracted by the first convolutional layer (Conv) and then high-level features are extracted by 16 modified residual blocks (ResBlocks) ‘(6)’ followed by one convolutional layer. Next, the sub-pixel structure ‘(5)’ which consists of one convolutional layer and one shuffle layer is adopted for up-scaling. Finally, the input HD SDR frame is converted to the global skip by the inverse of the workflow in Figure 1 and the global skip image is added to the output of the last convolutional layer to reconstruct the final output UHD HDR frame. Compared with the global skip, the final output image restores the lost information of high frequency, colors, and contrast. Both local residual learning and global residual learning are adopted to ease the difficulty of training of deep CNNs.

The training and testing datasets of JSGIN require both UHD HDR videos and their HD SDR versions. In our paper,10 different UHD HDR videos collected by ‘(10)’ are used as ground truth. Different from the automatic down-conversion process of YouTube used in ‘(10)’, we use the workflow in Figure 1 to down-convert the UHD HDR videos to their HD SDR versions. The 10 UHD HDR videos consisting of 59K frames are downloaded from YouTube. The durations of the videos vary from 47 seconds to 197 seconds. Following ‘(10)’, 7 UHD HDR videos consisting of 44K frames are used for training and $160 \times 160$ crops are randomly sampled from a video frame at the interval of about 45 frames. 28 different scenes selected from the remaining 3 UHD HDR videos are used for testing. Therefore, the UHD HDR videos used in the training set and the testing set are different.

We pre-process all the input and output video frames by converting digitally represented luminance and color-difference signals to normalized RGB color signals. Our JSGIN is trained by MSE loss function:

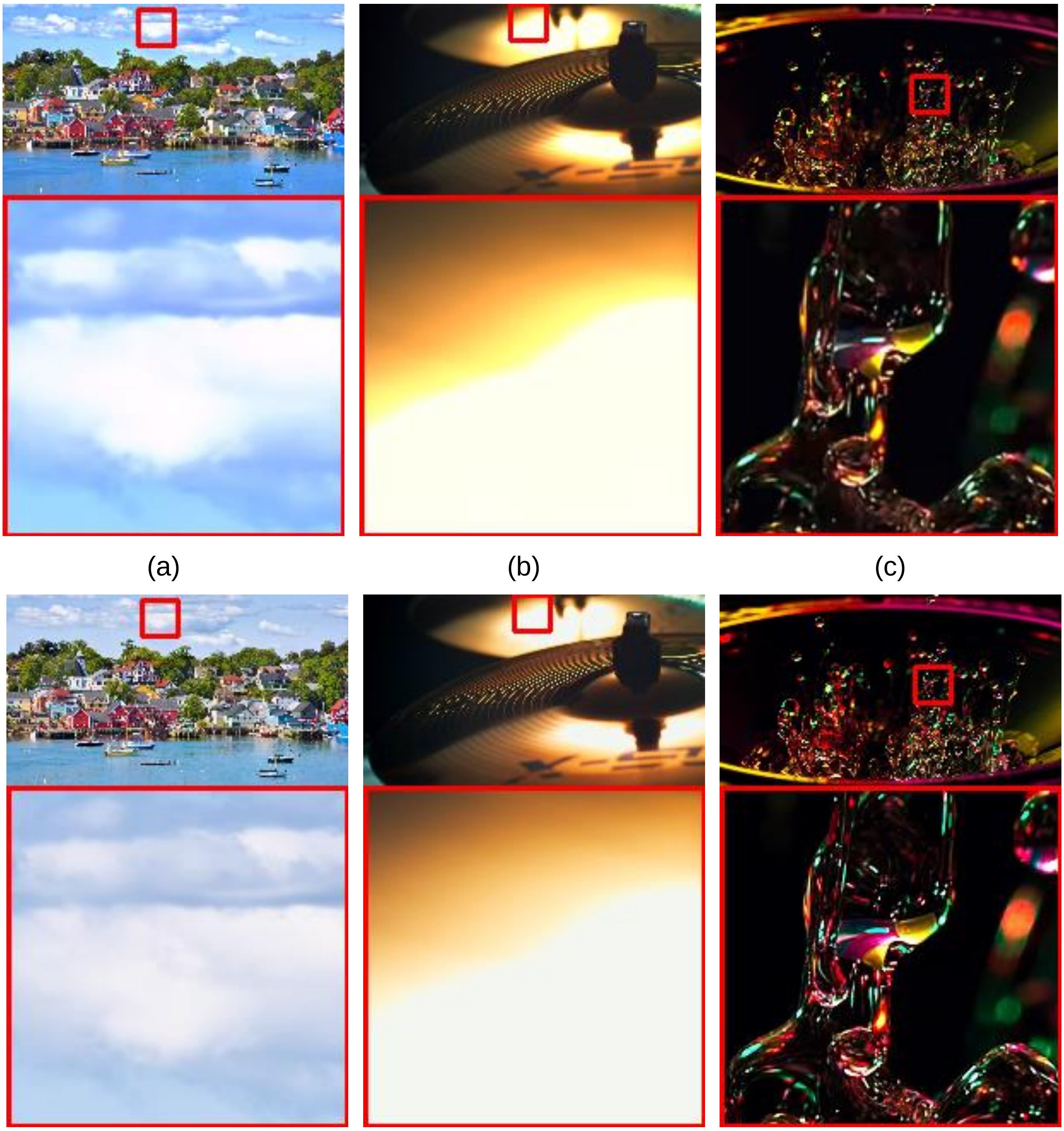
$$Loss(\theta) = \frac{1}{n}\sum_{i=1}^{n} \|f(x_i;\theta) - y_i\|^2,$$

where n represents the number of training samples, $f$ represents the end-to-end mapping function of JSGIN, $x_i$ represents the i-th HD SDR input frame, $\theta$ represent network parameters, and $y_i$ represents the i-th UHD HDR ground truth frame. We train our JSGIN using Adam optimizer ‘Kingma and Ba (15)’ with a mini-batch size of 16 for 320 epochs. Weights are randomly initialized as in ‘He et al (16)’. The initial learning rate is $10^{-6}$ and is divided by 10 at the 200th epoch and the 300th epoch.

## EXPERIMENTS

All experiments are conducted on an NVIDIA GeForce GTX 1080Ti.

### Comparison of Down-conversion



(a) (b) (c)

(d) (e) (f)

Figure 3 - Qualitative comparison of down-conversion between the automatic down-conversion process of YouTube and our workflow. 3 UHD HDR video frames are down-converted to their HD SDR versions. (a), (b), and (c) are generated by YouTube. (d), (e), and (f) are generated by our workflow.

The qualitative comparison of down-conversion between the automatic down-conversion process of YouTube and our workflow is shown in Figure 3. In Figure 3 (a), clouds seem overexposed and lose details. Figure 3 (b) exhibits color banding artifacts. The colors in Figure 3 (c) are washed out especially for red and green. In Figure 3 (d), (e), and (f), the video frames down-converted by our workflow preserve more information about colors and contrast.

### Comparison of Up-conversion

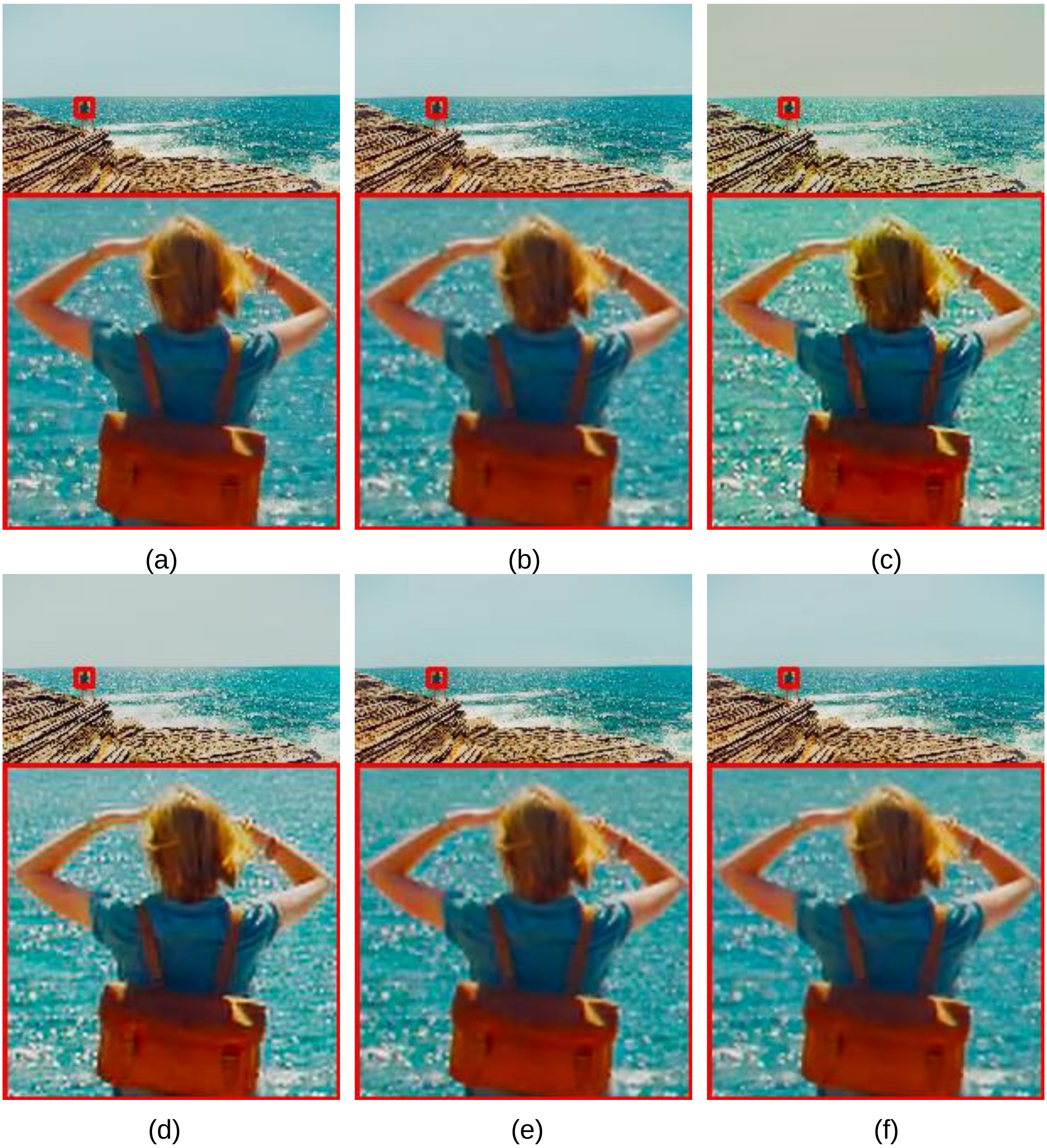
The quantitative performance is compared on 2 metrics: PSNR and SSIM 'Wang et al (17)'. Normalized RGB color signals are compared for PSNR and SSIM. Because Deep SR-ITM outperforms Multi-purpose CNN on both metrics '(10)', we only compare our JSGIN to Deep SR-ITM. As shown in Table 1, since original Deep SR-ITM is trained on a different dataset, it has poor quantitative performance on our testing set. For a fair comparison, we use the source code provided by '(10)' to retrain Deep SR-ITM on our dataset. Our JSGIN outperforms Deep SR-ITM retrained on our dataset on both metrics with fewer network parameters. For JSGIN, removing global skip connection leads to performance degradation.

| Method | Parameters | PSNR (dB) | SSIM |
|---|---|---|---|
| The inverse of the workflow in Figure 1 | 0 | 30.63 | 0.9351 |
| Original Deep SR-ITM | $2.50 \times 10^6$ | 26.97 | 0.8745 |
| Deep SR-ITM retrained on our dataset | $2.50 \times 10^6$ | 31.65 | 0.9433 |
| JSGIN | $1.37 \times 10^6$ | 32.16 | 0.9473 |
| JSGIN without global skip connection | $1.37 \times 10^6$ | 32.04 | 0.9461 |

Table 1 - Quantitative comparison of up-conversion.

The qualitative comparison of up-conversion is shown in Figure 4 and Figure 5. Following Deep SR-ITM, the MPC-HC player and the madVR are used to obtain the visualization of UHD HDR video frames. Original Multi-purpose CNN, original Deep SR-ITM, Deep SR-ITM retrained on our dataset, and JSGIN are trained with the same UHD HDR ground truth frames but different HD SDR input frames. Original Multi-purpose CNN and original Deep SR-ITM use the automatic process of YouTube to down-convert the UHD HDR frames to corresponding HD SDR input frames. In contrast, Deep SR-ITM retrained on our dataset and JSGIN uses our down-conversion workflow in Figure 1.

For the input frame from our testing set, the edges of arms predicted by our JSGIN in Figure 4 (f) is smoother than those in Figure 4 (b), (c), (d), and (e). For the input frame from the testing set of Multi-purpose CNN and Deep SR-ITM, the skies predicted by Multi-purpose CNN and Deep SR-ITM exhibit color banding artifacts in Figure 5 (c) and (d) respectively. In comparison, the UHD HDR frames predicted by Deep SR-ITM retrained on our dataset and our JSGIN enhance the visual experience without noticeable artifacts in Figure 5 (e) and (f) respectively, which indicates that networks trained on our dataset cause fewer artifacts.



(a) (b) (c)

(d) (e) (f)

Figure 4 - Qualitative comparison of up-conversion for the input frame from our testing set. (a) is the ground truth frame. (b) is predicted by the inverse of the workflow in Figure 1. (c) is predicted by original Multi-purpose CNN. (d) is predicted by original Deep SR-ITM. (e) is predicted by Deep SR-ITM retrained on our dataset. (f) is predicted by our JSGIN.

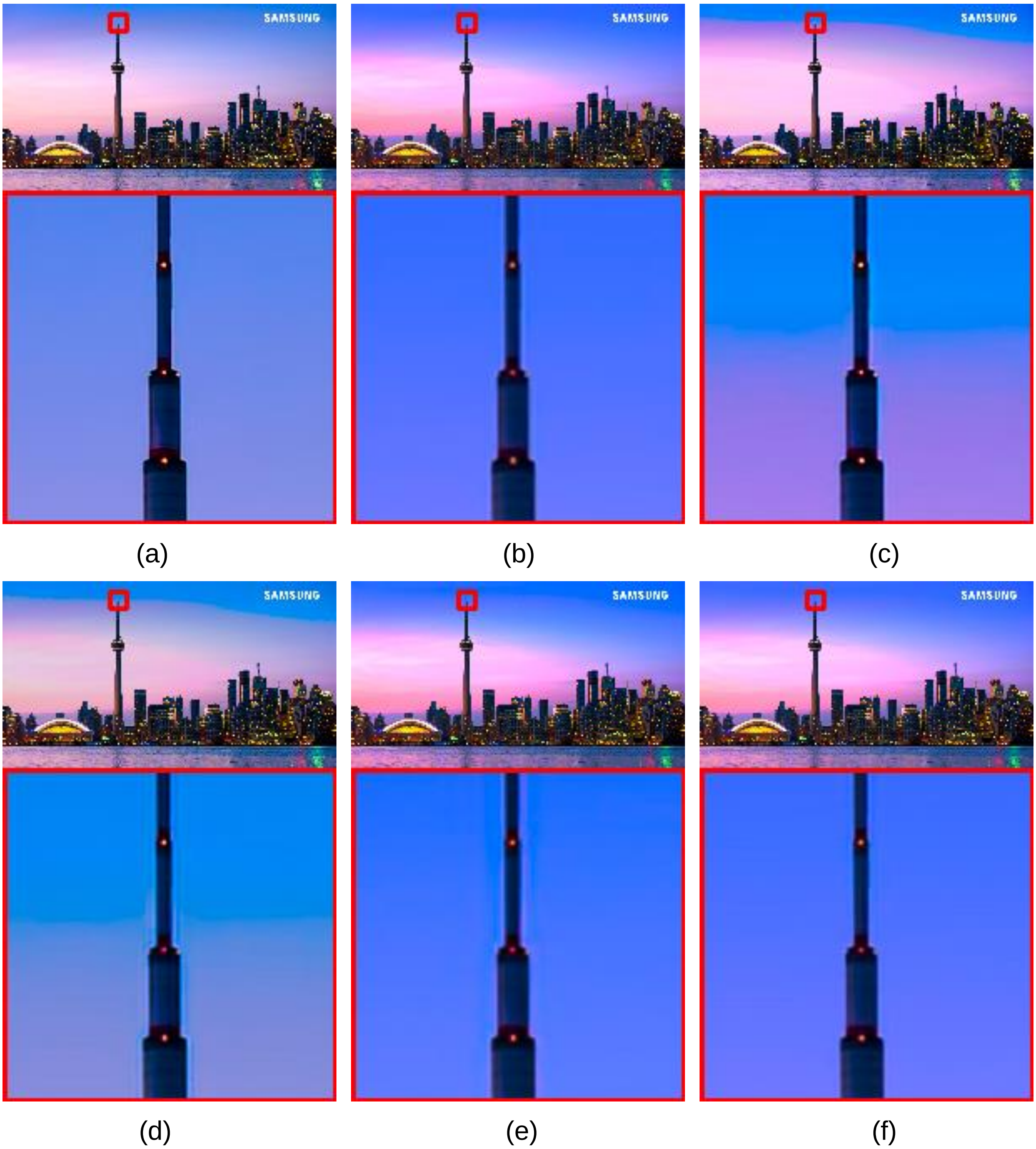


Figure 5 - Qualitative comparison of up-conversion for the input frame from the testing set of Multi-purpose CNN and Deep SR-ITM. (a) is the ground truth frame. (b) is predicted by the inverse of the workflow in Figure 1. (c) is predicted by original Multi-purpose CNN. (d) is predicted by original Deep SR-ITM. (e) is predicted by Deep SR-ITM retrained on our dataset. (f) is predicted by our JSGIN.

## CONCLUSION

In this paper, we first introduce a workflow to down-convert existing UHD HDR videos to their HD SDR versions. Compared with the automatic conversion process of YouTube, the HD SDR videos generated by our method preserve more information about colors and contrast. Then, we propose JSGIN to directly learn the up-conversion from the HD SDR videos to their UHD HDR versions. Both local residual learning and global residual learning are adopted in JSGIN to facilitate training convergence. Our JSGIN achieves better subjective visual quality than recent state-of-the-art methods. In the future, we plan to apply 3D convolution or optical flow to our JSGIN to utilize temporal information. With the help of temporal information, our JSGIN can generate temporally more consistent UHD HDR videos.

## REFERENCES

1. 2015. Parameter values for the HDTV standards for production and international programme exchange. ITU-R Rec. BT.709-6.

2. 2015. Parameter values for ultra-high definition television systems for production and international programme exchange. ITU-R Rec. ITU-R BT.2020-2.

3. 2018. Image parameter values for high dynamic range television for use in production and international programme exchange. ITU-R Rec. BT.2100-2.

4. Dong, C., Loy, C.C., He, K. and Tang, X., 2015. Image super-resolution using deep convolutional networks. IEEE Transactions on Pattern Analysis and Machine Intelligence. pp. 295 to 307.

5. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D. and Wang, Z., 2016. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. IEEE Conference on Computer Vision and Pattern Recognition. pp. 1874 to 1883.

6. Lim, B., Son, S., Kim, H., Nah, S. and Mu Lee, K., 2017. Enhanced deep residual networks for single image super-resolution. IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 136 to 144.

7. TAKEUCHI, M., SAKAMOTO, Y., YOKOYAMA, R., Heming, S.U.N., MATSUO, Y. and KATTO, J., 2019. A gamut-extension method considering color information restoration using convolutional neural networks. IEEE International Conference on Image Processing. pp. 774 to 778.

8. Eilertsen, G., Kronander, J., Denes, G., Mantiuk, R.K. and Unger, J., 2017. HDR image reconstruction from a single exposure using deep CNNs. ACM Transactions on Graphics. pp. 178.

9. Kim, S.Y. and Kim, M., 2018. A multi-purpose convolutional neural network for simultaneous super-resolution and high dynamic range image reconstruction. Asian Conference on Computer Vision. pp. 379 to 394.

10. Kim, S.Y., Oh, J. and Kim, M., 2019. Deep SR-ITM: Joint learning of super-resolution and inverse tone-mapping for 4K UHD HDR applications. IEEE International Conference on Computer Vision.

11. He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. IEEE Conference on Computer Vision and Pattern Recognition. pp. 770 to 778.

12. Kim, J., Kwon Lee, J. and Mu Lee, K., 2016. Accurate image super-resolution using very deep convolutional networks. IEEE Conference on Computer Vision and Pattern Recognition. pp. 1646 to 1654.

13. 2019. High dynamic range television for production and international programme exchange. ITU-R Rec. BT.2390-7.

14. 2011. Reference electro-optical transfer function for flat panel displays used in HDTV studio production. ITU-R Rec. BT.1886-0.

15. Kingma, D.P. and Ba, J., 2014. Adam: A method for stochastic optimization. International Conference on Learning Representations.

16. He, K., Zhang, X., Ren, S. and Sun, J., 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. IEEE International Conference on Computer Vision. pp. 1026 to 1034.

17. Wang, Z., Bovik, A.C., Sheikh, H.R. and Simoncelli, E.P., 2004. Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing. pp. 600 to 612.