Modernizing Old Photos Using Multiple References via Photorealistic Style Transfer

Agus Gunawan[1] Soo Ye Kim[1,2*] Hyeonjun Sim[1†] Jae-Ho Lee[3] Munchurl Kim[1‡]
[1]KAIST [2]Adobe Research [3]ETRI
{agusgun, flhy5836, mkimee}@kaist.ac.kr sooyek@adobe.com jhlee3@etri.re.kr

Figure 1. The results of old photo modernization produced by our method. Our method is able to generate more modern-looking images that resemble the style of input reference images **without the use of old photos during training**. Please visit our webpage at https://kaist-viclab.github.io/old-photo-modernization for dataset and code.

Abstract

This paper firstly presents old photo modernization using multiple references by performing stylization and enhancement in a unified manner. In order to modernize old photos, we propose a novel multi-reference-based old photo modernization (MROPM) framework consisting of a network MROPM-Net and a novel synthetic data generation scheme. MROPM-Net stylizes old photos using multiple references via photorealistic style transfer (PST) and further enhances the results to produce modern-looking images. Meanwhile, the synthetic data generation scheme trains the network to effectively utilize multiple references to perform modernization. To evaluate the performance, we propose a new old photos benchmark dataset (CHD) consisting of diverse natural indoor and outdoor scenes. Extensive experiments show that the proposed method outperforms other baselines in performing modernization on real old photos, even though no old photos were used during training. Moreover, our method can appropriately select styles from multiple references for each semantic region in the old photo to further improve the modernization performance.

1. Introduction

Old photos taken a long time ago may contain important information that carry cultural and heritage values, e.g., photos of Queen Elizabeth II's coronation. Such old images may contain multiple degradations, e.g., scratches, and old photo artifacts, e.g., color fading, often preventing people from understanding the scene. To restore these images, a skilled expert needs to perform laborious manual processes such as degradation restoration and modernization, i.e., colorization or enhancement, to make them look modern [46]. Consequently, early studies [9, 40] try to restore damaged old photos automatically by using traditional inpainting techniques. However, solely re-synthesizing damaged regions in the image is inadequate to ensure old photos look modern, as the overall style remains similar.

Recent work [29] formulates the task as time-travel rephotography which aims to translate old photos into a modern photos space. The authors considered a multi-task problem consisting of two main tasks: (i) restoration of old photos with both unstructured (noise, blur) and structured (scratch, crack) degradations; (ii) modernization which aims to change old photos' characteristics to look like modern images, e.g., better color saturation and contrast by using colorization [29, 50] or enhancement [44]. However, simply using an enhancement method [44] fails to modern-

*Soo Ye Kim is currently affiliated with Adobe Research.
†Hyeonjun Sim is currently affiliated with Qualcomm.
‡Corresponding author.

ize old photos, as shown in Fig. 1, since the overall look still remains similar to old photos, e.g., with a sepia color.

In this paper, we propose to modernize old color photos of natural scenes by changing their styles and enhancing them to look modern. For this, a novel unified framework is proposed which leverages multiple modern photo references in solving the modernization task of old photos by utilizing photorealistic style transfer (PST). Although one prior work [50] is also reference-based, it only relies on a single reference to colorize greyscale portrait photos. However, in natural scene cases, it is challenging to find a single modern photo as a reference that can well match the whole semantics of an old photo. Moreover, changing only the color is not sufficient to alter the overall look of an image [13]. Thus, our framework uses multiple references to modernize old photos by changing the *style* instead of only the color. Since there is no public old photos benchmark dataset of natural scenes, we propose a new Cultural Heritage Dataset (CHD) with 644 indoor and outdoor old color photos collected from various national museums in Korea.

Our multiple-reference-based old photo modernization framework (MROPM) consists of two main parts: (i) *MROPM-Net* and (ii) *a novel training strategy* that enables the network to utilize multiple references. The MROPM-Net consists of two different subnets: The first is a single stylization subnet that transfers both global and local styles without any semantic segmentation from a modern photo into an old photo; Specifically, we propose an improved version of WCT2 [53], inspired by its universal generalization, as the backbone of the single stylization subnet, and present a new architecture that can perform both global PST and local PST without requiring any semantic segmentation; The second is a merging-refinement subnet that merges multiple stylization results from multiple references based on semantic similarities and further refines the merged result to produce a modernized version of the old photo. To effectively train the MROPM-Net, we propose a synthetic data generation scheme that uses the style-variant (i.e., color jittering and unstructured degradation) and -invariant (i.e., rotation, flipping, and translation) transformations. Our MROPM can modernize old photos better than the state-of-the-art (SOTA) old photo restoration method [44], even without using any old photos during training, thanks to the generalization of PST. Our contributions are summarized as follows:

- We propose the *first* old photo modernization framework (MROPM) that allows the usage of *multiple references* to guide the modernization process.
- Our *photorealistic multi-stylization network* and training strategy enable the MROPM-Net to utilize multiple style references in modernizing old photos.
- Our training strategy based on synthetic data allows the MROPM-Net to modernize *real* old photos even without using any old photos during training.
- We propose a new old photo dataset of natural scenes, called Cultural Heritage Dataset (CHD), with 644 outdoor and indoor cultural heritage images.

2. Related Work

Reference-based color-related tasks. One way to change the overall look of an image is by changing color, which is one of the style components [13]. To change the color of old photos, one can employ two methods: *exemplar-based colorization* [11, 27, 41, 48, 51, 52] and *color transfer* or *recolorization* [1, 12, 21, 25]. However, exemplar-based colorization methods cannot utilize the color information in the input images for matching, although color is an important feature representing object semantics [39], limiting the methods for the modernization of old color photos. Color transfer aims to transfer the reference image's color statistics into the input image. Early deep learning works [12, 25] use deep feature matching from features extracted with pre-trained VGG19 [38] to perform the color transfer, which can also be extended to multi-reference cases [12]. Due to the long execution time of the optimization process, recent works develop end-to-end networks, where Lee *et al*. [21] utilize color histogram analogy, and Afifi *et al*. [1] utilize a color-controlled generative adversarial network (GAN). However, recent works can only use a single reference, where finding a single reference image containing similar semantics as the input old photo can be challenging. Thus, from the perspective of color transfer, our work is the first end-to-end network that can utilize multiple references to handle content mismatch without any slow optimization technique.

Photorealistic style transfer (PST). The PST aims at achieving photorealistic rendering of an image with the style of another image. Since the development of post-processing and regularization techniques [28, 32], PST has gained much popularity. Recent works can be categorized into architecture [2, 7, 8, 24, 36, 49, 53] and feature transformation [16, 22, 23] improvements to effectively and efficiently produce photorealistic results. Specifically, WCT2 [53] utilizes wavelet-based skip connection and progressive stylization to achieve better PST where the method can work universally without re-training to pre-defined styles. Due to these benefits, we base our network architecture on WCT2. However, WCT2 produces unnatural style transfer results when performing global and local stylization with unreliable semantic segmentation (shown in Supplementary Material), which hinders the application to old photos. Thus, our MROPM-Net is designed to enable local stylization without any semantic segmentation, which in consequence, can perform multi-style PST in one unified framework without specifying any masks. To the best of

our knowledge, this is the first work in multi-style PST, although there is one work in multi-style artistic style transfer (AST) [15]. Note that the AST is different from the PST in that it utilizes learning-dependent feature transformation, which can cause severe visual artifacts in PST.

Old photo restoration. Early works in old photo restoration focus on detecting and restoring structured degradation (scratch and crack) of images using traditional inpainting techniques [9, 40]. Besides the structured degradation, [26, 44, 50] incorporate additional spatially-uniform unstructured degradation, e.g., blur and noise, using synthetic degradation and formulate the problem as mixed degradation restoration. However, restoring mixed degradation is not enough to ensure that old photos look modern. Consequently, Luo *et al*. [29] formally introduce the time-travel rephotography problem, which aims to translate old photos to look like ones taken in the modern era. This problem adds modernization, synthesizing the missing colors and enhancing the details, on top of degradation restoration. To solve the modernization problem, Luo *et al*. [29] use a StyleGAN2-generated [19] sibling image to serve as a reference for old portrait photos. However, generating complex natural scene photos via GAN to be used as references is challenging [3], making the method unable to be applied to natural scene old photos. Another work [50] proposes to use a single reference image to colorize an old greyscale photo. However, using a single reference is not enough to cover the whole semantics of old photos (shown in Fig. 1). Thus, different from previous methods, we propose to modernize old photos by stylizing and enhancing old photos in a unified manner using *multiple references* to better cover the entire semantics of old photos.

3. Proposed Cultural Heritage Dataset (CHD)

Although some public datasets such as Historical Wiki Face Dataset (HWFD) [29] and RealOld [50] have been released recently, these datasets only contain portrait or face photos which are much simpler compared to natural scenes. In addition, these datasets only contain greyscale photos and disregard color photos produced during the 20th century using reversal films [30], which have specific degradations such as color dye fading and have not been analyzed before. Therefore, we propose a Cultural Heritage Dataset (CHD) consisting of 644 old color photos produced in the 20th century. Specifically, we collect these old photos in the form of reversal films or papers from three national museums in Korea, which are then scanned in resolutions varying from 4K to 8K. The photos have been well preserved and stored carefully due to their value, containing little structured degradation, e.g., scratches, but varying degrees of unstructured degradation, e.g., noises. These photos contain indoor and outdoor scenes of cultural heritage, such as special exhibitions and excavation ruins. After collection, all photos are divided into train and test sets by randomly splitting with a proportion of 8 (514 photos):2 (130 photos). The train set is only used for other baselines that need to be trained using real old photos. Since our task is reference-based old photo modernization, we further collect modern photos as references by crawling images with similar contexts from the internet. In total, we obtain 130 old photos in the test set, each of which has one or two references selected manually. Further details can be found in *Supplementary Material*.

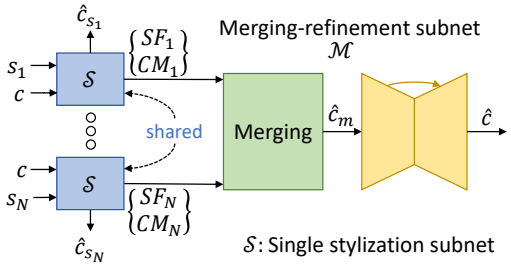

Figure 2. The overall framework of our multiple-reference-based old photo modernization network (MROPM-Net).

4. Proposed Method

4.1. Overall Framework

Fig. 2 shows our proposed multi-reference-based old photo modernization network (MROPM-Net) with a shared single stylization subnet $\mathcal{S}$ and a merging-refinement subnet $\mathcal{M}$. We denote an old photo input as content $c \in \mathbb{R}^{H\times W\times 3}$ and N number of modern photos as styles $\boldsymbol{s} = \{s_i\}_{i=1}^{N} \in \mathbb{R}^{N\times H\times W\times 3}$. Our goal is to modernize c using $\boldsymbol{s}$. In the first step, we utilize $\mathcal{S}$, which is built based on a photorealistic style transfer (PST) backbone, to stylize c using each s_i, yielding N stylized features and correlation matrices $\{SF_i, CM_i\}_{i=1}^{N}$. After having multiple stylization results, we merge the features $\{SF_i\}_{i=1}^{N}$ based on the semantic similarity $\{CM_i\}_{i=1}^{N}$ between c and $\boldsymbol{s}$ and further refine the merging result via $\mathcal{M}$. Specifically, $\mathcal{M}$ selects the appropriate styles for each semantic region based on multiple stylization results $\{SF_i\}_{i=1}^{N}$ to produce an intermediate merging image output $\hat{c}_m$, e.g., selecting the most appropriate feature for a sky region from SF_1 that contains a sky style, not from $SF_{i\neq 1}$, which do not contain sky styles. Then, $\hat{c}_m$ is further refined to get the final result $\hat{c}$. Given relevant references, $\hat{c}$ becomes a modern version with a modern style and enhanced details for old photo input c.

4.2. Network Architecture

Single stylization subnet $\mathcal{S}$. Fig. 3 shows a detailed structure of $\mathcal{S}$. For given multiple references, our single stylization subnet is shared for all input pairs and takes a single pair of an old photo c and a reference s_i at a time. Given a pair of (c, s_i), $\mathcal{S}$ stylizes c based on the style code of s_i locally and globally, resulting in a stylized feature SF_i,

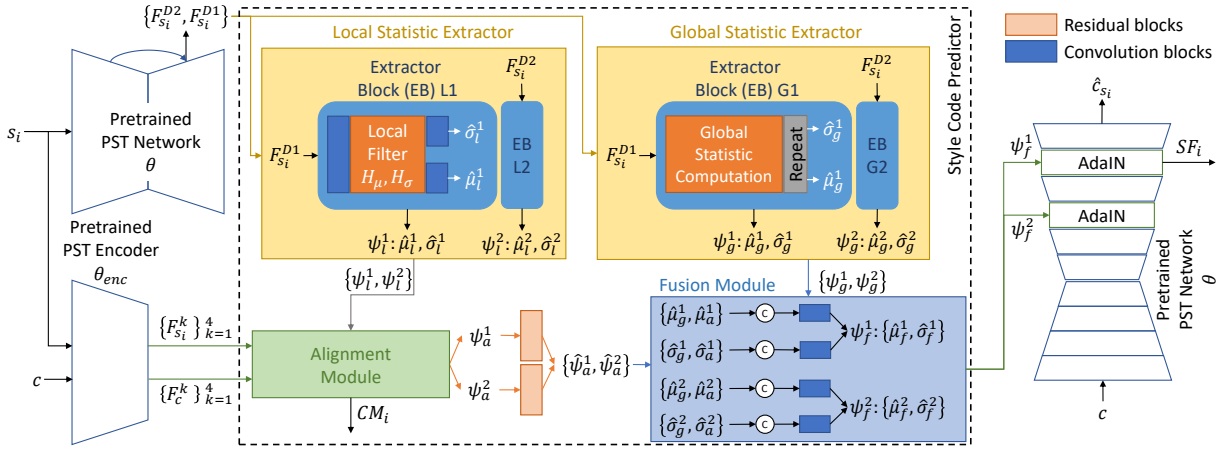

Figure 3. The architecture of the single stylization subnet $\mathcal{S}$.

a stylized old photo $\hat{c}_{s_i}$, and a correlation matrix CM_i. This subnet $\mathcal{S}$ consists of two main parts: (i) *an improved PST network* and (ii) *a style code predictor*.

For the PST network, we improve some drawbacks of the concatenated version of WCT2 [53]. We observed that the stylization only affects the last decoder block due to the design of its skip connection, where this issue is called a "short circuit" in [2]. Thus, instead of transferring three different high-frequency components as in the WCT2, we propose to simplify it by transferring a single high-frequency component in level-0 of the Laplacian pyramid representation [4]. Second, we only apply feature transformation in the network's decoder part, especially the last two decoder blocks, which achieves the best trade-off between the stylization effect and the photorealism. Third, we use the differentiable adaptive instance normalization (AdaIN) [14] instead of the non-differentiable WCT [23] to learn and predict the local style rather than compute it.

The second part of $\mathcal{S}$ is a style code predictor. This part aims to predict style codes $\psi = \{\mu, \sigma\}$ consisting of mean and standard deviation (std), which are statistics used to perform stylization in AdaIN [14]. We propose to predict ψ instead of computing it as in AdaIN to perform local style transfer without requiring any semantic segmentation. The first step (yellow) of the style code predictor is to extract local style codes $\psi_l^j = \{\hat{\mu}_l^j, \hat{\sigma}_l^j\}$ and global style codes $\psi_g^j = \{\hat{\mu}_g^j, \hat{\sigma}_g^j\}$ from the j-th level feature $\{F_{s_i}^{Dj}\}$ extracted by the last two decoder blocks ($j = 1, 2$) of the pre-trained PST network as shown in Fig. 3. In this regard, ψ_l^j is extracted using a local statistic extractor which consists of a local mean filter H_μ and local std filter H_σ with a kernel size of 3 and convolution blocks to refine both filtered outputs. Meanwhile, the global statistic extractor extracts ψ_g^j by computing channel-wise mean and std values, which are then spatially repeated to the same spatial size of ψ_l^j. After style code extraction, the second step (green) is to align ψ_l^j to c by using non-local attention [45]. Specifically, we extract multi-level feature maps $\{F_c^k\}_{k=1}^4$ and $\{F_{s_i}^k\}_{k=1}^4$ for both c and s_i, respectively, map them into the same feature space using shared convolution blocks, and perform matrix multiplication between mapped features to obtain correlation matrix CM_i. Then, we align ψ_l^j to c by using CM_i via matrix multiplication. The aligned style code ψ_a^j is further refined to prevent interpolation artifacts by using residual blocks [10], resulting in a refined version $\hat{\psi}_a^j = \{\hat{\mu}_a^j, \hat{\sigma}_a^j\}$. After obtaining $\hat{\psi}_a^j$, we fuse it with ψ_g^j via the fusion module to obtain a fused style code $\psi_f^j = \{\hat{\mu}_f^j, \hat{\sigma}_f^j\}$. The fusion module performs channel-wise concatenation for $\hat{\psi}_a^j$ and ψ_g^j, which is then fed into the following convolution blocks as shown in the blue part of Fig. 3.

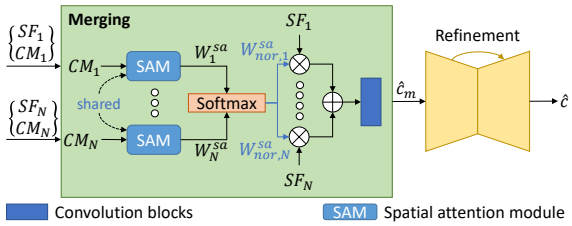

Figure 4. The architecture of the merging-refinement subnet $\mathcal{M}$.

Finally, after performing all the operations from the local and global statistic extractors to the fusion module, we obtain ψ_f^1 and ψ_f^2. These fused style codes are then used for stylizing c. We use our PST network with AdaIN to perform the stylization as shown in the right part of Fig. 3.

Merging-refinement subnet $\mathcal{M}$. After stylizing an old photo c with N different modern photos $\boldsymbol{s} = \{s_i\}_{i=1}^N$ using $\mathcal{S}$, we obtain multiple stylized features and correlation matrices $\{SF_i, CM_i\}_{i=1}^N$. The next step is to select the most appropriate styles from $\{SF_i\}_{i=1}^N$ for each semantic region via the merging part of $\mathcal{M}$, as shown in Fig. 4. For this, a spatial attention module (SAM) [47] is employed, which strengthens and dampens semantically related and unrelated spatial features, respectively, in the merging process of the stylized features. The SAM computes spatial

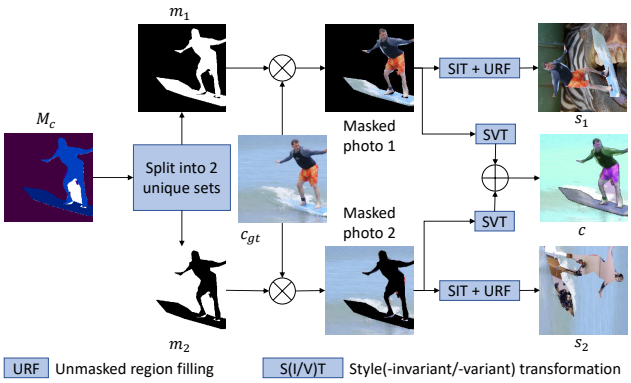

Figure 5. Our synthetic data generation pipeline.

attention weights W_i^{sa} by using CM_i for the corresponding SF_i. Then, we normalize all the spatial attention weights by using Softmax, thus having $\boldsymbol{W}_{nor}^{sa} = \{W_{nor,i}^{sa}\}_{i=1}^{N}$. All normalized attention weights $W_{nor,i}^{sa}$ are multiplied with their corresponding SF_i, whose results are summed and then fed into the final convolution blocks to obtain a merging result $\hat{c}_m$ as an intermediate multi-style PST image. We further refine $\hat{c}_m$ via the U-Net [37] based refinement subnet to produce a final modern version $\hat{c}$ for old photo input c.

4.3. Training Strategy

Synthetic data generation. Since there is no ground truth for multiple-reference-based old photo modernization tasks, we generate synthetic data for training the network in a self-supervised manner. For this, the COCO-stuff dataset [5] is utilized, which has a semantic segmentation mask for each image. Fig. 5 shows the pipeline of generating the synthetic data where each sample consists of a *synthetic* old photo c, its corresponding N different references $\boldsymbol{s} = \{s_i\}_{i=1}^{N}$, and its ground truth c_{gt}. We use two style references with $N = 2$ for each sample throughout our experiments. First, we randomly select a photo from the COCO-stuff dataset as ground truth c_{gt} and its corresponding semantic mask M_c at each iteration during the training. Then, all the semantic regions in M_c are randomly separated into N non-overlapping parts $\{m_i\}_{i=1}^{N}$. For the example shown in Fig. 5, M_c consists of two semantic regions, *surfer* and *sea*, and thus separated to: one mask m_1 with the surfer, and the other mask m_2 with the sea. The next step is to generate N different masked photos using $\{m_i\}_{i=1}^{N}$ by element-wise multiplication between m_i and c_{gt}. Then, we use these masked photos to generate multiple references $\{s_i\}_{i=1}^{N}$ and c via style-invariant (SIT) and -variant transformations (SVT) respectively.

The properties of style-variant and -invariant transformations are determined by whether the transformations alter the mean and std of any semantic region. Hence, we use random translation (only for the regions that can be translated), rotation, and flipping as our SIT. Meanwhile, random color jittering and unstructured degradation, i.e. blur, noise, resizing, and compression artifacts, are used for SVT. Other types of degradation, e.g., scratches can be included in the SVT to make the method able to generalize to these types of degradation. To generate c, we apply different SVTs for each masked photo and sum up the results. Meanwhile, to generate $\{s_i\}_{i=1}^{N}$, randomly selected SITs are applied for each masked photo, and then the unmasked region is filled (URF) with another photo randomly selected from the same COCO-Stuff dataset. Our MROPM-Net can work reasonably well for real old photos after training with this synthetic data. This is because the synthetic data make our MROPM-Net able to (i) robustly find local semantic correspondences between degraded synthetic old photo c and semantically confusing synthetic modern photo s_i, (ii) accurately transfer the styles of each s_i to c locally, and (iii) merge and refine multiple stylization results from multiple styles $\{s_i\}_{i=1}^{N}$ to produce an output similar to c_{gt}. Thus, our synthetic data creation pipeline can be effectively used for multi-reference-based old photo modernization.

Our MROPM-Net is trained in multiple stages: (i) **Stage 1**: Our PST network is trained using a similar training strategy to [53]; (ii) **Stage 2**: Our single stylization subnet $\mathcal{S}$ is trained, while the pre-trained PST network is freezed; (iii) **Stage 3**: We train our merging-refinement subnet $\mathcal{M}$, with both the pre-trained PST network and $\mathcal{S}$ freezed.

Loss function. In Stage 2 of training, our goal is to obtain a faithful stylization result from each of the style reference images $\{s_i\}_{i=1}^{N}$. Specifically, we use a weighted sum of the following different losses:

$$\begin{aligned}\mathcal{L}_{s_i}^{Stage2} = \lambda_{ML} \cdot \mathcal{L}_{ML}(\hat{c}_{s_i}, c_{gt}) + \lambda_p \cdot \mathcal{L}_p(\hat{c}_{s_i}, c_{gt}) \\ + \lambda_{CX} \cdot \mathcal{L}_{CX}(\hat{c}_{s_i}, c_{gt})\end{aligned} \tag{1}$$

where $\mathcal{L}_{ML}$, $\mathcal{L}_p$ and $\mathcal{L}_{CX}$ represent masked reconstruction, perceptual [18] and contextual [33] losses respectively, and the λ's control relative weights for their respective losses. We use the features extracted from VGG-19 [38] at layer $relu4\_1$ for $\mathcal{L}_p$, and $relu3\_1$ and $relu4\_1$ for $\mathcal{L}_{CX}$. Different from [33], GT image c_{gt} is used as the reference instead of style s_i for $\mathcal{L}_{CX}$ to compare with our output $\hat{c}_{s_i}$ because using s_i can cause severe structure distortion. $\mathcal{L}_{ML}$ in Eq. 1 can be expressed as:

$$\mathcal{L}_{ML}(\hat{c}_{s_i}, c_{gt}) = \|(\hat{c}_{s_i} - c_{gt}) \odot m_i\|_1 \tag{2}$$

where m_i is a mask used to generate s_i in our data generation scheme as shown in Fig. 5. Correspondingly, these three losses are used to encourage $\mathcal{S}$ (i) to faithfully stylize c at the pixel level for semantic regions that also appear in s_i and disregard other unrelated semantic regions, (ii) to faithfully stylize c at the semantic level, and (iii) to perform better semantic style transfer. In Stage 2 of training, we only use a single style reference for each c to reduce the computation complexity and stabilize the training of the single stylization subnet $\mathcal{S}$.

Method	PSNR↑	SSIM↑	LPIPS↓
ExColTran [52] + OPR-R	19.5796	0.7885	0.2563
ReHistoGAN [1] + OPR-R	20.0458	**0.7987**	0.2109
MAST [16] + OPR-R	19.0148	0.7853	0.2270
PCAPST [7] + OPR-R	19.1731	0.7908	0.2197
Ours	**21.2212**	0.7919	**0.2027**

Table 1. Quantitative results of modernization on synthetic dataset.

Method	NIQE↓	BRISQUE↓
OPR [44]	4.8705	21.4588
ExColTran [52] + OPR	4.9415	18.8971
ReHistoGAN [1] + OPR	4.8051	26.2557
MAST [16] + OPR	4.8111	18.9555
PCAPST [7] + OPR	4.7094	18.9860
Ours - Single	3.4737	15.5152
Ours - Multiple	**3.4487**	**15.4180**

Table 2. Quantitative results of modernization on real old photos.

In Stage 3, we train our merging-refinement subnet $\mathcal{M}$ by using weighted sum of four different losses:

$$\begin{aligned}\mathcal{L}^{Stage3} &= \lambda_{L1} \cdot \mathcal{L}_{L1}(\hat{c}, c_{gt}) + \lambda_p \cdot \mathcal{L}_p(\hat{c}, c_{gt}) \\ &+ \lambda_{sm} \cdot \mathcal{L}_{sm}(\hat{c}) + \lambda_{adv} \cdot \mathcal{L}_{adv}(\hat{c}, c_{gt})\end{aligned} \tag{3}$$

where $\mathcal{L}_{L1}$, $\mathcal{L}_p$, $\mathcal{L}_{sm}$ and $\mathcal{L}_{adv}$ are reconstruction, perceptual [18], local smoothness [54] and least square adversarial [31] losses respectively, and λ's control relative weights for corresponding losses. $\mathcal{L}_p$ in Eq. 3 and Eq. 1 refer to the same loss function. We use these four losses accordingly to encourage the merging-refinement subnet $\mathcal{M}$ to produce: (i) accurate merging and better refinement, (ii) perceptually plausible output, (iii) spatially smooth output, and (iv) realistic output.

5. Experiments

5.1. Experimental Settings

Training details. We use our proposed synthetic data generation scheme with the aforementioned multi-stage training strategy to train the network: (i) We train our PST network for five epochs; (ii) Then, we train our single stylization subnet $\mathcal{S}$ based on $\mathcal{L}^{Stage2}$ in Eq. 1 for two epochs, not to be overfitted for synthetic data, while freezing our PST network, where we set $\lambda_{ML} = 1$, $\lambda_p = 1$, and $\lambda_{CX} = 1$; (iii) Finally, we train our merging-refinement subnet $\mathcal{M}$ for three epochs which is sufficient while freezing both $\mathcal{S}$ and our PST network. The loss function to train $\mathcal{M}$ is $\mathcal{L}^{Stage3}$ in Eq. 3, where we set $\lambda_{L1} = 2$, $\lambda_p = 1$, $\lambda_{sm} = 3$, and $\lambda_{adv} = 0.2$. For all of the training, we use an ADAM optimizer [20] with a learning rate of $1e-4$ and batch size of 1 to optimize our network and discriminator (PatchGAN discriminator [17]). In addition, we apply a linear learning decay in the last epoch of the $\mathcal{M}$ training.

Baselines. Our work can be seen as handling a joint task of stylization and enhancement by using multiple references for old photo modernization. Since there are no baselines in reference-based old photo modernization, we compare to sequential models consisting of stylization then enhancement, which can perform the same task. Reversing the order (enhancement and then stylization), results in worse outcomes. For stylization, we employ four state-of-the-art (SOTA) methods as baselines: (i) exemplar-based colorization: transformer-based method (ExColTran [52]); (ii) recolorization: recolorization using color-controlled GAN (ReHistoGAN [1]); photorealistic style transfer (PST): semantic PST (MAST [16]) and PCA-based knowledge distillation PST (PCAPST [7]). Meanwhile, for enhancement, we employ SOTA no-reference-based old photo restoration method (OPR [44]), which can perform similar enhancement as ours. For the stylization baselines, we use their pre-trained models in the evaluation since they cannot be retrained with our synthetic data due to their different training strategies. However, note that these pre-trained models have already been trained to achieve the same goal of changing the overall look of input images based on the given reference image. Meanwhile, for enhancement, we use the pre-trained OPR model for real old photo evaluation, denoted as OPR, since it achieves better performance on real old photos, and retrain the OPR using our synthetic data and CHD training set for synthetic data evaluation, denoted as OPR-R. Since the four baselines can only utilize a single reference, we average the results of using different sets of references in quantitative evaluation, and randomly select one of two references for each input to the baseline networks in qualitative evaluation and user study.

Method	Top 1	Top 2	Top 3	Top 4	Top 5
OPR [44]	17.44	39.83	57.05	70.90	87.22
ExColTran [52] + OPR	1.62	5.13	10.77	24.87	47.27
ReHistoGAN [1] + OPR	7.91	32.27	61.84	83.80	96.92
MAST [16] + OPR	5.68	21.62	41.92	66.33	86.20
PCAPST [7] + OPR	10.98	28.50	44.87	61.97	84.66
Ours	**56.37**	**72.69**	**83.55**	**92.14**	**97.74**

Table 3. User study results. The percentage of user selection is shown.

Evaluation metrics. We evaluate all the methods on synthetically degraded and real old photos (CHD testing set). In synthetic degraded photos evaluation, we employ: 1) peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) to measure the pixel-level discrepancy between output and ground truth, 2) learned perceptual image patch similarity (LPIPS) [55] to measure the perceptual quality of the output. For evaluation on real old photos, we employ no-reference image quality assessment metrics such as NIQE [35] and BRISQUE [34], similar to [29,43,44], since the modernization ground truth photos do not exist.

5.2. Experimental Results

Quantitative comparison. We evaluate our method and baselines on a synthetic dataset and real-world old photos.

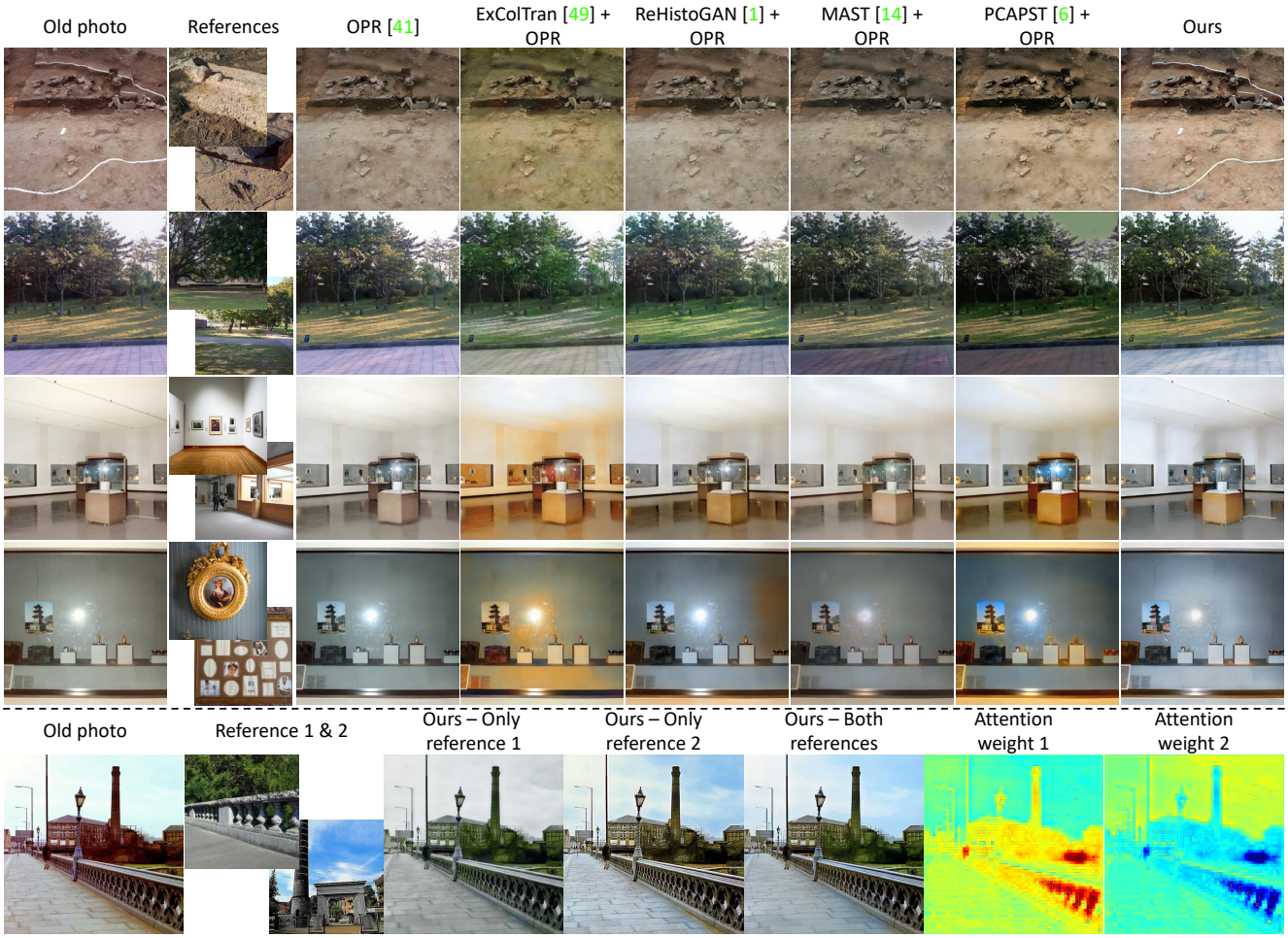

Figure 6. **Top**: qualitative results of modernization on real outdoor and indoor old photos. The baselines use top-left reference as their reference. **Bottom**: Attention weight visualization, blue (lowest)-red (highest) color coded. Our method can select appropriate styles from multiple references depending on the availability of similar objects to achieve better modernization. (Zoom in for a better view)

In synthetic dataset evaluation, we evaluate all methods, including ours, in a single-reference-based scenario since the baselines cannot utilize more than one reference by using ADE20K validation set [56] that includes semantic segmentation masks. Specifically, we generate 1,000 evaluation pairs, each consisting of a synthetic degraded photo and a reference image, by randomly degrading half set of the semantic regions using our synthetic data generation scheme, e.g., only the surfer in Fig. 5. Table. 5 shows our method achieves the best PSNR and LPIPS score, which means our method can effectively utilize the references to jointly stylize and enhance the synthetically degraded images, thus generating an output similar to the ground truth both in the pixel and semantic levels. In terms of SSIM, we achieve the second-best compared to recolorization (ReHistoGAN [1] + OPR-R) since our method can change the other aspects besides color, such as texture and luminance, which may result in a lower SSIM score. Interestingly, compared to other PST baselines, especially MAST [16], which is designed to perform semantic style transfer, our method achieves the most accurate stylization (PSNR and LPIPS) while still preserving the structure (SSIM), which are two important aspects in PST. A similar observation can be seen for real old photos evaluation shown in Table. 6, where our method outperforms other baselines significantly using a single reference and further improves the performance by using multiple references.

Qualitative comparison. As shown in Fig. 6, no-reference OPR [44] can restore both structured (SD) and unstructured degradations (UD). However, SD restoration cannot generalize well to real old photos since it significantly degrades the important regions of the original photos. In addition, it fails to modernize some old photos because the overall styles still remain similar to the original old photos. ExColTran [52], which can only use luminance information for semantic matching, fails to locally change the color of old photos, thus producing unnatural results. Meanwhile, ReHistoGAN [1] can better recolorize old photos,

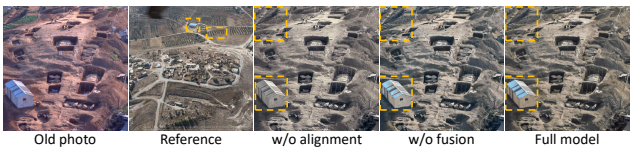

Figure 7. Ablation study on the single stylization subnet.

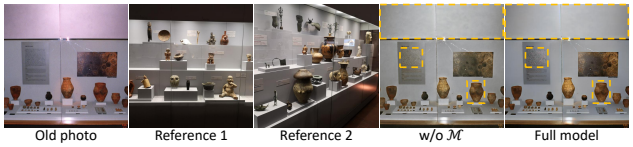

Figure 8. Ablation study on the merging-refinement subnet.

producing more modern-looking images than only OPR. Compared to other PST methods combined with OPR, and other baselines, our method achieves better local and global PST and yields enhancement, consequently achieving better modernization. Moreover, our method can utilize multiple references better in all examples, e.g., the second row of Fig. 6, where styles of tree leaves and road come from the first and second references, respectively. Meanwhile, the fourth row shows the generalization of our method, which can handle unrelated references well. The bottom part under the dotted line in Fig. 6 shows the visualization of spatial attention weights, where our method can select appropriate styles for each semantic object in an old photo from multiple references to achieve better modernization, e.g., the bridge and sky style in the first and second reference respectively. All in all, our method can modernize old photos better than the baselines by leveraging multiple modern photo references, even though it has not been trained with any old photos. These results also show that restoring the degradation of old photos cannot guarantee the outputs to look modernized, but changing their styles can contribute to the *modern* look more than restoring the degradation.

User study. We conduct a user study to compare the modernization results of our method with those of the baselines. Specifically, we select 130 photos from the CHD testing set and ask 18 users to rank the modernization results. As shown in Table. 3, our method outperforms other baselines with 56.37% chance selected as the best method.

5.3. Analysis

Ablation study on the single stylization subnet. We analyze the contribution of each module in the single stylization subnet $\mathcal{S}$. As shown in Fig. 7, the subnet fails to accurately transfer the local styles of objects, e.g., the styles of the blue building and grass, when the alignment module is removed. Even though the subnet can perform better local style transfer of building and grass regions with the alignment module, the stylization results are not smooth, which may produce unnatural results. Thus, adding a fusion module that merges global and local styles can produce smoother stylization locally and globally.

Figure 9. Limitation of our method.

Ablation study on the merging-refinement subnet. To evaluate the contribution of the merging-refinement subnet $\mathcal{M}$, we change this subnet to a simple concatenation between multiple stylization features and feed the concatenated features into several convolution blocks (denoted as w/o $\mathcal{M}$). As shown in Fig. 8, without $\mathcal{M}$, the network cannot select appropriate styles from different references and fail to enhance the results. In addition, retraining the network is required to use a different number of references between inference and training. With our $\mathcal{M}$, we can adaptively choose the number of references without retraining. Results of other ablation studies and modernization using more than two references are in *Supplementary Material*.

Limitation. The limitation of our work mainly comes from the selection of references. As shown in Fig. 9, our method may produce unsatisfying modernization when a related object in the references has a style that does not enhance the old photo. However, finding references that contain a similar local object with a modern style in an automated way is highly challenging using existing image retrieval methods. Moreover, using VGG feature space matching similar to [11] fails to produce semantically similar references due to the domain gap between old and modern photos.

6. Conclusion

In this paper, we first proposed old photo modernization by using multiple references. In order to perform modernization, we proposed MROPM, which performs old photo stylization using multiple references via photorealistic style transfer and enhancement in one unified framework. Thanks to the generalization of PST and our synthetic data generation scheme, our work outperforms baselines for real-world old photos, even without using any old photos during the training. Furthermore, we analyze that our method can select appropriate styles from multiple references, further improving the modernization performance. Also, we propose an old color photos dataset CHD consisting of natural indoor and outdoor scenes to spur future research in the domain.

Acknowledgment. This research was supported by Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2020 (Project Name: CHIC, Project Number: R2020040045, Contribution Rate: 100%). We would like to thank Gimhae, Jeju, and National Museum of Korea for the old photos.

References

[1] Mahmoud Afifi, Marcus A Brubaker, and Michael S Brown. Histogan: Controlling colors of gan-generated and real images via color histograms. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7941–7950, 2021. 2, 6, 7, 15, 16, 19, 20, 21, 22, 23, 28, 29, 35, 36, 37, 38

[2] Jie An, Haoyi Xiong, Jun Huan, and Jiebo Luo. Ultrafast photorealistic style transfer via neural architecture search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10443–10450, 2020. 2, 4, 14

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2018. 3

[4] Peter J Burt and Edward H Adelson. The laplacian pyramid as a compact image code. In *Readings in computer vision*, pages 671–679. Elsevier, 1987. 4, 14

[5] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018. 5

[6] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*, 2022. 16

[7] Tai-Yin Chiu and Danna Gurari. Pca-based knowledge distillation towards lightweight and content-style balanced photorealistic style transfer models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7844–7853, 2022. 2, 6, 14, 15, 16, 19, 20, 22, 23, 26, 28, 29, 35, 36, 37, 38

[8] Tai-Yin Chiu and Danna Gurari. Photowct2: Compact autoencoder for photorealistic style transfer resulting from blockwise training and skip connections of high-frequency residuals. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2868–2877, 2022. 2

[9] Ioannis Giakoumis, Nikos Nikolaidis, and Ioannis Pitas. Digital image processing techniques for the detection and removal of cracks in digitized paintings. *TIP*, 15(1):178–188, 2005. 1, 3

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[11] Mingming He, Dongdong Chen, Jing Liao, Pedro V Sander, and Lu Yuan. Deep exemplar-based colorization. *ACM Transactions on Graphics (TOG)*, 37(4):1–16, 2018. 2, 8, 12

[12] Mingming He, Jing Liao, Dongdong Chen, Lu Yuan, and Pedro V Sander. Progressive color transfer with dense semantic correspondences. *ACM Transactions on Graphics (TOG)*, 38(2):1–18, 2019. 2

[13] Zhiyuan Hu, Jia Jia, Bei Liu, Yaohua Bu, and Jianlong Fu. Aesthetic-aware image style transfer. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3320–3329, 2020. 2

[14] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 4, 14, 17

[15] Zixuan Huang, Jinghuai Zhang, and Jing Liao. Style mixer: Semantic-aware multi-style transfer network. In *Computer Graphics Forum*, volume 38, pages 469–480. Wiley Online Library, 2019. 3

[16] Jing Huo, Shiyin Jin, Wenbin Li, Jing Wu, Yu-Kun Lai, Yinghuan Shi, and Yang Gao. Manifold alignment for semantically aligned style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14861–14869, 2021. 2, 6, 7, 15, 16, 19, 20, 22, 23, 28, 29, 35, 36, 37, 38

[17] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 6, 18

[18] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 5, 6

[19] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 3

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[21] Junyong Lee, Hyeongseok Son, Gunhee Lee, Jonghyeop Lee, Sunghyun Cho, and Seungyong Lee. Deep color transfer using histogram analogy. *The Visual Computer*, 36(10):2129–2143, 2020. 2

[22] Xueting Li, Sifei Liu, Jan Kautz, and Ming-Hsuan Yang. Learning linear transformations for fast image and video style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3809–3817, 2019. 2

[23] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. *Advances in neural information processing systems*, 30, 2017. 2, 4, 14, 17

[24] Yijun Li, Ming-Yu Liu, Xueting Li, Ming-Hsuan Yang, and Jan Kautz. A closed-form solution to photorealistic image stylization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 453–468, 2018. 2

[25] Jing Liao, Yuan Yao, Lu Yuan, Gang Hua, and Sing Bing Kang. Visual attribute transfer through deep image analogy. *ACM Transactions on Graphics (TOG)*, 36(4):1–15, 2017. 2

[26] Jixin Liu, Rui Chen, Shipeng An, and Heng Zhang. Cg-gan: Class-attribute guided generative adversarial network for old photo restoration. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5391–5399, 2021. 3

[27] Peng Lu, Jinbei Yu, Xujun Peng, Zhaoran Zhao, and Xiaojie Wang. Gray2colornet: Transfer more colors from reference image. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3210–3218, 2020. 2

[28] Fujun Luan, Sylvain Paris, Eli Shechtman, and Kavita Bala. Deep photo style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4990–4998, 2017. 2

[29] Xuan Luo, Xuaner Zhang, Paul Yoo, Ricardo Martin-Brualla, Jason Lawrence, and Steven M Seitz. Time-travel rephotography. *ACM Transactions on Graphics (TOG)*, 40(6):1–12, 2021. 1, 3, 6, 13, 14

[30] Octavian-Mihai Machidon and Mihai Ivanovici. Digital color restoration for the preservation of reversal film heritage. *Journal of Cultural Heritage*, 33:181–190, 2018. 3, 13

[31] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017. 6

[32] Roey Mechrez, Eli Shechtman, and Lihi Zelnik-Manor. Photorealistic style transfer with screened poisson equation. *arXiv preprint arXiv:1709.09828*, 2017. 2

[33] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. The contextual loss for image transformation with non-aligned data. In *Proceedings of the European conference on computer vision (ECCV)*, pages 768–783, 2018. 5

[34] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012. 6, 12, 13, 21

[35] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 6, 21

[36] Ying Qu, Zhenzhou Shao, and Hairong Qi. Non-local representation based mutual affine-transfer network for photorealistic stylization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7046–7061, 2021. 2

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 5, 18

[38] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2, 5, 12, 13

[39] Aditya Singh, Alessandro Bay, and Andrea Mirabile. Assessing the importance of colours for cnns in object recognition. In *NeurIPS 2020 Workshop SVRHM*, 2020. 2

[40] F Stanco, Giovanni Ramponi, and A De Polo. Towards the automated restoration of old photographic prints: a survey. In *The IEEE Region 8 EUROCON 2003. Computer as a Tool.*, volume 2, pages 370–374. IEEE, 2003. 1, 3

[41] Jheng-Wei Su, Hung-Kuo Chu, and Jia-Bin Huang. Instance-aware image colorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7968–7977, 2020. 2

[42] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 18

[43] Ziyu Wan, Bo Zhang, Dongdong Chen, and Jing Liao. Bringing old films back to life. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17694–17703, 2022. 6

[44] Ziyu Wan, Bo Zhang, Dongdong Chen, Pan Zhang, Dong Chen, Jing Liao, and Fang Wen. Bringing old photos back to life. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2747–2757, 2020. 1, 2, 3, 6, 7, 15, 16, 19, 20, 23, 29, 35, 36, 37, 38

[45] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 4, 18

[46] Phillip Whitt. *Beginning photo retouching and restoration using GIMP*. Springer, 2014. 1

[47] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 4, 18

[48] Yanze Wu, Xintao Wang, Yu Li, Honglun Zhang, Xun Zhao, and Ying Shan. Towards vivid and diverse image colorization with generative color prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14377–14386, 2021. 2

[49] Xide Xia, Meng Zhang, Tianfan Xue, Zheng Sun, Hui Fang, Brian Kulis, and Jiawen Chen. Joint bilateral learning for real-time universal photorealistic style transfer. In *European Conference on Computer Vision*, pages 327–342. Springer, 2020. 2

[50] Runsheng Xu, Zhengzhong Tu, Yuanqi Du, Xiaoyu Dong, Jinlong Li, Zibo Meng, Jiaqi Ma, Alan Bovik, and Hongkai Yu. Pik-fix: Restoring and colorizing old photo. *arXiv preprint arXiv:2205.01902*, 2022. 1, 2, 3, 12, 13, 14, 26, 27

[51] Zhongyou Xu, Tingting Wang, Faming Fang, Yun Sheng, and Guixu Zhang. Stylization-based architecture for fast deep exemplar colorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9363–9372, 2020. 2

[52] Wang Yin, Peng Lu, Zhaoran Zhao, and Xujun Peng. Yes," attention is all you need", for exemplar based colorization. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2243–2251, 2021. 2, 6, 7, 14, 15, 16, 19, 20, 21, 23, 28, 29, 35, 36, 37, 38

[53] Jaejun Yoo, Youngjung Uh, Sanghyuk Chun, Byeongkyu Kang, and Jung-Woo Ha. Photorealistic style transfer via wavelet transforms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9036–9045, 2019. 2, 4, 5, 14, 16, 17

[54] Bo Zhang, Mingming He, Jing Liao, Pedro V Sander, Lu Yuan, Amine Bermak, and Dong Chen. Deep exemplar-based video colorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8052–8061, 2019. 6

[55] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[56] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 7

Supplementary Material

Figure 10. Diverse examples of outdoor and indoor scenes from our dataset. The images contain various kinds of degradations, especially color fading.

7. Details of Proposed CHD Dataset

7.1. Details of Our CHD Dataset

In order to curate our Cultural Heritage Dataset (CHD), we collect old photos produced in the 20th century. Specifically, we collect these old photos in the form of reversal films or papers from three national museums in Korea, i.e., the National Museum of Korea, Gimhae National Museum, and Jeju National Museum. After collecting old photos, we scan the photos in resolution varying from 4K to 8K. The content of the photos is indoor and outdoor scenes of cultural heritage, such as special exhibitions and excavation ruins. For the degradation, the photos contain a little scratch and crack degradations since they have been well preserved and stored carefully due to their important values. However, they contain various degrees of unstructured degradations and color fading. Fig. 10 shows the diversity of indoor and outdoor scenes in our dataset with varying degrees of degradations such as blur, noise, scratch and crack, and color fading. In addition to these degradations, our dataset also contains various real-world artifacts that may provide benefits for the community such as reflection, flash, etc.

After the collection, we filter out several old photos that contain sensitive information, e.g., front-facing faces, distinguished faces, and license plates. In total, 644 old color photos are obtained through filtering, where 383, 147, and 114 photos are from the National Museum of Korea, Gimhae National Museum, and Jeju National Museum respectively. Then, we randomly divide these 644 old photos into train and test sets with a proportion of 8:2. Note that we also preserve the same ratio of images in the train and test set for each museum name. In total, we obtain 514 photos for the train set and 130 photos for the test set. The train set is used to train the old photo restoration baseline that needs to be trained using real old photos since it works by reducing the domain gap between real and synthetic old photos. Meanwhile, our method does not use any old photos during the training since our method utilizes photorealistic style transfer that can work on any photo, including old photos. Since the scanned photos have a resolution of 4K to 8K, we further preprocess the photos. Specifically, we resize these photos to make the short side (width or height) have a resolution of 1024, then we center-crop the images, resulting in a resolution of 1024×1024.

Since our task is reference-based old photo modernization, we further collect photos as references by automatically crawling CC-Licensed images with similar contexts from an internet search using the crawling tool[1] for the test set, where each old photo serves as the query of the search. Approximately 100 reference photos for each old photo in the test set are obtained. Then, we select one to two modern photos manually as the references for each old photo, which are then resized and center-cropped, similar to the resize and crop operation applied to the old photos resulting in a resolution of 1024×1024. We tried to perform the selection automatically by selecting reference photos that have the largest cosine similarity in the VGG-19 [38] feature space, using a similar idea to [11, 50], and the best BRISQUE [34] score. However, Fig. 11

[1] https://github.com/hardikvasa/google-images-download

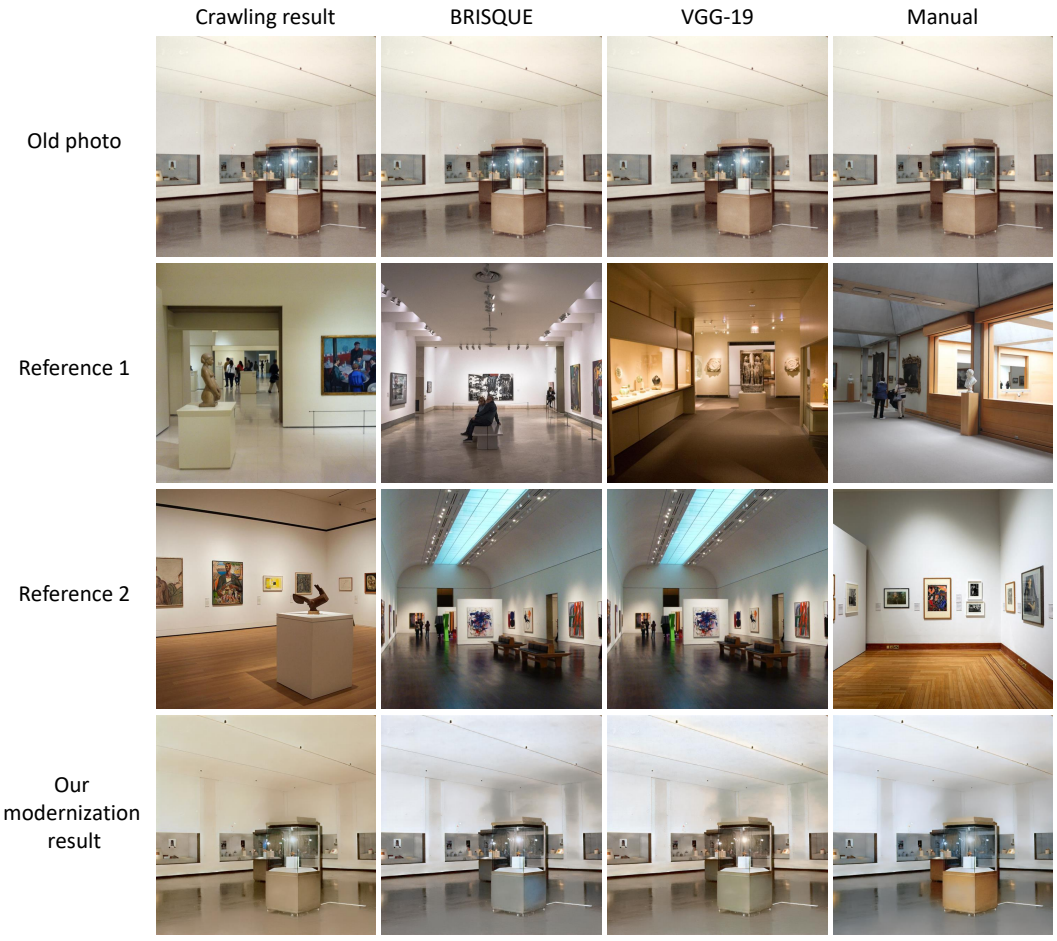

Figure 11. Automatic references selection trial. We try to automate reference selection by using BRISQUE [34] and VGG-19 [38] feature similarity.

shows that the automatic selection sometimes fails to obtain modern photo references with a modern style. The references from the crawling result have a similar context to the old photo. However, the references have the characteristic of old photos, i.e., hazy and unsaturated colors. This observation is similar to the automatic selection using the VGG-19 feature space, where the selection algorithm tends to select photos with similar old photos characteristics, e.g., sepia color. Meanwhile, selecting references with the best BRISQUE score cannot obtain references with similar contexts, e.g., no similar objects for the showcase. Note that since we do not own any of the reference photos, we will only release the link for the reference images and the attribution in the dataset.

7.2. Comparison of CHD and Other Old Photos Dataset

There are two other public old photo datasets, such as the Historical Wiki Face Dataset (HWFD) [29] and RealOld [50]. However, when this paper was submitted, RealOld [50] had not been published yet. Table. 4 shows the comparison between our and other datasets. Our dataset is mainly focused on old color photos produced during the 20th century using reversal film [30], which have specific degradations such as color fading (shown in Fig. 10)

	HWFD [29]	RealOld [50]	Ours
Number of images	224	200	644
Era	19-20th century	-	20th century
Content type	Face	Portrait	Indoor & outdoor natural scenes
Color space	Greyscale	Greyscale	Color
Resolution	133 × 133 until 1024 × 1024	-	1024 × 1024
Expert ground-truth	✗	✓	✗

Table 4. Comparison between our dataset and other public old photos datasets in several factors.

and have not yet been analyzed before. Meanwhile, other datasets contain greyscale photos. Regarding the diversity of content, our dataset contains complex and diverse scenes of indoor and outdoor natural scenes, as shown in Fig. 10. In contrast, other datasets only contain portrait and face photos which are much simpler than natural scene photos. In addition to the complexity of the scenes, our dataset also has a larger number of images compared to the two other datasets. Fig. 12 shows some visual examples of our and

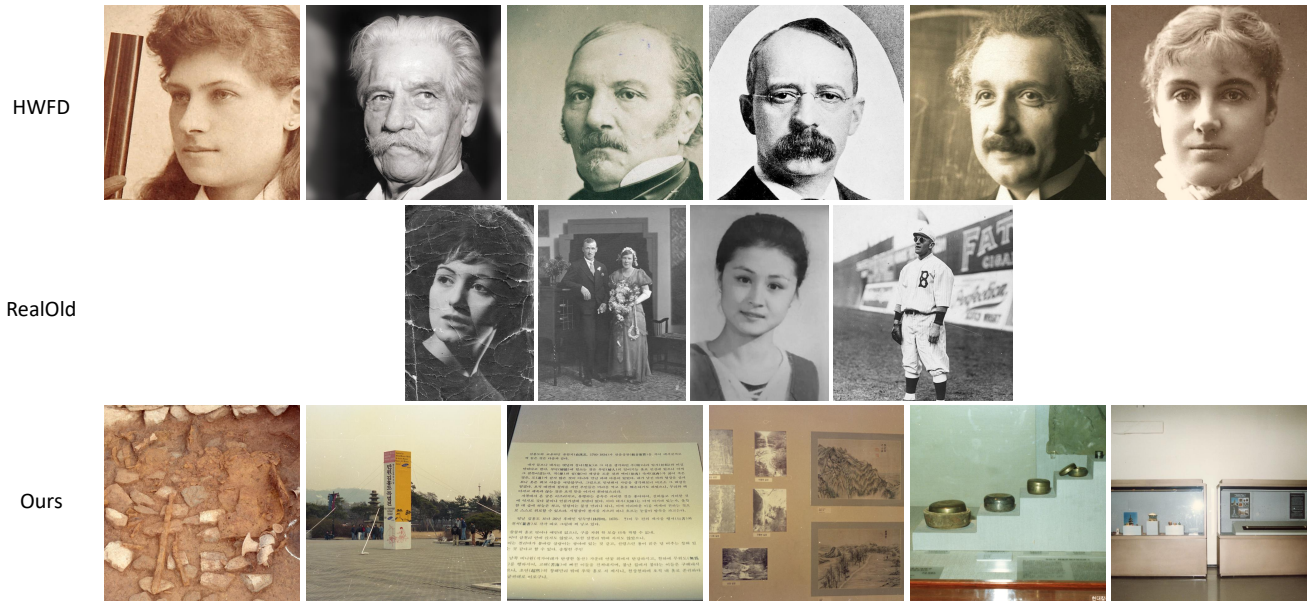

Figure 12. Comparison between our CHD dataset and other datasets (HWFD [29] and RealOld [50]). Our CHD dataset has the most complex and diverse scenes compared to other datasets. In addition, our dataset also contains unique color fading artifacts.

other datasets.

8. Results on Real Old Photos in The Wild

Fig. 13 and Fig. 14 show the generalization and robustness of our method when applied to real old photos in the wild. The first and second examples of Fig. 13 show that our method outperforms other baselines in modernizing old color photos and even can work for greyscale photos. Interestingly, our method can achieve natural modernization results on greyscale old photos (second example) even when compared to the colorization baseline (ExColTran [52] + OPR). For the third example in Fig. 13, our method achieves the second-best performance compared to 'PCAPST [7] + OPR', where our method can better stylize the trees but fail to stylize the big castle, caused by our alignment module that may think that castle and building are different.

Fig. 14 shows additional examples of modernization on greyscale old photos in the wild. The same observation can be seen where our method outperforms other baselines even when compared to the colorization baseline (ExColTran [52] + OPR). Interestingly, we can handle paper blotches in the second example of Fig. 14 even though our method is not trained with this kind of artifact. Meanwhile, the baseline OPR trained with this kind of artifact further highlights the paper blotches artifact instead of removing them. In addition, compared to other reference-based methods, our method can better match similar semantic regions between the old photo and references even though the viewpoint and scale are significantly different. For example, the Eiffel tower in the old photo of the first and second examples of Fig. 14 have different viewpoints and scales compared to the references. However, our method can faithfully match the Eiffel tower in the old photo and references, thus resulting in better stylization and modernization results. Note that our network can achieve all these results without using old photos during training.

9. Details & Analyses of Our Photorealistic Style Transfer (PST) Network

9.1. Comparison Between Our Photorealistic Style Transfer (PST) Network and WCT2 [53]

Fig. 15 shows the comparison between our PST network and WCT2 [53] architecture. We propose our photorealistic style transfer (PST) network to address some drawbacks of the concatenated version of the WCT2 network with progressive stylization (style transfer). Specifically, our PST network only transfers a single high-frequency component in level-0 of the Laplacian pyramid representation [4]. Meanwhile, WCT2 [53] transfers three different high-frequency components of wavelet-based skip connection. This modification addresses the "short circuit" issue explored in [2], which makes the stylization of WCT2's decoder part only has an effect when applied in the last decoder block (shown in Fig. 18). Then, we only apply progressive stylization in the decoder part, especially the last two decoder blocks, which achieves the best trade-off between the stylization effect and the photorealism. The last improvement is we use differentiable adaptive instance normalization (AdaIN) [14] instead of the non-differentiable whitening-and-coloring transformation (WCT) [23] to enable the learning and prediction of local style, which can enable our single stylization subnet to perform local style transfer without any semantic segmentation mask. Further analyses of the drawbacks of WCT2 [53] are explained in the following subsection.

9.2. Additional Analyses

Limitations of WCT2 [53] for old photo modernization. There are two main limitations of WCT2 [53] that can prevent its application on real-world old photos. The first limitation is the unnatural global style transfer results shown in Fig. 16, where this unnatural result may make the photo look like an old photo instead of mod-

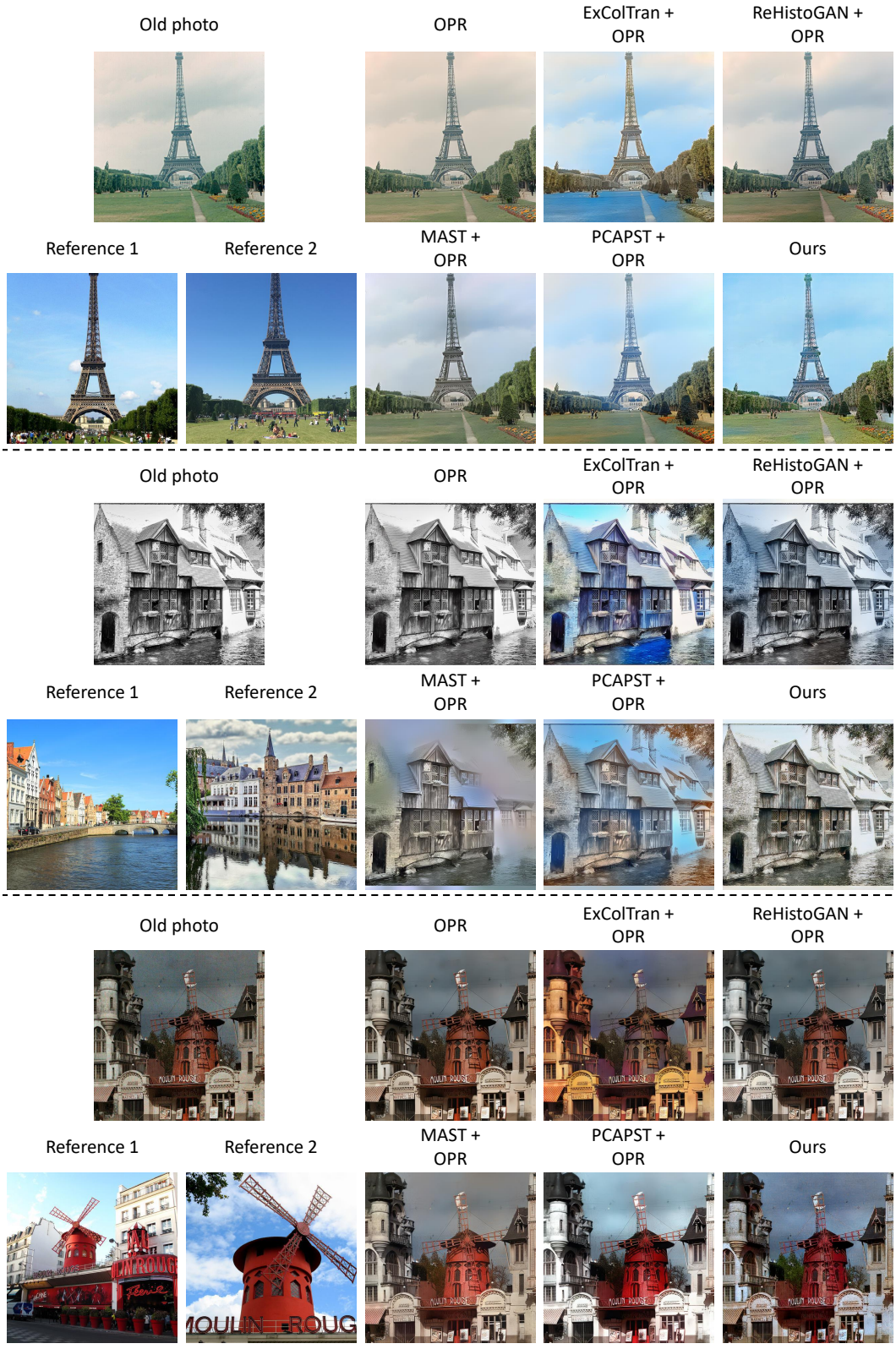

Figure 13. Comparison of modernization on old photos in the wild between our method and other baselines. In most cases, our method outperforms other methods (OPR [44], ExColTran [52] + OPR, ReHistoGAN [1] + OPR, MAST [16] + OPR, and PCAPST [7] + OPR) in modernizing old photos in the wild showing the robustness of our method. Other reference-based baselines use reference 1 as their reference.

Figure 14. Comparison of modernization on greyscale old photos in the wild between our method and other baselines. Our method outperforms other methods (OPR [44], ExColTran [52] + OPR, ReHistoGAN [1] + OPR, MAST [16] + OPR, and PCAPST [7] + OPR) in modernizing greyscale old photos in the wild showing the generalization of our method. Other reference-based baselines use reference 1 as their reference. In these examples, our method can better match the corresponding semantic regions between the old photo and multiple references even though the **viewpoint and scale are significantly different** (e.g., the viewpoint and scale of the tower).

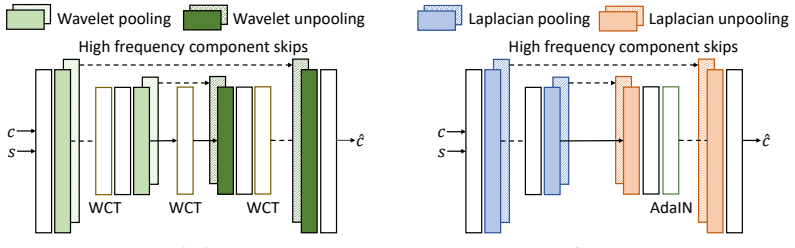

Figure 15. Comparison between our PST network and WCT2 [53].

ernizing them. Meanwhile, Fig. 17 shows the second limitation of WCT2. We generate the semantic segmentation masks using VIT-Adapter [6], which is one of the state-of-the-art models in the semantic segmentation task for the examples in Fig. 17. As can be seen, WCT2 needs a near-perfect semantic segmentation mask to produce satisfactory results of local style transfer. However, generating a near-perfect segmentation mask moreover for old photos is highly challenging even with one of the SOTA networks. There-

Figure 16. Unnatural global style transfer result of WCT2 [53].

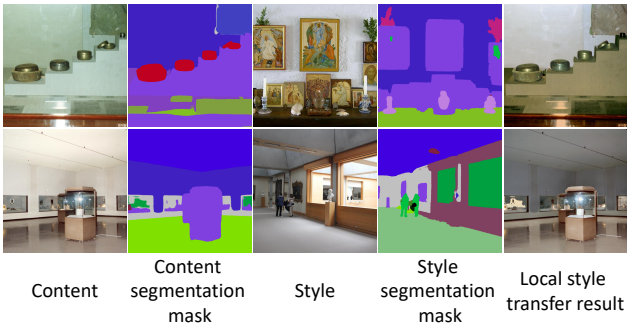

Figure 17. Unnatural local style transfer result of WCT2 [53].

fore, we propose our single stylization subnet to overcome these limitations, especially to perform local style transfer without any semantic segmentation mask and produce natural global and local style transfer results.

The trade-off between stylization and photorealism. Fig. 18 shows that applying progressive stylization in different decoder blocks does not affect the original concatenated version of the WCT2 [53] network. Thus, we modify the skip connection using the aforementioned laplacian-based skip connection to overcome this limitation. In terms of progressive stylization, we only apply feature transformation on the decoder part using AdaIN to perform style transfer, especially the last two decoder blocks, to achieve the best trade-off between stylization and photorealism. As shown in Fig. 18, applying feature transformation in the deep feature space (D4 or D3) produces stylization artifacts in the output, making them look non-photorealistic. Thus, applying feature transformation in the shallow feature space (D1 and D2) achieves the best stylization and photorealistic results.

Comparison between different feature transformations. In our PST network, we use AdaIN instead of WCT, which is commonly used as the feature transformation to perform photorealistic style transfer. We observe that using AdaIN achieves more improved stylization with better color saturation which can help us to achieve superior modernization as shown in Fig. 19. In addition, AdaIN feature transformation is also differentiable, which

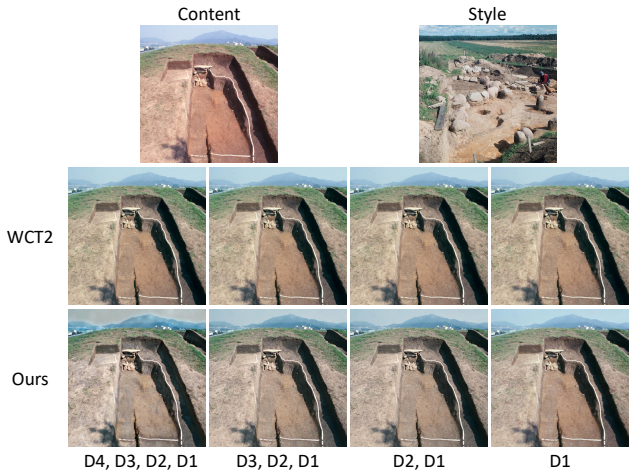

Figure 18. Applying feature transformation in different decoder blocks of WCT2 [53] and our PST network. D denotes the decoder block, while the number denotes the decoder block number (a higher number denotes the decoder block in deeper feature space).

can help us achieve local style transfer without any semantic segmentation mask since we want to learn and predict the local styles instead of computing them.

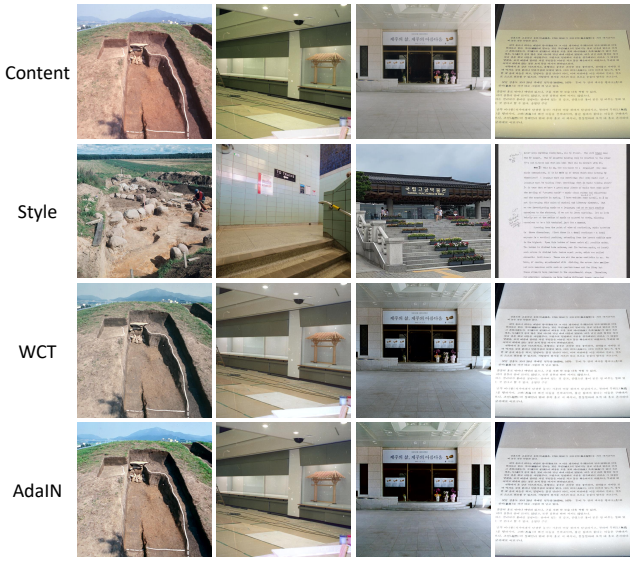

Figure 19. Visual results comparison between AdaIN [14] and WCT [23] feature transformations.

10. Details of MROPM-Net Architecture

10.1. Single Stylization Subnet

The detailed architecture of our single stylization subnet $\mathcal{S}$ can be seen in Fig. 20. We only describe additional parts that have not been described in the main paper, which is the alignment module (green part). Given an old photo c, modern style reference s_i, and extracted local style code (ψ_l^1, ψ_l^2), the alignment module aligns

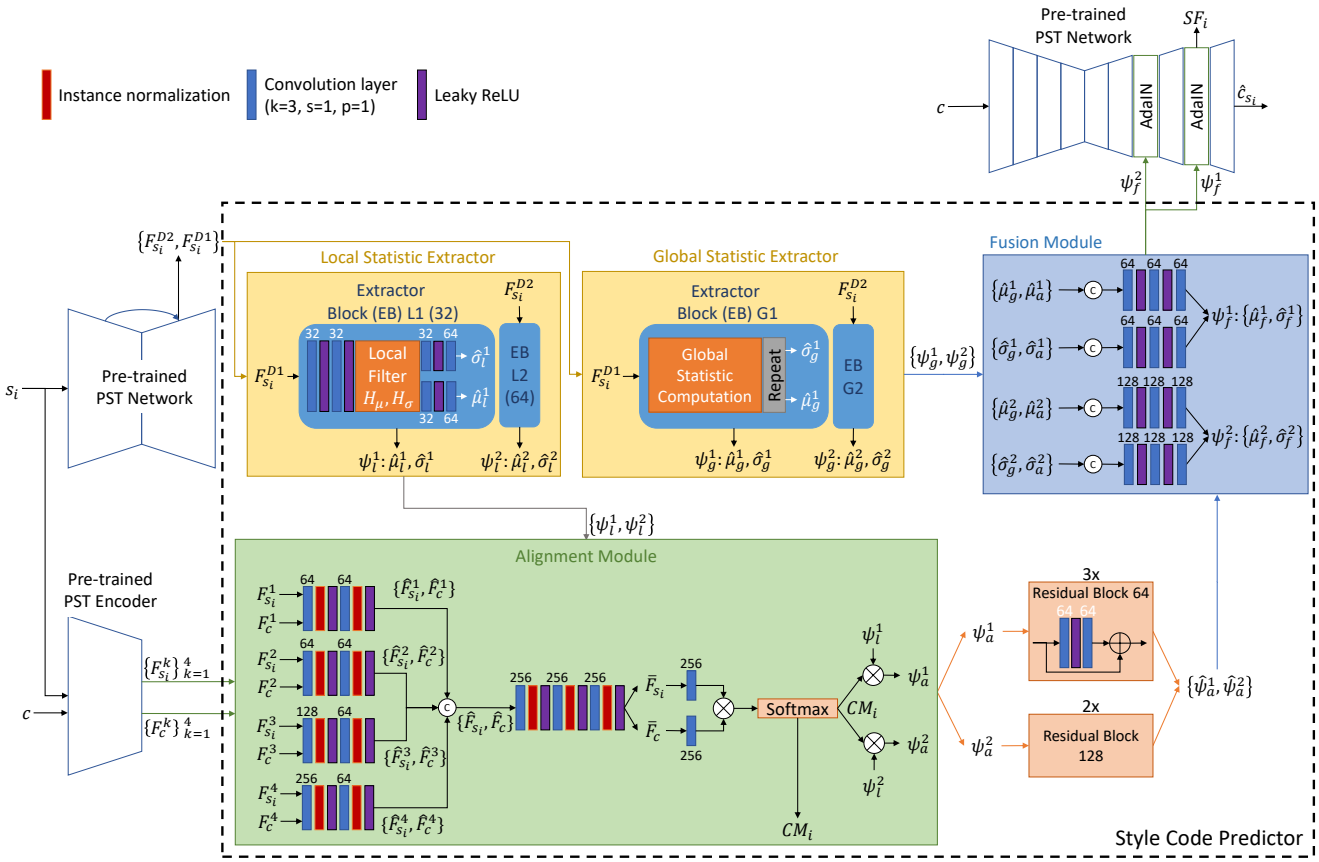

Figure 20. Detailed architecture of our single stylization subnet $\mathcal{S}$.

the local style code of s_i to c. In the alignment module, we map the extracted multi-level feature maps $\{F_c^k\}_{k=1}^4$ and $\{F_{s_i}^k\}_{k=1}^4$ for both c and s_i, respectively using shared convolution blocks, and perform matrix multiplication between mapped features to obtain correlation matrix CM_i similar to non-local attention [45]. Since different feature maps have different spatial resolutions, we map them into the same spatial resolution, which is the spatial resolution of the deepest features, i.e., the spatial resolution of F_c^4, using nearest neighbor interpolation. The next step is to align the local style code (ψ_l^1, ψ_l^2) using correlation matrix CM_i via matrix multiplication, thus resulting in aligned style codes (ψ_a^1, ψ_a^2). Since different $\{\psi_l^k\}_{k=1}^2$ have a different spatial resolution than CM_i, we use nearest neighbor interpolation to map $\{\psi_l^k\}_{k=1}^2$ to the same spatial resolution of CM_i and then map it back to the original spatial resolution after multiplication with CM_i. Then, we use three residual blocks to refine ψ_a^1 and two residual blocks to refine ψ_a^2.

10.2. Merging-Refinement Subnet

Fig. 21 shows the detailed architecture of our merging-refinement subnet $\mathcal{M}$. We show the details of the spatial attention module [47]. Additionally, we show the details of convolution blocks that consist of several convolution layers and leaky ReLU activation, in order to get the intermediate merging output $\hat{c}_m$. For the details of the refinement subnet, we follow the notation of U-Net [37] architecture in [17]. Specifically, the encoder-decoder architecture is based on the following:

encoder:
$C64 - C128 - C256 - C512 - C512 - C512 - C512$

decoder:
$CD512 - CD512 - CD512 - CD256 - CD128 - CD64$

The activation functions in the encoder are leaky ReLUs with a slope of 0.2, while the activation functions in the decoder are ReLUs. Then, we use a single convolution layer, followed by a single Tanh function, to map the features of the last layer decoder to the RGB channels representing modernized images. We use instance normalization layers [42] in the U-Net architecture.

11. Additional Details of Synthetic Data Generation Scheme

In this section, we describe additional details of the synthetic data generation scheme, such as the style variant transformation which includes the color jittering and unstructured degradation, and the details of the style invariant transformations. To get the degraded images via the style variant transformation, we first perform color jittering on the images by randomly changing the brightness, contrast, saturation, and hue with the magnitude of 0.2, 0.2, 0.4, and 0.4, respectively. In addition, we also apply a random sequence of mixed unstructured degradation after color jittering. Specifically, we choose a random sequence of the following degradations:

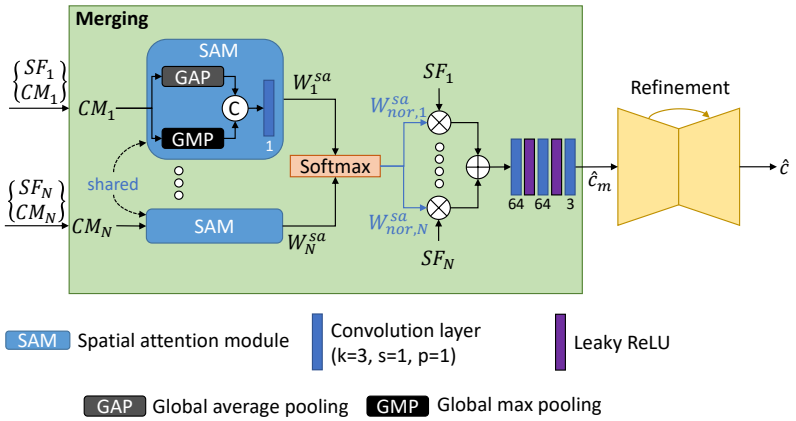

Figure 21. Detailed architecture of our merging-refinement subnet $\mathcal{M}$.

- Gaussian blur with a probability of 50%, where the kernel size is chosen randomly between 3, 5, and 7 and the standard deviation $\sigma = 0.004 - 0.02$.
- Random noise with a probability of 50%, where we choose randomly between gaussian noise with ($\mu = 0, \sigma = 0.02 - 0.04$), and speckle noise with ($\mu = 0, \sigma = 0.02 - 0.08$).
- Random resizing artifacts with a probability of 50%. The resizing artifact is generated by downsampling the spatial resolution of the image to the half size using bicubic downsampling and then upsampling the downsampled image back to the original spatial resolution by using nearest or bilinear interpolation, which is chosen randomly.
- Random JPEG artifact with a probability of 50% where the compressed quality is a random number between 40% to 100% (no artifact).

How to adapt to new degradation. In this work, we focus more on unstructured degradation since at the time this work was published, no public scratches data were available. However, one can easily add new degradation or special artifacts into our style variant transformation. By doing so, the network can adapt and handle new degradation or artifacts.

For the style invariant transformations, we apply a sequence of the following operators:

1. Random $k \times 90^\circ$ rotations chosen randomly between 90°, 180°, and 270°.
2. Random translation for regions that can be translated (the translated regions still remain inside the boundary of the image after the translation).
3. Random left-to-right flipping.
4. Random up-to-down flipping.

12. Additional Experiments on Baselines

In the main paper, we propose to use a sequence of stylization and enhancement as the baselines to compare with our method since our method can perform both stylization and enhancement jointly. In this section, we first show the results of retraining the baseline OPR [44] using our synthetic data and our CHD training set since we use the original pretrained baseline model in the main paper. Then, we show the results of using only stylization to show that an enhancement method is required to further improve the results. Furthermore, we show the results of using a sequence of enhancement and stylization (reversed order) as the baselines compared to a sequence of stylization and enhancement.

Baselines. We choose four different state-of-the-art (SOTA) stylization methods, from exemplar-based colorization [52], recolorization [1], and photorealistic style transfer [7, 16]. Even though exemplar-based colorization and recolorization can only change the color, we still use them as the baseline since changing the color can also affect the look of an image. Our user study also shows that the recolorization baseline achieves better results than other baselines. Specifically, we choose the following baselines that act as the stylization:

- exemplar-based colorization: transformer-based method (ExColTran [52])
- recolorization: color-controlled GAN method (ReHistoGAN [1])
- photorealistic style transfer (PST): semantic PST (MAST [16]) and PCA-based knowledge distillation PST (PCAPST [7])

For the enhancement, we use the SOTA of old photo restoration (OPR [44]) as the baseline. Note that, OPR is used for enhancement since OPR can handle both unstructured degradation and structured degradation. Thus, it is used as an enhancement method in conjunction with stylization baselines for fair comparison since our method can perform both stylization and enhancement.

Qualitative results of retraining the baseline OPR [44]. As mentioned in the main paper, we use the pretrained model of baseline OPR [44] rather than retraining it for real old photo evaluation. The results with the retrained baseline OPR [44] are shown in Fig. 22 both when using their synthetic data and our CHD training set (denoted as OPR-R-Old) and when using our synthetic data and our CHD training set (denoted as OPR-R). Interestingly, training using their synthetic data and our CHD training set (OPR-R-Old) results in worse performance in real old photos. This may suggest that OPR [44] requires a large number of old photos since the authors trained the OPR network with 5,718 private real old photos. Meanwhile, our CHD training set only contains 514 old photos. Another hypothesis of the training failure is that our collection of

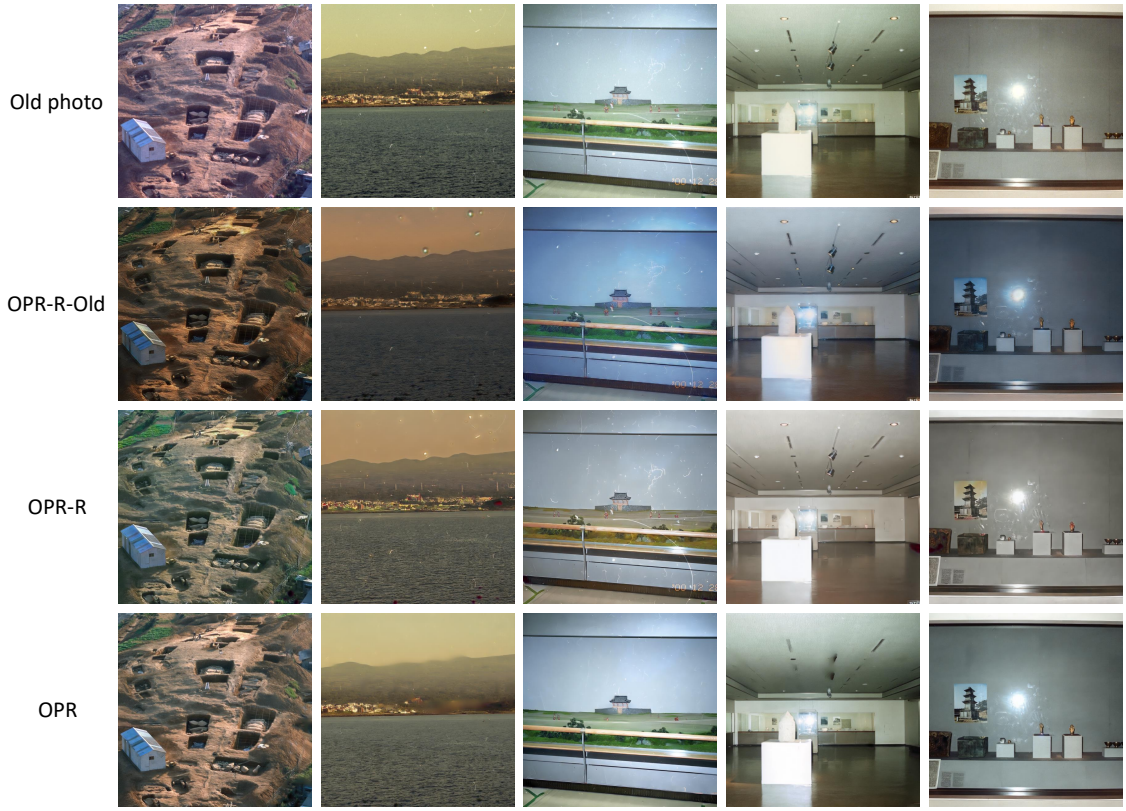

Figure 22. The results of retraining OPR. OPR [44] denotes the original pretrained model. OPR-R-Old denotes the model retrained using the original synthetic training data and our CHD training set, while OPR-R denotes the model retrained using our proposed synthetic data and our CHD training set.

old photos has a larger diversity compared to the portrait photos used to train the original OPR network [44]. In addition, since the baseline OPR is not capable of handling diverse color jittering degradation, the results of the retrained baseline OPR using our synthetic training data and CHD training set (OPR-R) are inferior to those of the pretrained baseline OPR model [44] in real old photos evaluation.

Comparison between a sequence of 'stylization + enhancement', 'enhancement + stylization', and only 'stylization'. We show additional results of performing old photo modernization using three different variations: 1) 'stylization + enhancement', 2) 'enhancement + stylization', and 3) 'stylization'. Specifically, we provide additional quantitative results on a synthetic dataset and real old photos, and qualitative results on real old photos to show that 'stylization + enhancement' is the best baseline over other variations. Table 5 shows the quantitative results of old photo modernization on the synthetic dataset, where on average, the 'stylization + enhancement' baselines achieve better results than other baselines' variations. Even though 'ExColTran' [52] achieves higher PSNR and SSIM than other baselines, we still choose 'stylization + enhancement' as the main baselines since this sequence provides the most stable results for all of the baselines. In addition, Table. 5 also shows that pre-trained OPR [44] is only better for real old photo evaluation, while worse for synthetic data evaluation. Compared to retrained OPR (OPR-R), using the pre-trained OPR (OPR) decreases PSNR, SSIM, and increases LPIPS by an average of 0.980, 0.079, and 0.060 respectively for all stylization baselines.

Method	PSNR↑	SSIM↑	LPIPS↓
ExColTran [52]	<u>20.1637</u>	**0.8123**	0.2735
ReHistoGAN [1]	19.8240	<u>0.8044</u>	0.2467
MAST [16]	17.5653	0.7685	0.2726
PCAPST [7]	17.3873	0.7834	0.2671
Average	18.7351	0.7922	0.2650
ExColTran [52] + OPR	18.9152	0.7144	0.3044
ReHistoGAN [1] + OPR	18.9767	0.7220	0.2748
MAST [16] + OPR	18.1063	0.7042	0.2855
PCAPST [7] + OPR	17.8949	0.7061	0.2874
Average	18.4733	0.7117	0.2880
ExColTran [52] + OPR-R	19.5796	0.7885	0.2563
ReHistoGAN [1] + OPR-R	20.0458	0.7987	<u>0.2109</u>
MAST [16] + OPR-R	19.0148	0.7853	0.2270
PCAPST [7] + OPR-R	19.1731	0.7908	0.2197
Average	19.4533	0.7908	0.2285
OPR-R + ExColTran [52]	20.1565	0.7989	0.2400
OPR-R + ReHistoGAN [1]	19.8990	0.7932	0.2115
OPR-R + MAST [16]	17.6050	0.7591	0.2374
OPR-R + PCAPST [7]	17.7387	0.7702	0.2317
Average	18.8498	0.7803	0.2302
Ours	**21.2212**	0.7919	**0.2027**

Table 5. Quantitative results of modernization on synthetic dataset.

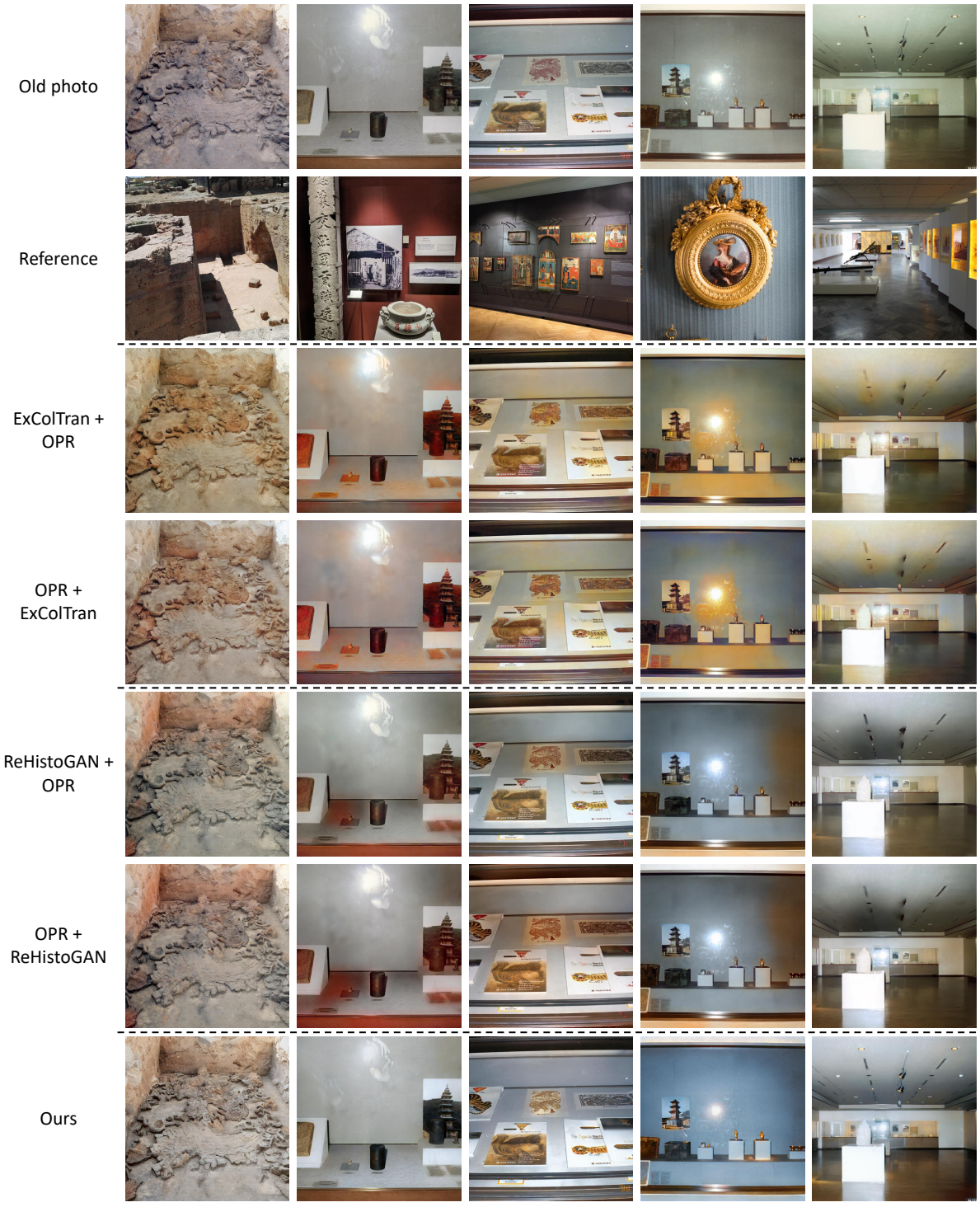

Figure 23. Comparison between 'stylization + enhancement' and 'enhancement + stylization' for ExColTran [52] and ReHistoGAN [1].

The same observation can also be seen in the quantitative results of modernization on real old photos shown in Table. 6. On average, other variations: 'enhancement + stylization' and 'stylization' achieves better (lower) average NIQE [35] scores and worse (higher) BRISQUE [34] scores compared to 'stylization + enhancement'. In our observation, the BRISQUE score is a better metric for real old photo evaluation that better matches the qualitative results of modernization on real old photos. For example, we show the qualitative results of OPR and OPR-R in Fig. 22, where OPR achieves better results. However, the NIQE performance of OPR-R is better than OPR, even though the qualitative results show the opposite. In addition, the qualitative results of

Figure 24. Comparison between 'stylization + enhancement' and 'enhancement + stylization' for MAST [16] and PCAPST [7].

'ReHistoGAN [1] + OPR' are also better than 'OPR + ReHistoGAN [1]' shown in Fig. 23. All in all, we show the qualitative results of both 'stylization + enhancement' and 'enhancement + stylization' for every baseline in Fig. 23 and Fig. 24, where the results show that 'stylization + enhancement' achieves better results than the 'enhancement + stylization'. In addition, we also show the qualitative results of 'stylization + enhancement' and only 'stylization' for every baseline in Fig. 25. The results show that using

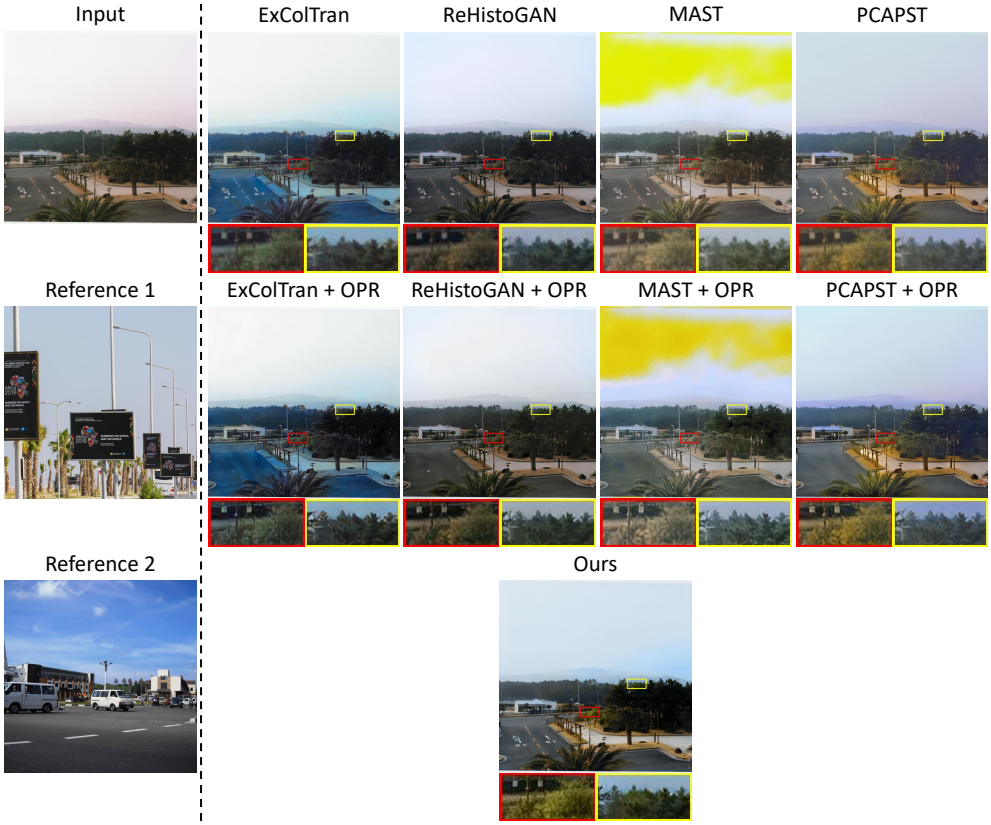

Figure 25. Comparison between only 'stylization' and 'stylization + enhancement' for all of the stylization baselines: ExColTran [52], ReHistoGAN [1], MAST [16], and PCAPST [7]. OPR [44] is used for the enhancement method.

Method	NIQE↓	BRISQUE↓
OPR [44]	4.8705	21.4588
OPR-R	3.8616	25.2025
ExColTran [52]	3.3852	28.5359
ReHistoGAN [1]	**3.2115**	32.4907
MAST [16]	3.4060	26.6633
PCAPST [7]	<u>3.2264</u>	24.8812
Average	3.3073	28.1428
ExColTran [52] + OPR	4.9415	18.8971
ReHistoGAN [1] + OPR	4.8051	26.2557
MAST [16] + OPR	4.8111	18.9555
PCAPST [7] + OPR	4.7094	18.9860
Average	4.8168	20.7736
OPR + ExColTran [52]	5.1461	22.7619
OPR + ReHistoGAN [1]	3.3192	33.7882
OPR + MAST [16]	4.7573	22.5228
OPR + PCAPST [7]	4.7087	22.7718
Average	4.4829	25.4612
Ours - Single	3.4737	<u>15.5152</u>
Ours - Multiple	3.4487	**15.4180**

Table 6. Quantitative results of modernization on real old photos.

enhancement ('stylization + enhancement') improves the stylization output, making it look cleaner and sharper, and have a better color (yellow and red boxes). Thus, choosing 'stylization + enhancement' as the sequence for the baselines is the better choice to provide a fair comparison.

The results of spatial concatenation as the baseline. One naive way to make single-reference baselines able to handle multi-reference is by spatially concatenating multiple references into a single reference. We show the results of spatial concatenation baselines in Fig. 26. The results show that using a single reference for all of the baselines is mostly better compared to using the spatial concatenation of multiple references since the results of concatenation look more unnatural in most cases, e.g., unnatural tree color. This is likely caused by the inability of the baselines to perform local style/color transfer properly.

13. Additional Ablation Studies

Ablation study on loss functions for single stylization subnet. Fig. 27 shows the visual results of the ablation study on loss functions for the single stylization subnet. Training the subnet with only $\mathcal{L}_{ML}$ is insufficient, making the subnet produce severe artifacts far from photorealistic results. Meanwhile, adding $\mathcal{L}_p$ can reduce the artifact and enable the subnet to achieve faithful stylization at the semantic level, e.g., the wall and the painting, but the results still have some style artifacts. Changing $\mathcal{L}_p$ to $\mathcal{L}_{CX}$ can produce better semantic style transfer with fewer artifacts. However, it produces weird artifacts, e.g., black dots in the wall region of the second row in Fig. 27. By applying all three losses $\mathcal{L}_{ML}$, $\mathcal{L}_p$, and $\mathcal{L}_{CX}$ to train the subnet, we achieve the best photorealistic style transfer results that can faithfully stylize the old photos both

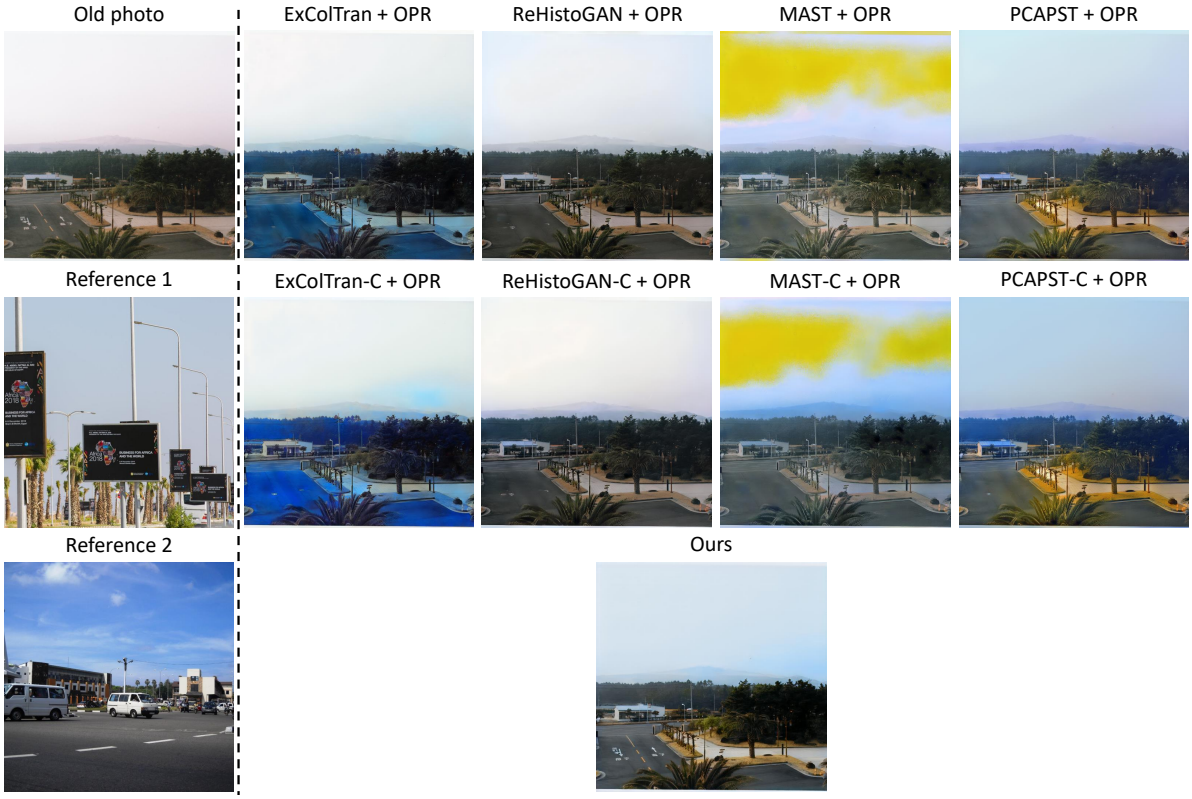

Figure 26. The results of multi-reference stylization using spatial concatenation and single-reference stylization. Baseline, e.g., ReHistoGAN denotes the result of performing single-reference stylization using reference 1. Meanwhile, Baseline-C, e.g., ReHistoGAN-C denotes the result of performing multi-reference stylization by spatially concatenating references 1 and 2 into a single reference.

Figure 27. Ablation study on loss functions for single stylization subnet.

on pixel and semantic levels, and can perform local style transfer without any semantic segmentation mask.

Ablation study on loss functions for merging-refinement subnet. Fig. 28 shows the visual results of the ablation study on loss functions for the merging-refinement subnet. Training the subnet with only reconstruction loss $\mathcal{L}_{L1}$ can make the subnet produce accurate merging and better refinement. However, it produces several artifacts, e.g., rough textures around the wall regions. Even though adding the local smoothness loss $\mathcal{L}_{sm}$ can produce spatially smooth output, it still contains some artifacts, e.g., the bluish color around the painting frame in the second row of Fig. 28. All of the artifacts can be removed by additionally adding a perceptual loss $\mathcal{L}_p$, but it has dull and unattractive (unsaturated) colors and blurry texture. Adding a GAN loss $\mathcal{L}_{adv}$ to the loss function further makes the modernization results more realistic so that the texture becomes sharper and the saturation of color increases, making the output look more like modern images.

Exploration study for the merging-refinement subnet. In this

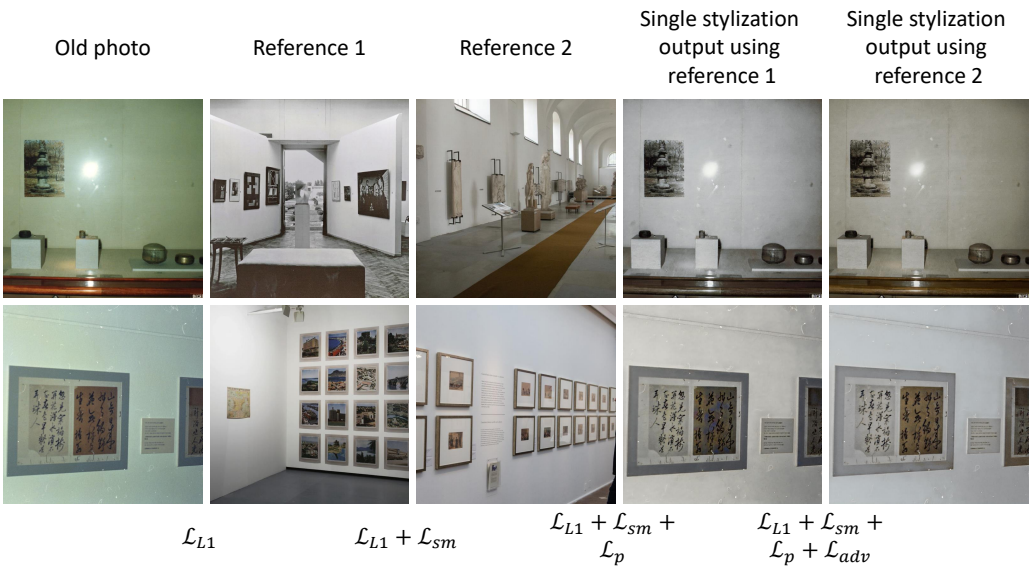

Figure 28. Ablation study on loss functions for merging-refinement subnet.

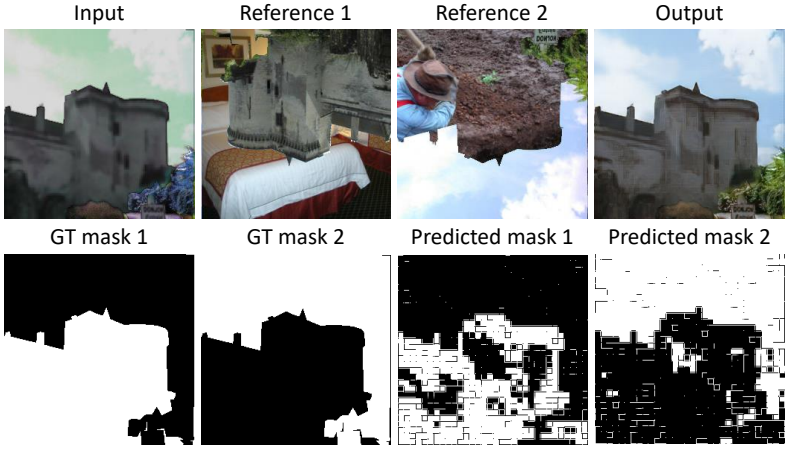

Figure 29. Study on the capability of merging-refinement subnet $\mathcal{M}$ to select relevant regions from multiple references to transfer their styles to the corresponding regions in the input.

study, we explore the capability of the merging-refinement subnet $\mathcal{M}$ in selecting relevant regions from multiple references to transfer their styles to the corresponding regions in the input. To evaluate this capability, we use a synthetic sample generated using our synthetic data generation pipeline, where we can get the ground truth segmentation mask. Since our merging-refinement subnet uses spatial attention, we can generate the prediction mask by simple thresholding of the attention weight. This prediction mask denotes the regions in the input where the corresponding style from multiple references will be transferred to. Furthermore,

we can use the mIoU (mean intersection over union) between the predicted masks and ground truth masks to measure the accuracy of the merging-refinement subnet $\mathcal{M}$. As shown in Fig. 29, our $\mathcal{M}$ can select relevant regions in the input where it achieves an average of 70.70% mIoU for both predictions.

14. Study on the Method's Capability

Nature and capability of the enhancement in this work. In this work, the enhancement primarily focuses on unstructured degradation (UD) restoration such as deblurring, denoising, and artifact removal, commonly found in old photos. The capability of our enhancement can be seen in Fig. 25, where the output of our method is sharper and less noisy compared to the baselines denoting better enhancement capability. Despite primarily focusing on UD, we find that our method can still generalize to some extent to structured degradation (SD). We provide additional results on an old photo from RealOld dataset [50] with severe SD and UD in Fig. 30 to better show the enhancement capability of our method. All the stylization baselines coupled with OPR can restore SD (scratches and holes) better than ours (red boxes) since it is explicitly trained for such degradations, even though the baselines also fail to remove all SDs like ours. Nevertheless, our method can enhance the image by restoring small scratches and UD (blur and noise) better than the baselines without excessive blurry artifacts like the results of MAST + OPR (yellow boxes).

Stylization on modern images. We provide stylization results on modern photos which have no degradations using our network (MROPM-Net) and other stylization baselines. As shown in Fig. 31, our MROPM-Net (denoted as Our – Full) achieves the best local style transfer on all images. In addition, our PST network (without style code predictor and merging-refinement subnet) achieves faithful stylization as shown in the first and second examples of Fig. 31 and is on par with the SOTA PST network (PCAPST [7]) results.

Modernization results using unrelated references. Fig. 32 shows the visual examples of the robustness of our method when the references are highly unrelated. Our method outperforms other baselines in terms of handling unrelated references. We further show the internal working of our MROPM-Net when handling one of the unrelated references in Fig. 33. In this example, our single stylization subnet can robustly find a better style that can modernize the specific regions in the old photos, e.g., the style of concrete to stylize the wall region instead of the red wall in the first reference. In addition, the merging-refinement subnet can further select the first reference style to stylize the wall region compared to the yellowish wall style in the second reference.

Modernization results using more than two references. We provide additional results when using more than two references in Fig. 34, Fig. 35, Fig. 36, and Fig. 37. As shown in all of the figures, our MROPM-Net can adaptively select appropriate styles from multiple references to further improve modernization performance. Some results show distinctive improvement in specific regions shown inside yellow dashed boxes. In some other results, the overall improvement of the old photos can also be seen outside the yellow dashed boxes. Users can choose which region is important and accordingly choose references that can improve the specific regions depending on the availability of similar objects in references. Since using more than four references with the resolution of 1024×1024 could not be processed with our GPU (NVIDIA RTX 3090), we resize the images (old photo and references) into the resolution of 512×512 to handle more than two references.

Some examples of user study results. We provide some examples of user study results with varying user voting percentages. The results are shown in Fig. 38, Fig. 39, Fig. 40, and Fig. 41. In most cases, the results produced by our method are more preferably selected by the users compared to other baselines.

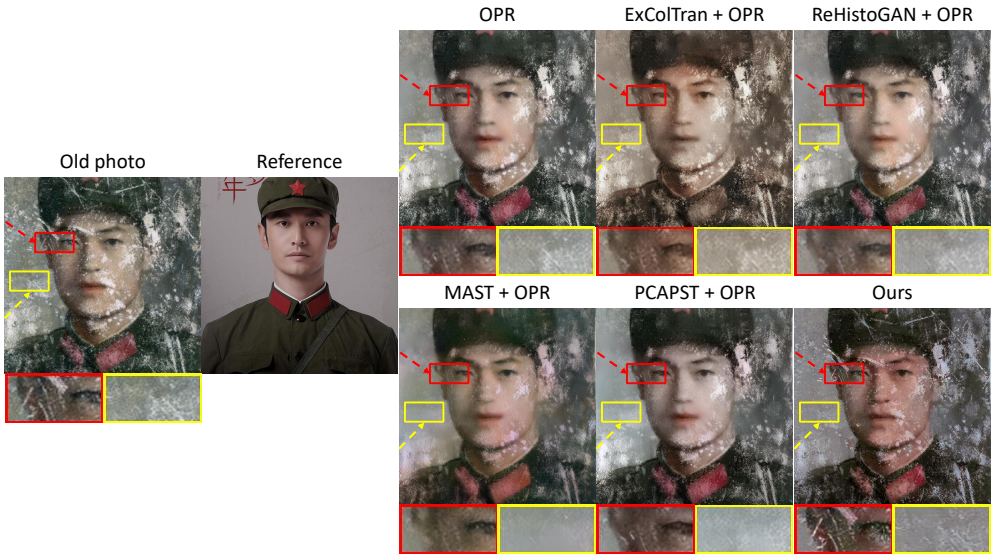

Figure 30. Results of our method compared to other baselines on an old photo from RealOld dataset [50] with severe structured and unstructured degradations.

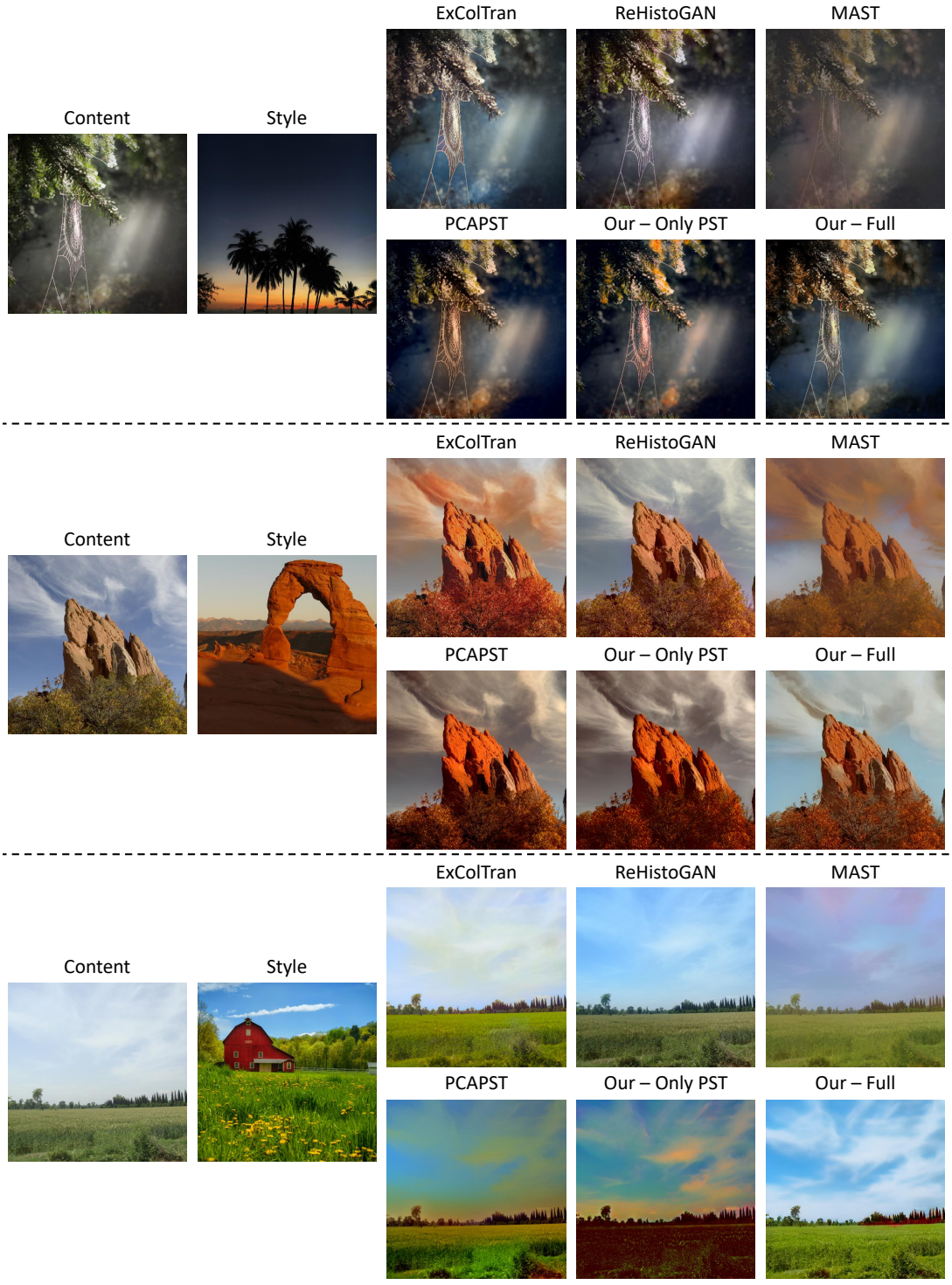

Figure 31. Comparison of stylization results on modern images. Our MROPM-Net, denoted as Our – Full, achieves the best local PST results on all examples compared to other PST baselines such as MAST [16] and PCAPST [7], and other baselines such as ExColTran [52] and ReHistoGAN [1]. Our – Only PST denotes the results of PST using our PST network (without style code predictor and merging-refinement subnet).

Figure 32. Old photo modernization results using unrelated references. In most cases, our method outperforms baselines (OPR [44], ExColTran [52] + OPR, ReHistoGAN [1] + OPR, MAST [16] + OPR, and PACPST [7] + OPR) even though the references are unrelated with the old photo. Reference-based baselines use reference 1 as their reference.

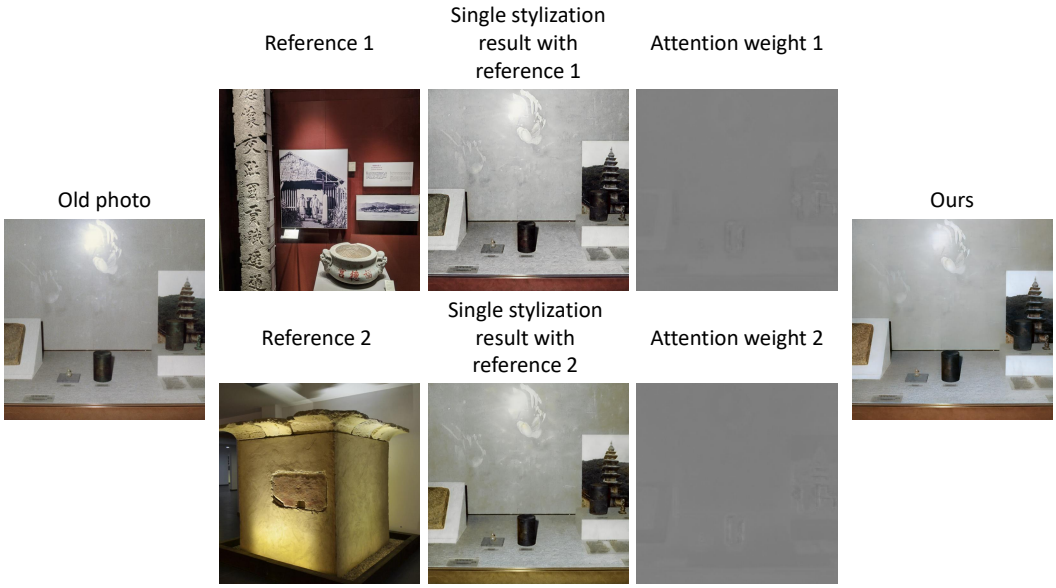

Figure 33. The internal working of our MROPM-Net when handling unrelated references.

Figure 34. Progressive old photo modernization results using three references. Some regions with distinctive improvements are shown inside yellow boxes.

Figure 35. Progressive old photo modernization results using three references. Some regions with distinctive improvements are shown inside yellow boxes.

Figure 36. Progressive old photo modernization results using four references. Some regions with distinctive improvements are shown inside yellow boxes.

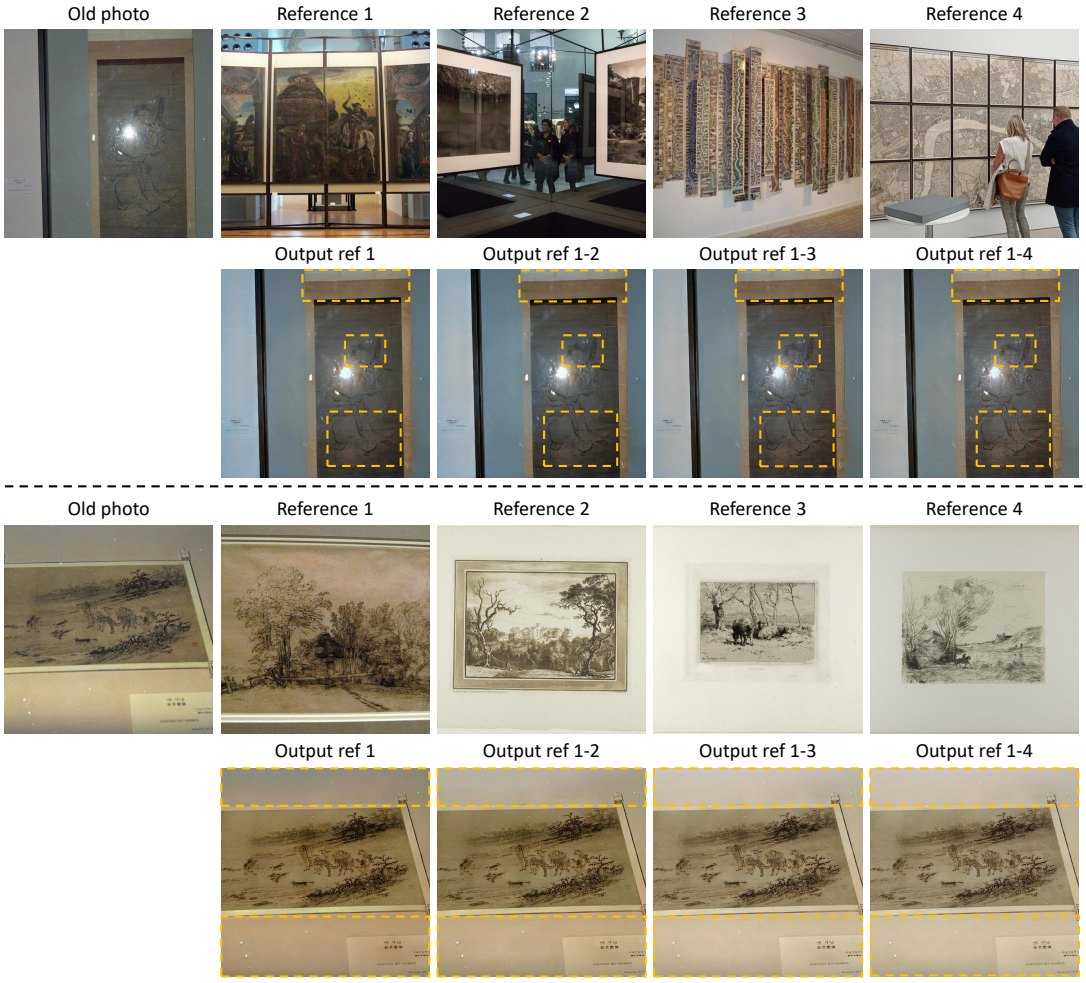

Figure 37. Progressive old photo modernization results using four references. Some regions with distinctive improvements are shown inside yellow boxes.

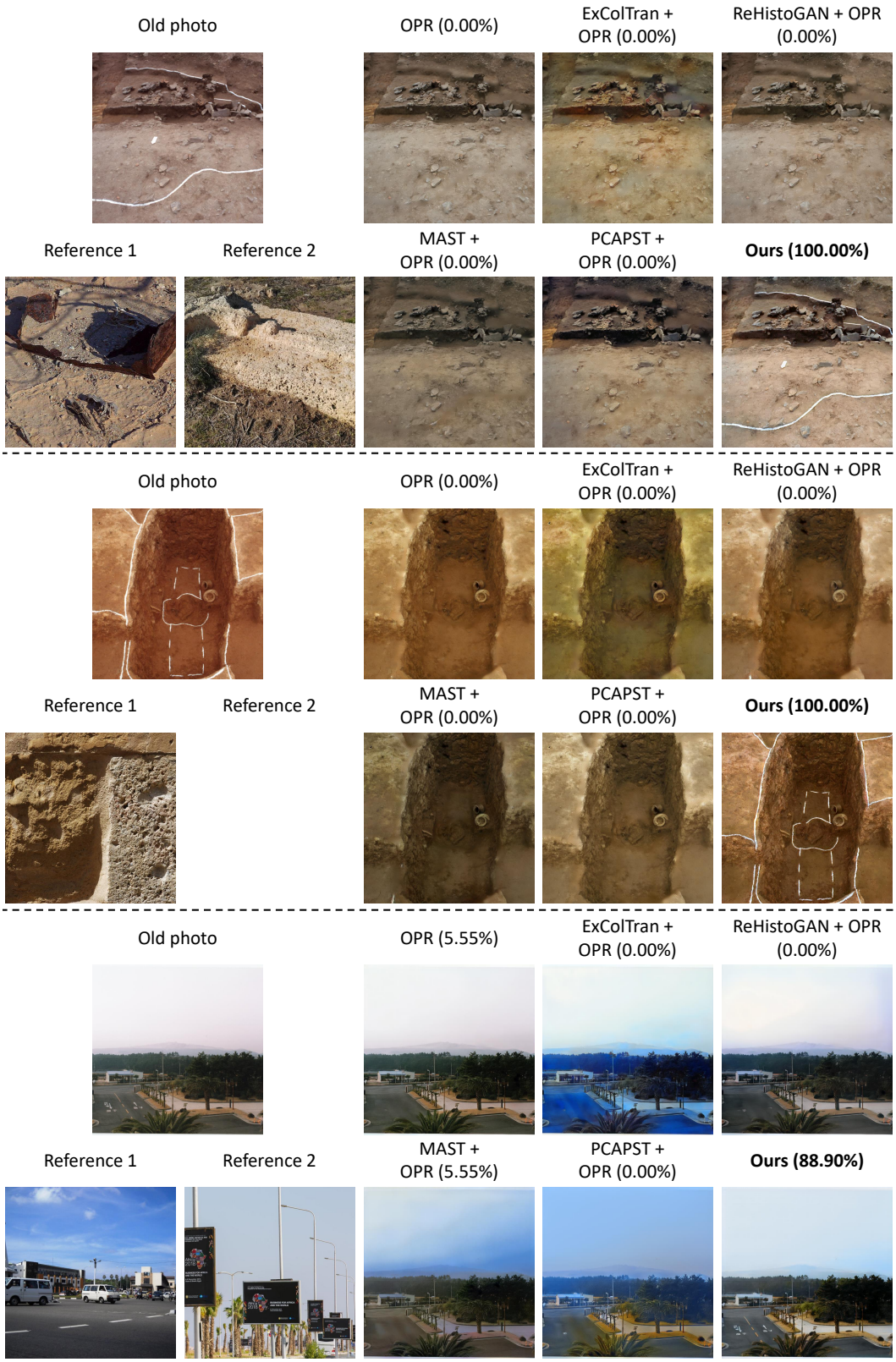

Figure 38. User study results with the percentage of user voting. Our method compares favorably against baselines (OPR [44], ExColTran [52] + OPR, ReHistoGAN [1] + OPR, MAST [16] + OPR, and PCAPST [7] + OPR). Reference-based baselines use reference 1 as their reference.

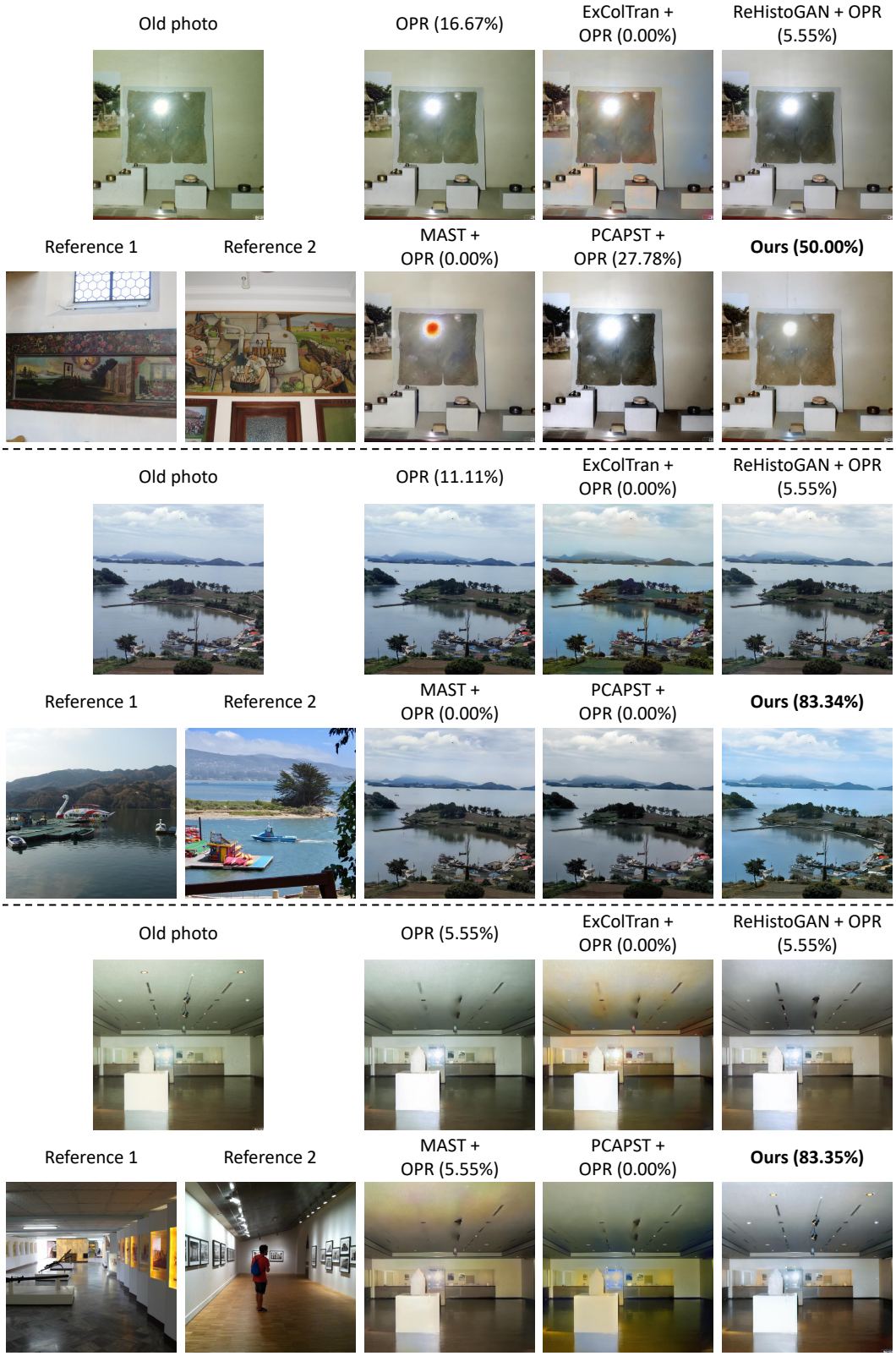

Figure 39. User study results with the percentage of user voting. Our method compares favorably against baselines (OPR [44], ExColTran [52] + OPR, ReHistoGAN [1] + OPR, MAST [16] + OPR, and PCAPST [7] + OPR). Reference-based baselines use reference 1 as their reference.

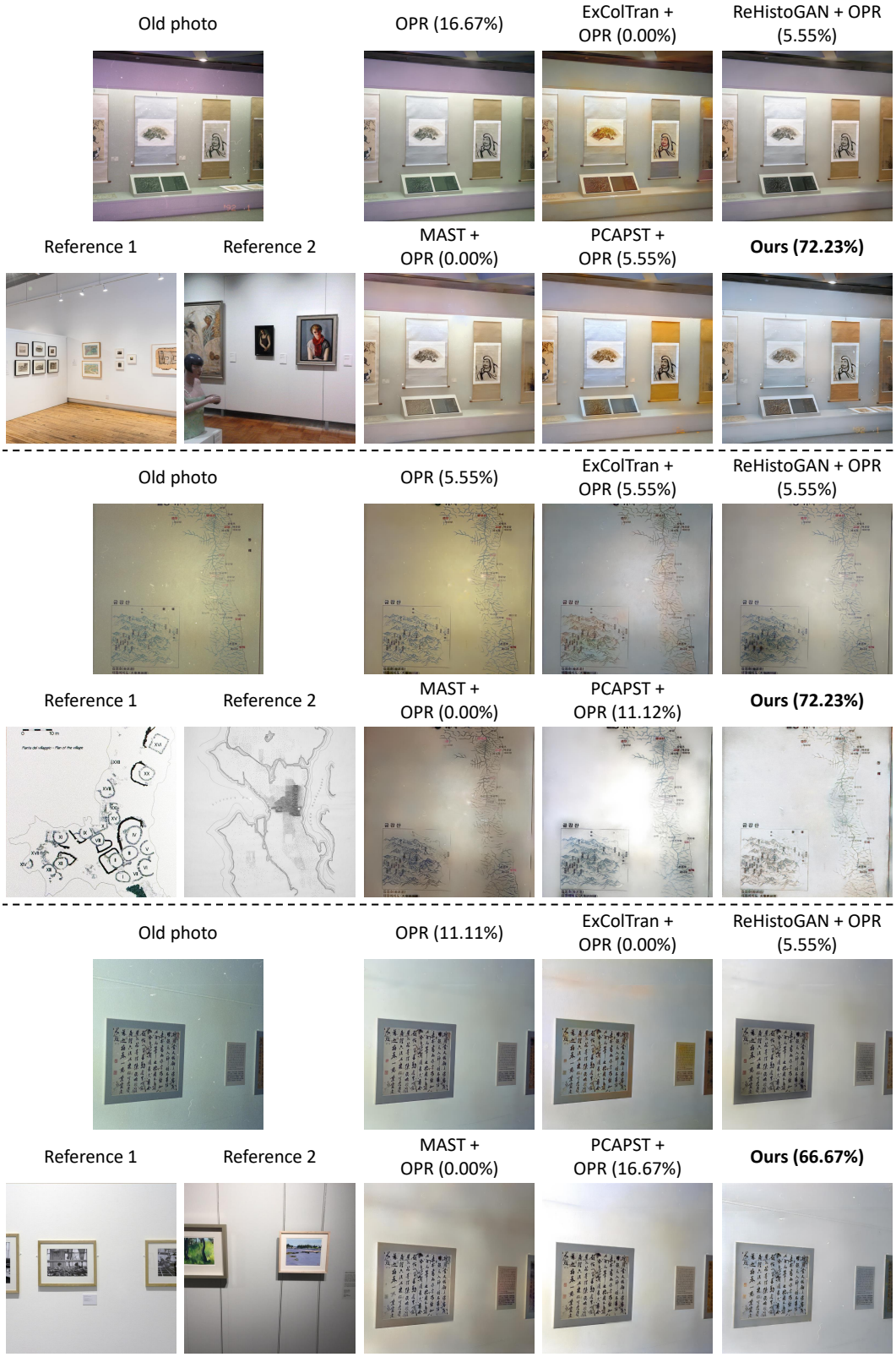

Figure 40. User study results with the percentage of user voting. Our method compares favorably against baselines (OPR [44], ExColTran [52] + OPR, ReHistoGAN [1] + OPR, MAST [16] + OPR, and PCAPST [7] + OPR). Reference-based baselines use reference 1 as their reference.

Figure 41. User study results with the percentage of user voting. Our method compares favorably against baselines (OPR [44], ExColTran [52] + OPR, ReHistoGAN [1] + OPR, MAST [16] + OPR, and PCAPST [7] + OPR). Reference-based baselines use reference 1 as their reference.