A Large-scale Film Style Dataset for Learning Multi-frequency Driven Film Enhancement

Zinuo Li1, Xuhang Chen12, Shuqiang Wang2, Chi-Man Pun1

1University of Macau
2Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
cmpun@umac.mo

Abstract

Film, a classic image style, is culturally significant to the whole photographic industry since it marks the birth of photography. However, film photography is time-consuming and expensive, necessitating a more efficient method for collecting film-style photographs. Numerous datasets that have emerged in the field of image enhancement so far are not film-specific. In order to facilitate film-based image stylization research, we construct FilmSet, a large-scale and high-quality film style dataset. Our dataset includes three different film types and more than 5000 in-the-wild high resolution images. Inspired by the features of FilmSet images, we propose a novel framework called FilmNet based on Laplacian Pyramid for stylizing images across frequency bands and achieving film style outcomes. Experiments reveal that the performance of our model is superior than state-of-the-art techniques. The link of code and data is https://github.com/CXH-Research/FilmNet.

1 Introduction

Film imaging is a special chemical process [Teubner and Brückner, 2019] that generates a unique color and graininess, which is different from that of digital cameras. It is the graininess of film-style images that gives the whole picture a unique charm. The beauty shown in film photographs has demonstrated its charm. Hence, film has become a prominent kind of photography in the minds of many. Years of adjusting by film makers have made it possible to present films in colors that meet the aesthetics of the public. As a result, people have a better sense of the colors presented by film, which has significant implications for the field of enhancing the beauty of images.

Although film style is appealing, film photography is time-consuming, labor-intensive, and expensive. Therefore, many individuals begin to minimize the professionalism of film photography to save time and funds by digitally simulating film styles. Designed to replace tedious human labor, the Look-Up-Table (LUT) is a reliable tool for automatic image color grading. The fundamental premise underlying them is transforming input into a certain output value using efficient lookup and interpolation algorithms. In recent years, deep learning has pushed the development of LUTs, resulting in an explosion of fascinating research [Haoyuan Wang and Lau, 2022; Liu et al., 2022; Wang et al., 2022; Song et al., 2021; Kim et al., 2021; Kim et al., 2020; Yang et al., 2022].

Despite the above-mentioned studies, film stylization has not yet been investigated. A great number of individuals have been drawn to film’s timeless and alluring visual qualities. Nevertheless, many old film cameras lack the ability to export digital images, and digital cameras cannot replicate film imagery so that they must rely on LUT for film simulation.

The challenge is that the current LUTs or some methods focus mostly on global color and lighting, while other de-
tailed operations may get less consideration. Consequently, the visual quality may deteriorate and they may become invalid, as shown in Figure 1 (c). In addition, the existing image enhancement datasets MIT FiveK [Bychkovsky et al., 2011] and HDR Plus [Hasinoff et al., 2016] are designed for broad use and do not adequately address our issue.

Therefore, our objective is to extract additional features and enhance photographs to resemble film. Thus, we construct a large scale film-specific dataset that allows us to facilitate relative film style research, namely the FilmSet. We found that the features of film style images are very suitable for the enhancement in multi-frequency, inspired this, we propose a novel framework utilizing Laplacian Pyramid [Burt and Adelson, 1987]. We summarize our major contributions as follows:

1. We are the first to construct a large-scale high-quality dataset with 3 groups of different film style and a total of 5,285 high-quality images, called FilmSet.
2. To learn the features in FilmSet properly, we present FilmNet, a novel multi-frequency framework based on Laplacian Pyramid for simulating film styles and subsequently retouching normal photos.
3. We demonstrate our model is superior to the state-of-the-art methods via extensive experiments on our dataset and other publicly accessible benchmark datasets.

2 Related Work

2.1 Lookup tables

A LUT is an array that supplants run-time calculation with a more straightforward array indexing process. Once the LUT is created, input images can be retouched using only the memory access and interpolation without further recalculation.

Previous works focus on mastering LUTs to simulate the color adjustment curves of well-known picture editing software [Song et al., 2021; Kim et al., 2021; Guo et al., 2020; Bianco et al., 2020]. By learning a large number of image-independent basic LUTs and combining them with image-dependent weights, it is possible to predict LUTs that are adaptable to a variety of picture contents. Therefore, it is viable to use digital LUT to replicate the film imaging process. However, more refinements are needed to learn the features properly instead of simply using a single LUT, such as refining in different frequency bands or focus both on detailed and global features.

2.2 Photo retouching methods

Previously, experts and professional image editing systems are required for photo retouching to optimize global tuning and adjust local aspects. Nowadays, deep learning models are widely used to retouch photos.

Inspired by bilateral grid processing and local affine color transforms, HDRNet is proposed to use in smartphones [Gharbi et al., 2017]. Then, Hui Zeng et al. proposed to learn 3D LUTs from annotated data using pairwise or unpaired learning [Zeng et al., 2020]. SepLUT [Yang et al., 2022], separated a single color transform into component-independent and component-correlated sub-transforms to enhance images. Liang et al. propose LPTN based on Laplacian Pyramid [Liang et al., 2021]. Nonetheless, few studies have concentrated on film stylization.

2.3 Image enhancement datasets

Recently, there has appeared many datasets that enable learning photo enhancement and retouching. MIT FiveK [Bychkovsky et al., 2011] is a general-purpose dataset which consists of 5,000 original images of broad situations and five versions of retouched targets. Another example is HDR Plus [Hasinoff et al., 2016]. HDR Plus is a burst photography dataset which consists of 3640 bursts (made up of 28461 images in total), organized into sub-folders.

Despite these significant efforts, the preceding datasets were constructed in general scenarios and film-style photographs were not included. Therefore, the models trained on them are inappropriate for the film stylization task. In this study, we produce an expansive FilmSet dataset to help this endeavor.

3 FilmSet Dataset

As previously stated, current datasets and models for photo retouching cannot meet the needs of film stylization. To address these issues, we develop a vast and high-quality dataset, called FilmSet. Visual samples are available in Figure 2. We have three styles in total: Cinema, Classical Negative (Class-Neg) and Velvia, each style contains 5285 images.
3.1 Challenges
To develop a worthwhile film-style dataset that meets real-world needs, we must surmount a number of obstacles. First, the photographs should be in high-quality raw format. In contrast to the ubiquitous and widely accessible JPG photographs, raw photos are far more difficult to get. Second, the dataset should be extensive and encompass a broad variety of real-world scenarios in terms of shooting purpose, diverse settings and portraits, lighting circumstances, and film type, hence increasing the cost of data collecting.

3.2 Data collection and selection
To gather as many raw film-style photographs as feasible, we collected license-free samples from individual photographers and professional photography studios. Additionally, we supplemented some raw format images taken by ourselves.

When gathering data, we carefully inspected the variety of raw images in terms of the shooting occasion, the portrait, background scenes and other possible variants. Figure 2 illustrates the variety of images obtained.

3.3 Film simulation
Initially, we amassed over 8000 raw images, after which we undertook multiple rounds of curation. We initially discarded photos with poor quality, such as significant motion blur or out-of-focus, as well as those containing improper information. In addition, we meticulously examined photos group by group, eliminating outliers and duplicates. We acquired a total of 5285 photographs after the screening.

Using Capture One [One, 2023], we independently applied three film recipes to each of these 5000 images as ground truth. Fujifilm is renowned for its world-class film manufacturing and quality, and Capture One’s film stylization LUT imitates three Fujifilm film styles flawlessly due to their extensive collaboration.

4 Proposed Method: FilmNet
Following the core idea of multi-frequency optimization, initially, our network splits the input picture into two distinct areas using Laplacian Pyramid (LP): two high-frequency regions representing textures and edges, and a colored low-frequency region. The 512 × 512 input image is downsampled to 128 × 128 and sent into the Nonlinear Stylization Remapping (NSR) block to precisely adjust the color details, which substantially increases computing efficiency. The high-frequency sections are input individually into a cascade network, and a mask is learned and expanded, which saves our computation volume and enables us to optimize the high-frequency regions more effectively. Finally, these three images are recombined into a single picture \(I_{out}\), downsampled to a Low-Resolution (LR) image \(I_{L,R}\) and fed into Triple Tri-linear Regulator (TTR) alongside the \(I_{out}\), and the weights of the TTR are adjusted using a CNN to produce the final film output. Each node in our network lightweight algorithm parameters, resulting a more efficient calculation. The overall framework of FilmNet can be seen in Figure 3.

4.1 Multi-frequency Style Transferring
Laplacian Pyramid
Since the characteristics of film style images are very suitable for multi-frequency enhancement and inspired by [Liang et al., 2021], Laplacian Pyramid (LP) [Burt and Adelson, 1987] is applied in the first phase, which is a time-tested image processing method. This allows us to refine images in various frequency domains and obtain high quality results. The LP stores the picture difference between each level’s blurred version and it consists of linearly decomposing a picture into a series of high- and low-frequency bands, from which the original image may be precisely rebuilt.

Given an \(H \times W\) input image \(I\), it creates a low-pass prediction \(\hat{I} \in R^{H/2 \times W/2}\) in which each pixel is a weighted average of its nearby pixels using a specified kernel. In order to provide reversible reconstruction, the LP stores the high-frequency residual \(h_{0}\) as \(h_{0} = I_{0} - \hat{I}_{0}\), where \(I_{0}\) represents the upsamled picture from \(\hat{I}\). To further lower the sample rate and picture resolution, LP repeatedly performs the preceding procedures on \(\hat{I}\) to provide a series of low-frequency and high-frequency components. The whole process can be wrote as Equation 1:

\[
L_{i} = G_{i} - \text{PyrUp}(\text{PyrDown}(G_{i}))
\]

where PyrUp and PyrDown represent the upsample and downsample operations, respectively. \(G_{i}\) is the source image and \(L_{i}\) is the corresponding LP image.

Nonlinear Stylization Remapping
The low-frequency part \(I_{L}\) of the input image is sent to a UNet-like architecture for detailed color transfer, namely Nonlinear Stylization Remapping (NSR), which output refined result \(\hat{I}_{L}\). Inspired by [Chen et al., 2022], NSR is a simplified nonlinear network with lightweight parameters, which eliminates extraneous activation functions such as Sigmoid, Softmax, ReLU, etc. and it has been demonstrated the performance will not drop. The NSR block begins with the addition of LayerNorm inspired by [Ba et al., 2016] to stabilize the training process, followed by two convolutions. After the deformable convolution, SimpleGate and Simplified Channel Attention (SCA) are utilized to improve the performance.

As shown in Figure 3, SimpleGate separates the features straight into two pieces along the channel dimension and multiplies them together and SCA utilizes a direct \(1 \times 1\) convolution technique to transmit data across channels. SimpleGate can be described as Equation 2:

\[
\text{SimpleGate}(X, Y) = X \odot Y
\]

where \(X\) and \(Y\) are identically sized feature maps, \(\odot\) is an element-wise multiplication.

Given a fully-connected layer \(W\), \(pool\) represents the global average pooling procedure that combines spatial data into channels, \(\ast\) indicates a channel-wise multiplication, SCA can be described as Equation 3:

\[
SCA(X) = X \ast W \text{pool}(X)
\]
High-frequency Refinement

In this section, we learn a mask on the concatenation of $[I_{i-1}, up(I_i), up(I_i)]$, as shown in Figure 3, where $up(·)$ indicates bilinear upsampling. This mask is gradually enlarged to improve the remaining high-frequency components based on the inherent property of LP. Given an input high-frequency image $H$ and mask $M$, the output is described as:

$$H_{out} = H \otimes M$$

where $\otimes$ represents the pixel-wise multiplication. This is a simpler method for optimizing global correction compared to mixed-frequency images since high-frequency bands vary only little, allowing us to reduce the calculation volume [Liang et al., 2021].

Then, for better refinement, we introduce Multi-Scale Reconstruction Module (MSRM). There are two components in MSRM: Global-Aware Convolution and Attentive Aggregation Node [Cun et al., 2019]. Inspired by [Xu et al., 2022], Global-Aware Convolution is designed as a two-branch structure with a lightweight convolution [Wu et al., 2019] and a standard convolution, focusing on both light and dark areas of images. The lightweight convolution seeks to learn the color mapping in the lighter area, while conventional convolution seeks to learn the color mapping in the darker part. The outputs of these two branches are then blended and added to the input characteristics using a shortcut in order to increase contextual similarity and decrease learning difficulty.

The Attentive Aggregation Node is designed for feature aggregation and attention aggregation. Each aggregation node utilizes a squeeze-and-excitation block [Hu et al., 2018] to weight the significance of each feature channel. Then, a $3 \times 3$ convolution is used to compress the features and match the original channels. At the end of the MSRM, a spatial pooling pyramid (SPP) [He et al., 2015] is introduced to facilitate the remixing of multi-context features. As shown in Figure 3, the $1/r$ in SPP represents the Average Pooling with $stride = r$.

4.2 Global Refinement

After completing the above procedures, the three images are combined into one and sent into a lightweight TTR module. TTR has been designed to further enhance the tone of film styles. It consists of three 3-dimensional lookup table (3DLUT) weight matrices, which are used to perform trilinear interpolation. Given a source image $I$, we first send it to an Style-Aware Adjuster. The Style-Aware Adjuster is a CNN network and the weight is borrowed from [Zeng et al., 2020]. As seen in Figure 3, the Style-Aware Adjuster modifies the weight of three basis 3D LUTs on the left side based on the input LR image. Given a source image with RGB color $I(x,y,z)_r$, a LUT is performed in order to determine its position $(x, y, z)$ in the 3D LUT lattice as Equation 4:

$$x = \frac{r}{s}, y = \frac{g}{s}, z = \frac{b}{s}$$

where $s = \frac{C_{max}}{M}$, $C_{max}$ refers to the maximum color value and $M$ indicates the number of bins in each color channel.
In the whole training phase, we use $MSE$ and $SSIM$ as loss functions as Equation 5 and 6:

$$L_{MSE} = \frac{1}{n} \sum_{i=1}^{n} (f(x) - y)^2$$ (5)

$$L_{SSIM} = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$ (6)

where $\mu_x$ and $\mu_y$ represent the mean of images $X$ and $Y$, $\sigma_{xy}$ indicates the covariance between images $X$ and $Y$, $\sigma_x$ and $\sigma_y$ represent the standard deviation of images $X$ and $Y$. Normally, $C_1 = (K_1 \times L)^2$ and $C_2 = (K_2 \times L)^2$ with $K_1$, $K_2$ and $L$ set to 0.01, 0.03 and 255.

We set the weight of $SSIM$ function to 0.4, so our total loss function can be wrote as Equation 7:

$$L_{total} = L_{MSE} + 0.4 \cdot L_{SSIM}$$ (7)

5 Experiments

5.1 Experimental Setup

Datasets

In this section, three datasets are used for training and evaluation in total: MIT-Adobe FiveK [Bychkovsky et al., 2011], HDR+ [Hasinoff et al., 2016] and our FilmSet. The MIT-Adobe FiveK dataset is the largest image enhancement dataset available, consisting of five retouched versions of 5,000 original pictures under varied conditions. The 3640-scene HDR+ collection from Google Camera Group for high dynamic range and low-light enhancement is a burst photography dataset. And our FilmSet is a vast high-quality dataset including over 8000 images with three distinct film genres. It is configured with 4657 training samples and 638 testing samples. For easier training and validation, all images are transformed to $512 \times 512$ resolution and standard PNG format. For FiveK and HDR+, we use the same dataset configuration as [Zeng et al., 2020] and transform all images to the
Figure 5: Visual comparison of various image enhancement methods on the FiveK dataset. Our result (g) is aesthetically superior in terms of color tone and specifics. The DPE (b) and UPE (c) results greatly deviate from the objective. Their color, exposure and reproduction of fine details are unsatisfactory. 3D-LUT (e), SepLUT (f) and LPTN (d) are visually better, however the tone mapping is typically too bright or too dark compared to the target (h).

<table>
<thead>
<tr>
<th>Method</th>
<th>Fivek PSNR↑</th>
<th>Fivek SSIM↑</th>
<th>Fivek ΔE↓</th>
<th>HDR+ PSNR↑</th>
<th>HDR+ SSIM↑</th>
<th>HDR+ ΔE↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDRNet</td>
<td>19.93</td>
<td>0.798</td>
<td>14.42</td>
<td>23.04</td>
<td>0.879</td>
<td>8.97</td>
</tr>
<tr>
<td>DPE</td>
<td>17.66</td>
<td>0.725</td>
<td>17.71</td>
<td>22.56</td>
<td>0.872</td>
<td>10.45</td>
</tr>
<tr>
<td>UPE</td>
<td>21.88</td>
<td>0.853</td>
<td>10.80</td>
<td>21.21</td>
<td>0.816</td>
<td>13.05</td>
</tr>
<tr>
<td>DeepLPF</td>
<td>24.55</td>
<td>0.846</td>
<td>8.62</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>3D-LUT</td>
<td>24.59</td>
<td>0.846</td>
<td>8.30</td>
<td>23.54</td>
<td>0.885</td>
<td>7.93</td>
</tr>
<tr>
<td>STAR-DCE</td>
<td>24.50</td>
<td>0.893</td>
<td>N/A</td>
<td>26.50</td>
<td>0.883</td>
<td>5.77</td>
</tr>
<tr>
<td>LPTN</td>
<td>22.19</td>
<td>0.878</td>
<td>11.90</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>SepLUT</td>
<td>25.02</td>
<td>0.873</td>
<td>7.91</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>25.20</strong></td>
<td><strong>0.906</strong></td>
<td><strong>7.62</strong></td>
<td><strong>28.06</strong></td>
<td><strong>0.916</strong></td>
<td><strong>5.41</strong></td>
</tr>
</tbody>
</table>

Table 1: Quantitative comparisons on the MIT FiveK and HDR+ dataset of different image enhancement methods. "N/A" indicates that the result is unavailable and the top result is highlighted in red.

<table>
<thead>
<tr>
<th>Method</th>
<th>Cinema</th>
<th>ClassNeg</th>
<th>Velvia</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PSNR↑</td>
<td>SSIM↑</td>
<td>ΔE↓</td>
</tr>
<tr>
<td>HDRNet</td>
<td>35.18</td>
<td>0.990</td>
<td>2.81</td>
</tr>
<tr>
<td>DPE</td>
<td>3.98</td>
<td>0.358</td>
<td>47.58</td>
</tr>
<tr>
<td>UPE</td>
<td>22.81</td>
<td>0.946</td>
<td>4.22</td>
</tr>
<tr>
<td>DeepLPF</td>
<td>36.34</td>
<td>0.984</td>
<td>1.96</td>
</tr>
<tr>
<td>3D-LUT</td>
<td>35.49</td>
<td>0.990</td>
<td>1.86</td>
</tr>
<tr>
<td>STAR-DCE</td>
<td>28.12</td>
<td>0.949</td>
<td>6.91</td>
</tr>
<tr>
<td>LPTN</td>
<td>36.55</td>
<td>0.985</td>
<td>2.12</td>
</tr>
<tr>
<td>SepLUT</td>
<td>35.82</td>
<td>0.986</td>
<td>2.42</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>40.07</strong></td>
<td><strong>0.993</strong></td>
<td><strong>1.61</strong></td>
</tr>
</tbody>
</table>

Table 2: Quantitative comparisons on the FilmSet dataset of different image enhancement methods. The top result is highlighted in red.

more common 480p resolution and standard PNG format.

**Evaluation Metrics**

In this section, we analyze frameworks utilizing the Peak Signal-to-Noise Ratio (PSNR), Structural Similarity (SSIM),
and $\Delta E$ metrics. $\Delta E$ is a measure of color variation as experienced by humans in the CIELab color space [Backhaus et al., 2011]. Greater PNSR and SSIM values imply increased performance, whereas a lower $\Delta E$ value indicates enhanced color appearance.

### Implementation Details

Our implementation is based on the PyTorch. The typical Adam optimizer with its default parameters is used to train our model by NVIDIA RTX A6000. The batch size is set to 1 and the learning rate is set to $1e-4$. Random cropping, horizontal flipping, and tweaks to brightness and saturation are used to enrich data. Visual results of FilmSet and FiveK are available in Figure 4 and 5.

### 5.2 Comparisons with State-of-the-Arts

A total of eight state-of-the-art methods are selected in this section: HDRNet [Gharbi et al., 2017], DPE [Chen et al., 2018], UPE [Wang et al., 2019], DeepLPF [Moran et al., 2020], 3D-LUT [Zeng et al., 2020], STAR-DCE [Zhang et al., 2021], LPTN [Liang et al., 2021] and SepLUT [Yang et al., 2022]. We utilized SOTA models with their provided pretrained weights in FiveK and HDR+ datasets. For FilmSet, we trained SOTA models by utilizing their own training strategies. As demonstrated in Tables 1 and 2, our method exceeds others across all metrics among the three datasets. Note that DPE produces bad results, which may be because its framework is not conducive to learning the distribution of film style, so we eliminate the visual sample of DPE. Comparing FilmSet to other datasets reveals that all approaches provide excellent results, indicating that our dataset’s distribution is stable and not chaotic like the manually enhanced dataset.

### 5.3 Ablation Studies

In this section, we separate the different parts from our architecture and set $\text{Depth}_{LP}$ to 2 and 3. We do not set $\text{Depth}_{LP}$ to 1 due to the CUDA Memory limitation. In Table 3 and Table 4, “NSR” is Nonlinear Stylization Remapping; “D” represents the depth of LP, i.e., the $\text{Depth}_{LP}$; “A” stands for Attentive Aggregation Node and “TTR” is Triple Trilinear Regulator. The architecture of the best results here is $\text{Depth}_{LP} = 2 + \text{NSR} + \text{Aggregation} + \text{TTR}$.

In the ablation experiment, $\text{Depth}_{LP}$ is increased to 3 and 4. The module’s controlled variable experiment is conducted on D4. It is evident that the results do not improve when the $\text{Depth}_{LP}$ is raised. When the $\text{Depth}_{LP}$ is set to 4, removing the TTR module slightly diminishes the results. Likewise, deleting the A resulted in a small drop as well. When both A and TTR are eliminated, the data show a slighter decline. Following this, we substitute NSR with UNet and see a significant decline in outcomes, indicating that NSR plays a rather significant influence. In summary, increasing the $\text{Depth}_{LP}$ does not mean a better performance and every component in our architecture is helpful for improving the results.

### 6 Conclusion

In this paper, we construct a new dataset FilmSet, a large-scale and high-quality library of film styles. Our dataset consists of three distinct film types and over 5000 photos captured in the field in raw format. In order to learn the features of the FilmSet images more properly, we propose the FilmNet, a new framework based on Laplacian Pyramid for refining multi-frequency pictures and achieving high-quality results. We demonstrate that the performance of our model is superior to the state-of-the-art strategies through extensive experiments. This may facilitate the film style transferring researches in deep learning methods.

---

### Tables

**Table 3**: Ablation studies on the MIT FiveK and HDR+ dataset of different image enhancement methods. The top result is highlighted in red.

<table>
<thead>
<tr>
<th>Method</th>
<th>Fivek PSNR↑</th>
<th>Fivek SSIM↑</th>
<th>Fivek $\Delta E$↑</th>
<th>HDR+ PSNR↑</th>
<th>HDR+ SSIM↑</th>
<th>HDR+ $\Delta E$↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>D2+NSR+A</td>
<td>25.18</td>
<td>0.902</td>
<td>7.67</td>
<td>27.22</td>
<td>0.905</td>
<td>6.02</td>
</tr>
<tr>
<td>D2+NSR+TTR</td>
<td>25.14</td>
<td>0.903</td>
<td>7.63</td>
<td>27.08</td>
<td>0.906</td>
<td>6.24</td>
</tr>
<tr>
<td>D2+NSR</td>
<td>25.15</td>
<td>0.901</td>
<td>7.64</td>
<td>26.95</td>
<td>0.888</td>
<td>6.65</td>
</tr>
<tr>
<td>D2+UNet+A+TTR</td>
<td>21.98</td>
<td>0.856</td>
<td>11.56</td>
<td>21.67</td>
<td>0.842</td>
<td>11.51</td>
</tr>
<tr>
<td>D3+NSR+A+TTR</td>
<td>25.06</td>
<td>0.899</td>
<td>7.68</td>
<td>26.70</td>
<td>0.901</td>
<td>6.78</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>25.20</strong></td>
<td><strong>0.906</strong></td>
<td><strong>7.62</strong></td>
<td><strong>28.06</strong></td>
<td><strong>0.916</strong></td>
<td><strong>5.41</strong></td>
</tr>
</tbody>
</table>

**Table 4**: Ablation studies on the FilmSet dataset of different image enhancement methods. The top result is highlighted in red.

<table>
<thead>
<tr>
<th>Method</th>
<th>Cinema PSNR↑</th>
<th>Cinema SSIM↑</th>
<th>Cinema $\Delta E$↑</th>
<th>ClassNeg PSNR↑</th>
<th>ClassNeg SSIM↑</th>
<th>ClassNeg $\Delta E$↑</th>
<th>Velvia PSNR↑</th>
<th>Velvia SSIM↑</th>
<th>Velvia $\Delta E$↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>D2+NSR+A</td>
<td>39.18</td>
<td>0.992</td>
<td>1.61</td>
<td>37.40</td>
<td>0.991</td>
<td>1.55</td>
<td>37.54</td>
<td>0.976</td>
<td>2.06</td>
</tr>
<tr>
<td>D2+NSR+TTR</td>
<td>39.46</td>
<td>0.992</td>
<td>1.61</td>
<td>38.65</td>
<td>0.992</td>
<td>1.53</td>
<td>37.60</td>
<td>0.978</td>
<td>2.12</td>
</tr>
<tr>
<td>D2+NSR</td>
<td>39.46</td>
<td>0.992</td>
<td>1.91</td>
<td>37.55</td>
<td>0.990</td>
<td>1.71</td>
<td>37.49</td>
<td>0.977</td>
<td>2.61</td>
</tr>
<tr>
<td>D2+UNet+A+TTR</td>
<td>32.10</td>
<td>0.987</td>
<td>3.84</td>
<td>30.14</td>
<td>0.975</td>
<td>4.50</td>
<td>33.59</td>
<td>0.969</td>
<td>3.52</td>
</tr>
<tr>
<td>D3+NSR+A+TTR</td>
<td>39.29</td>
<td>0.992</td>
<td>1.94</td>
<td>37.50</td>
<td>0.991</td>
<td>1.74</td>
<td>37.36</td>
<td>0.980</td>
<td>2.21</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>40.07</strong></td>
<td><strong>0.993</strong></td>
<td><strong>1.61</strong></td>
<td><strong>38.89</strong></td>
<td><strong>0.992</strong></td>
<td><strong>1.47</strong></td>
<td><strong>37.60</strong></td>
<td><strong>0.981</strong></td>
<td><strong>2.05</strong></td>
</tr>
</tbody>
</table>
Ethical Statement
There are no ethical issues.

Acknowledgments
This work was supported in part by the University of Macau under Grant MYRG2022-00190-FST, in part by the Science and Technology Development Fund, Macau SAR, under Grant 0034/2019/AMJ, Grant 0087/2020/A2 and Grant 0049/2021/A, in part by the National Natural Science Foundations of China under Grants 62172403 and in part by the Distinguished Young Scholars Fund of Guangdong under Grant 2021B1515020019.

Contribution Statement
Zinuo Li and Xuhang Chen contributed equally to this work as co-first authors. Zinuo Li and Xuhang Chen made the dataset, designed and performed the experiments, analyzed the data and wrote the manuscript. Shuqiang Wang and Chi-man Pun served as corresponding authors. Shuqiang Wang, the primary corresponding author, conceived and supervised the project, provided overall guidance and critical revisions of the manuscript. Chi-man Pun, the primary corresponding author, provided key suggestions on the experimental design and revised the manuscript. Chi-man Pun, the primary corresponding author, conceived and supervised the project, provided critical feedback and edited the manuscript.

References


