Modernizing Old Photos Using Multiple References via Photorealistic Style Transfer – Supplementary Material –

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

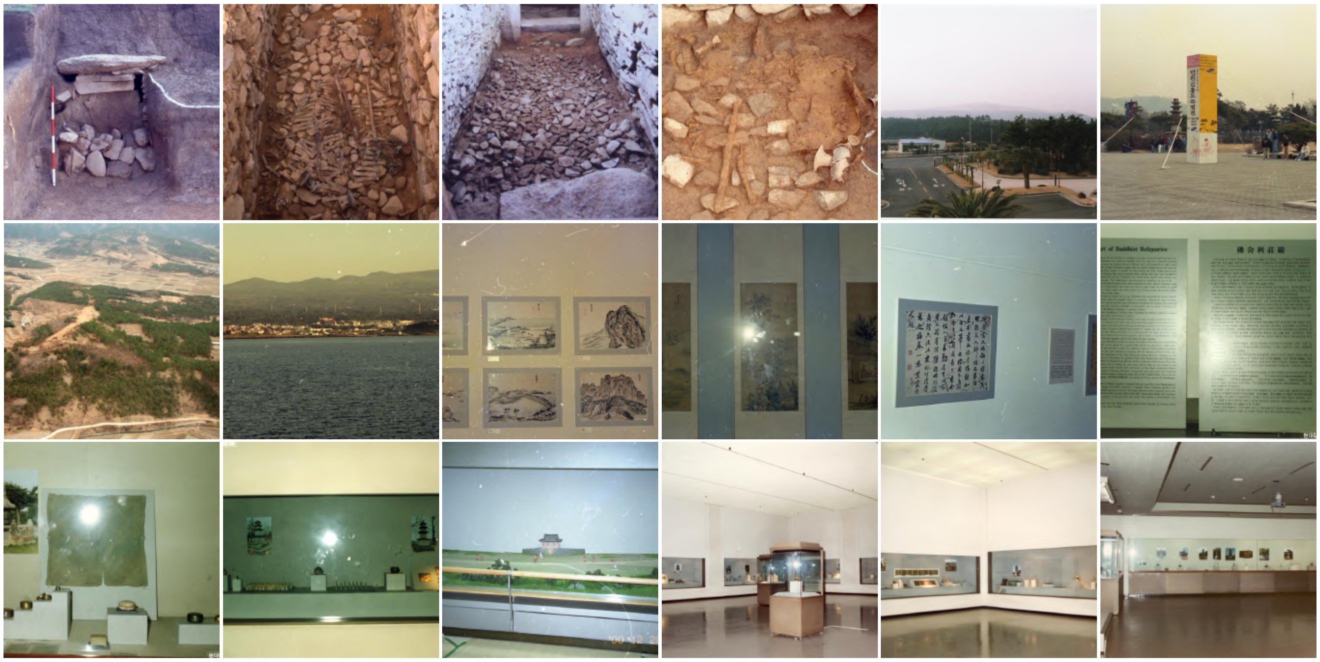


Figure 1. Diverse examples of outdoor and indoor scenes from our dataset. The images contain various kinds of degradations, especially color fading.

1. Details of Proposed CHD Dataset

1.1. 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. 1 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.

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,

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

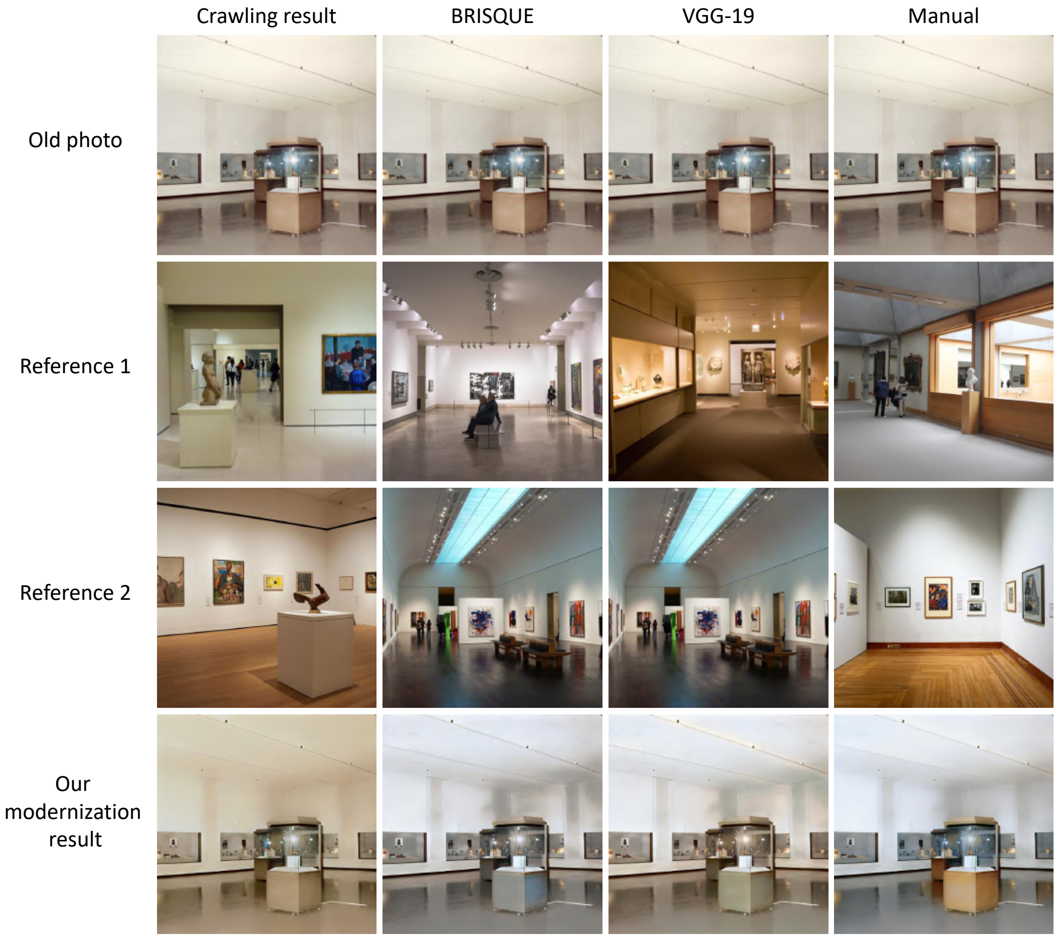


Figure 2. Automatic references selection trial. We try to automate reference selection by using BRISQUE [13] and VGG-19 [16] feature similarity.

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 [16] feature space, using a similar idea to [6, 21], and the best BRISQUE [13] score. However, Fig. 2 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

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

	HWFD [11]	RealOld [21]	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 1. Comparison between our dataset and other public old photos datasets in several factors.

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.

1.2. Comparison of CHD and Other Old Photos Dataset

There are two other public old photo datasets, such as the Historical Wiki Face Dataset (HWFD) [11] and RealOld [21]. However, when this paper was submitted, RealOld [21] had not been published yet. Table. 1 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 [12], which have specific degradations such as color fading (shown in Fig. 1) 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. 1. 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. 3 shows some visual examples of our and other datasets.

2. Results on Real Old Photos in The Wild

Fig. 4 and Fig. 5 show the generalization and robustness of our method when applied to real old photos in the wild. The first and second examples of Fig. 4 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 [22] + OPR). For the third example in Fig. 4, our method achieves the second-best performance compared to ‘PCAPST [5] + 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. 5 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 [22] + OPR). Interestingly, we can handle paper blotches in the second example of Fig. 5 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. 5 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.

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

3.1. Comparison Between Our Photorealistic Style Transfer (PST) Network and WCT2 [23]

Fig. 6 shows the comparison between our PST network and WCT2 [23] 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 [3]. Meanwhile, WCT2 [23] 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. 9). 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) [7] instead of the non-differentiable whitening-and-coloring transformation (WCT) [10] 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 [23] are explained in the following subsection.

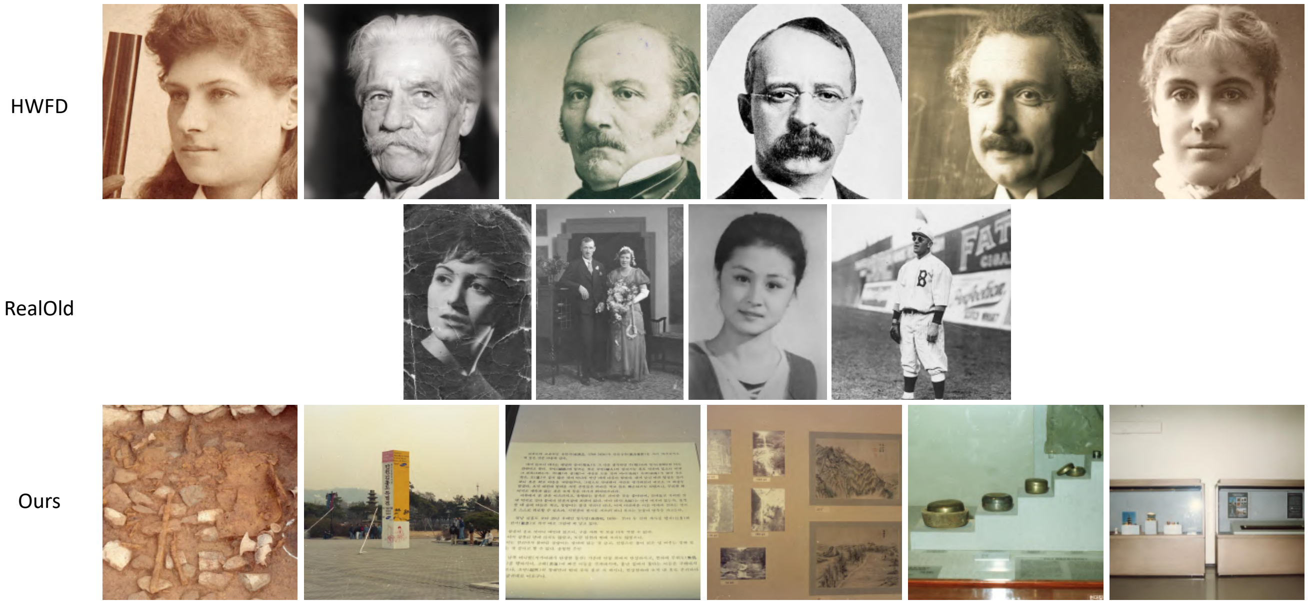


Figure 3. Comparison between our CHD dataset and other datasets (HWFD [11] and RealOld [21]). Our CHD dataset has the most complex and diverse scenes compared to other datasets. In addition, our dataset also contains unique color fading artifacts.

3.2. Additional Analyses

Limitations of WCT2 [23] for old photo modernization. There are two main limitations of WCT2 [23] that can prevent its application on real-world old photos. The first limitation is the unnatural global style transfer results shown in Fig. 7, where this unnatural result may make the photo look like an old photo instead of modernizing them. Meanwhile, Fig. 8 shows the second limitation of WCT2. We generate the semantic segmentation masks using VIT-Adapter [4], which is one of the state-of-the-art models in the semantic segmentation task for the examples in Fig. 8. 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. Therefore, 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. 9 shows that applying progressive stylization in different decoder blocks does not affect the original concatenated version of the WCT2 [23] 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. 9, 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. 10. In addition, AdaIN feature transformation is also differentiable, which 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.

4. Details of MROPM-Net Architecture

4.1. Single Stylization Subnet

The detailed architecture of our single stylization subnet $\mathcal{S}$ can be seen in Fig. 11. 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 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 atten-

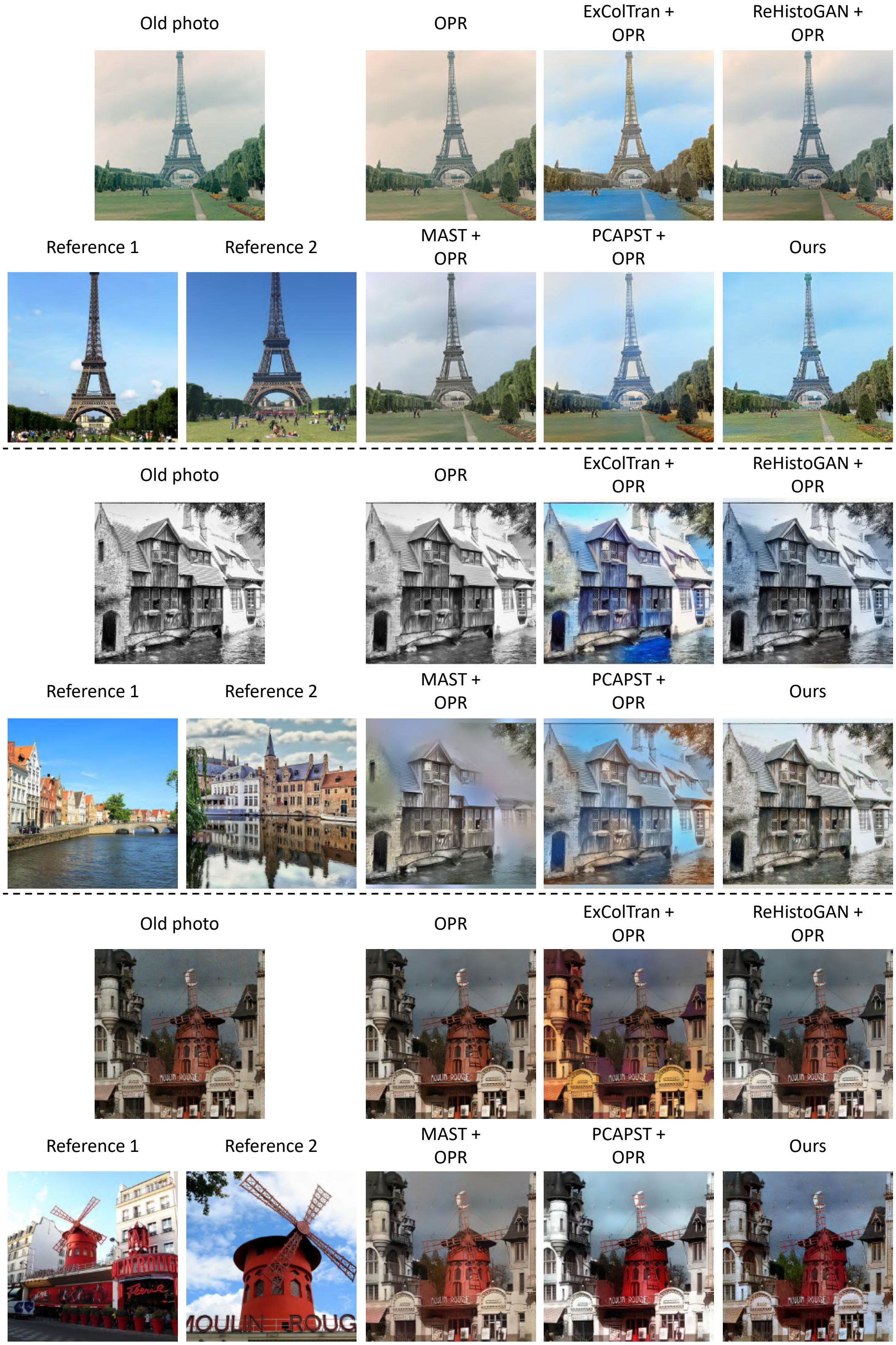
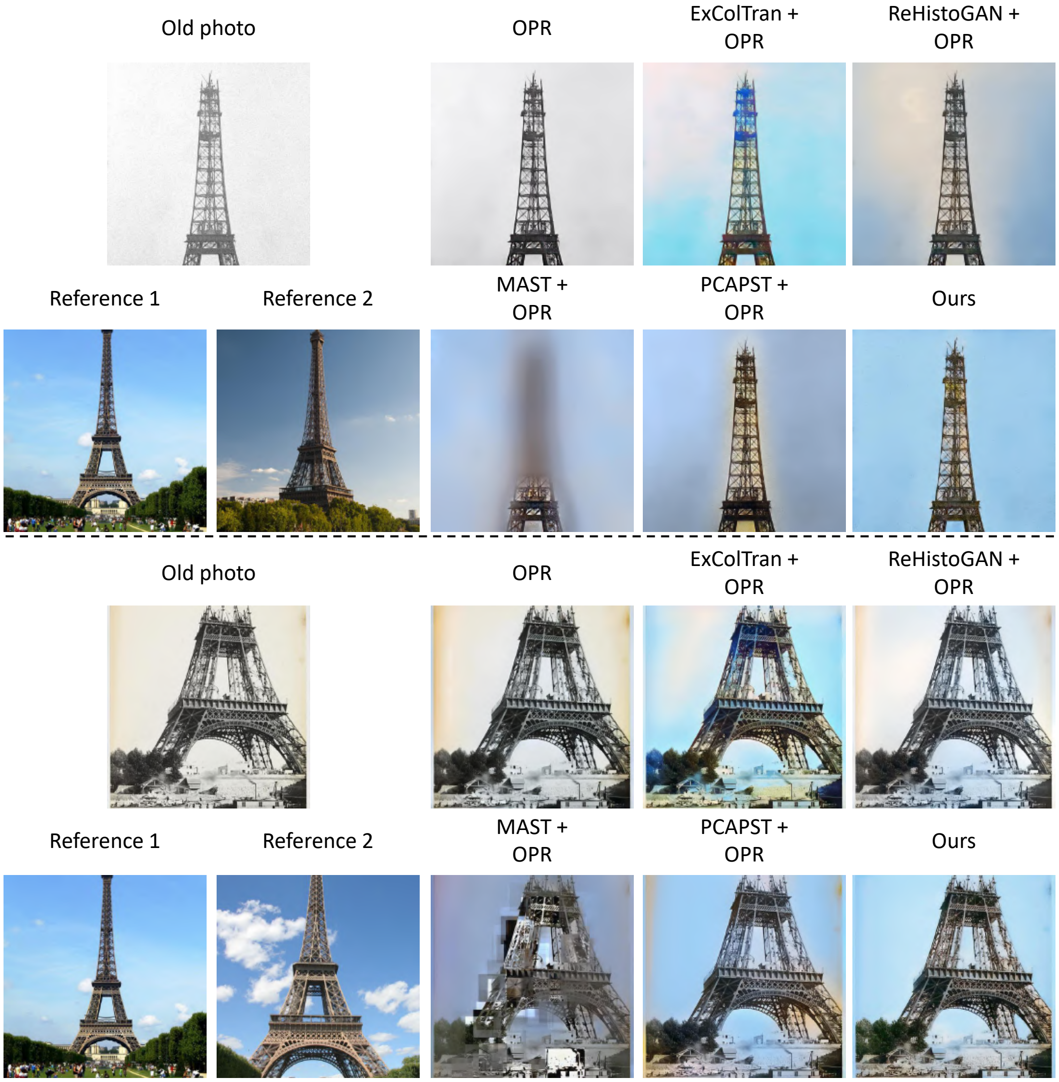


Figure 4. Comparison of modernization on old photos in the wild between our method and other baselines. In most cases, our method outperforms other methods (OPR [18], ExColTran [22] + OPR, ReHistoGAN [1] + OPR, MAST [8] + OPR, and PCAPST [5] + OPR) in modernizing old photos in the wild showing the robustness of our method. Other reference-based baselines use reference 1 as their reference.



Old photo

OPR

ExColTran + OPR

ReHistoGAN + OPR

Reference 1

Reference 2

MAST + OPR

PCAPST + OPR

Ours

Old photo

OPR

ExColTran + OPR

ReHistoGAN + OPR

Reference 1

Reference 2

MAST + OPR

PCAPST + OPR

Ours

Figure 5. Comparison of modernization on greyscale old photos in the wild between our method and other baselines. Our method outperforms other methods (OPR [18], ExColTran [22] + OPR, ReHistoGAN [1] + OPR, MAST [8] + OPR, and PCAPST [5] + 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).

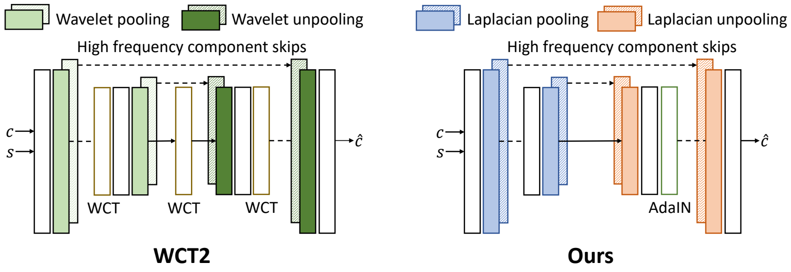


Figure 6. Comparison between our PST network and WCT2 [23].

tion [19]. 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 multi-

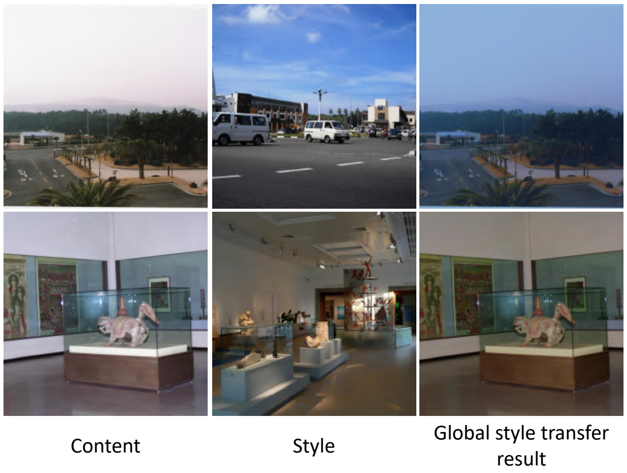


Figure 7. Unnatural global style transfer result of WCT2 [23].

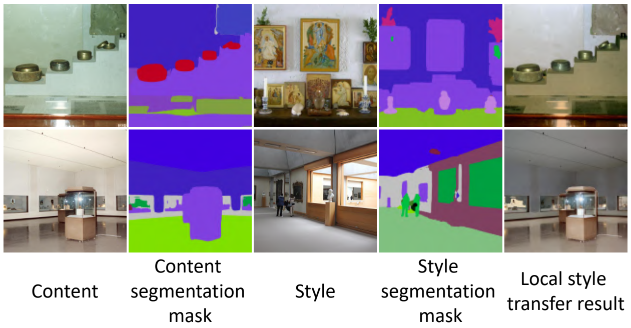


Figure 8. Unnatural local style transfer result of WCT2 [23].

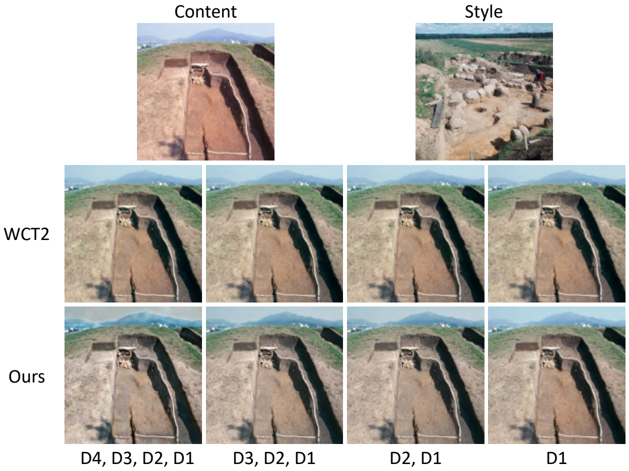


Figure 9. Applying feature transformation in different decoder blocks of WCT2 [23] and our PST network. D denotes the decoder block, while the number denotes the decoder block number (a higher number denotes the decoder block in deeper feature space).

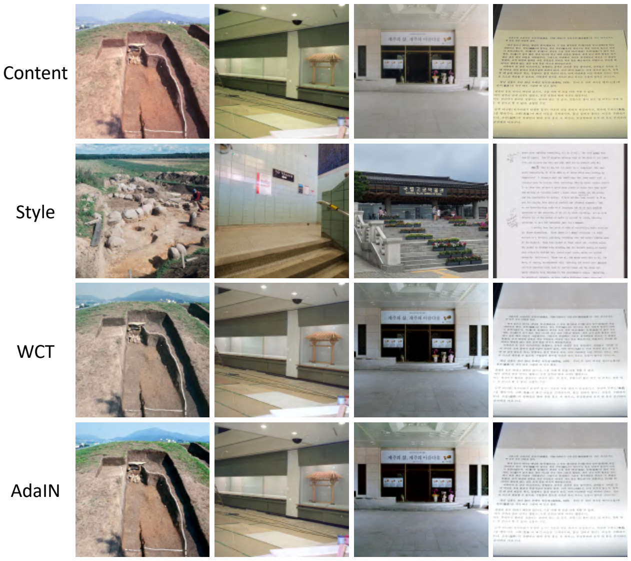


Figure 10. Visual results comparison between AdaIN [7] and WCT [10] feature transformations.

plication, 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.

4.2. Merging-Refinement Subnet

Fig. 12 shows the detailed architecture of our merging-refinement subnet $\mathcal{M}$. We show the details of the spatial attention module [20]. 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 [15] architecture in [9]. 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 [17] in the U-Net architecture.

5. 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

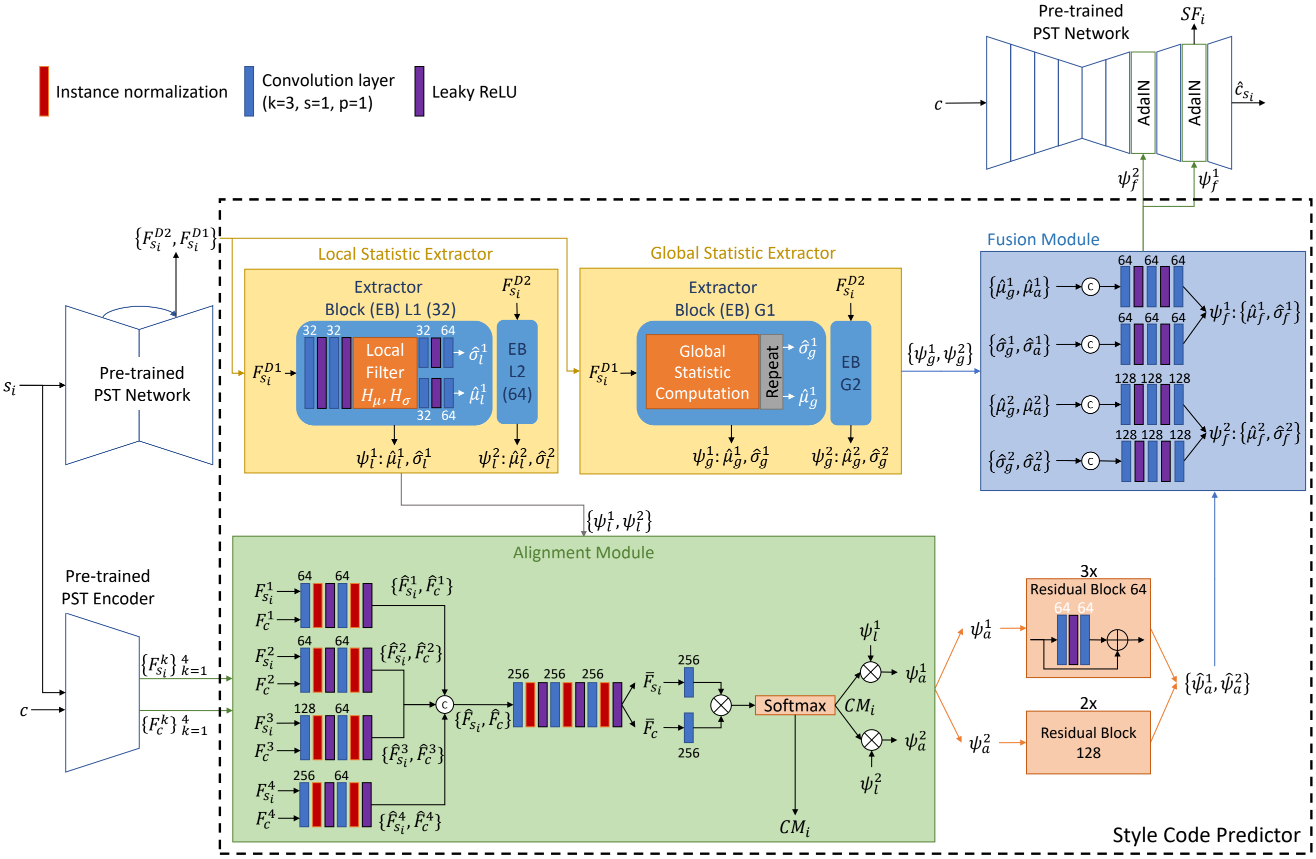

Instance normalization
Convolution layer (k=3, s=1, p=1)
Leaky ReLU
Pre-trained PST Network
SF_i
c
AdaIN
AdaIN
$\hat{c}_{s_i}$
ψ_f^2
ψ_f^1
$\{F_{s_i}^{D2}, F_{s_i}^{D1}\}$
Local Statistic Extractor
Extractor Block (EB) L1 (32)
Local Filter H_μ, H_σ
EB L2 (64)
$\psi_l^1: \hat{\mu}_l^1, \hat{\sigma}_l^1$
$\psi_l^2: \hat{\mu}_l^2, \hat{\sigma}_l^2$
Global Statistic Extractor
Extractor Block (EB) G1
Global Statistic Computation
Repeat
EB G2
$\psi_g^1: \hat{\mu}_g^1, \hat{\sigma}_g^1$
$\psi_g^2: \hat{\mu}_g^2, \hat{\sigma}_g^2$
$\{\psi_g^1, \psi_g^2\}$
Fusion Module
$\psi_f^1: \{\hat{\mu}_f^1, \hat{\sigma}_f^1\}$
$\psi_f^2: \{\hat{\mu}_f^2, \hat{\sigma}_f^2\}$
s_i
Pre-trained PST Network
Pre-trained PST Encoder
$\{F_{s_i}^k\}_{k=1}^4$
$\{F_c^k\}_{k=1}^4$
$\{\psi_l^1, \psi_l^2\}$
Alignment Module
Softmax
CM_i
ψ_a^1
ψ_a^2
3x Residual Block 64
2x Residual Block 128
$\{\hat{\psi}_a^1, \hat{\psi}_a^2\}$
Style Code Predictor

Figure 11. Detailed architecture of our single stylization subnet $\mathcal{S}$.

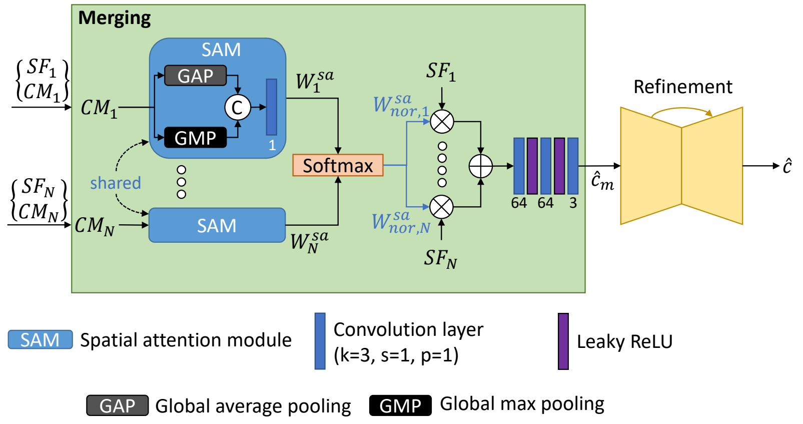


Figure 12. Detailed architecture of our merging-refinement subnet $\mathcal{M}$.

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:

- 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.

6. 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 [18] 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 [22], recolorization [1], and photorealistic style transfer [5, 8]. 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 [22])

- recolorization: color-controlled GAN method (ReHistoGAN [1])

- photorealistic style transfer (PST): semantic PST (MAST [8]) and PCA-based knowledge distillation PST (PCAPST [5])

For the enhancement, we use the SOTA of old photo restoration (OPR [18]) 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 [18]. As mentioned in the main paper, we use the pretrained model of baseline OPR [18] rather than retraining it for real old photo evaluation. The results with the retrained baseline OPR [18] are shown in Fig. 13 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 [18] 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 old photos has a larger diversity compared to the portrait photos used to train the original OPR network [18]. 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 [18] 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 2 shows the quantitative results of old photo modernization on the synthetic dataset, where on

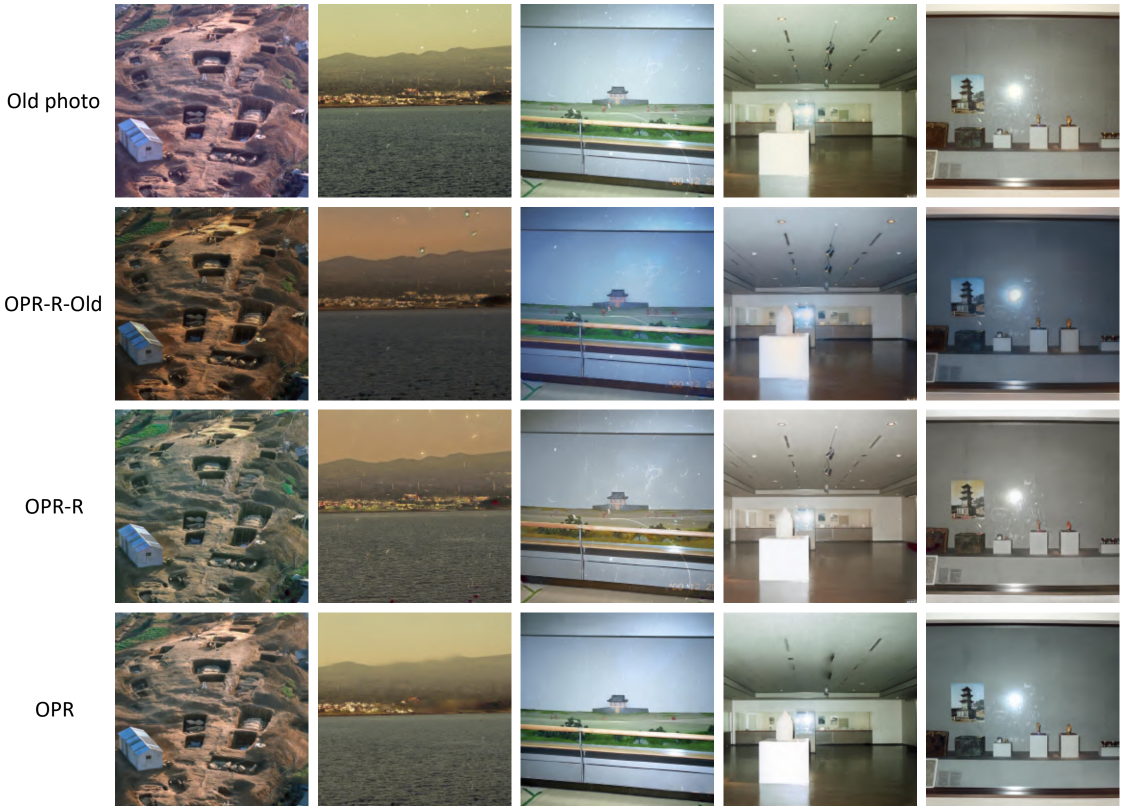


Figure 13. The results of retraining OPR. OPR [18] denotes the original pretrained model. OPR-R-Old denotes the model retrained using the original synthetic training data and our CHD training set, while OPR-R denotes the model retrained using our proposed synthetic data and our CHD training set.

Method	PSNR↑	SSIM↑	LPIPS↓
ExColTran [22]	20.1637	**0.8123**	0.2735
ReHistoGAN [1]	19.8240	0.8044	0.2467
MAST [8]	17.5653	0.7685	0.2726
PCAPST [5]	17.3873	0.7834	0.2671
Average	18.7351	0.7922	0.2650
ExColTran [22] + OPR	18.9152	0.7144	0.3044
ReHistoGAN [1] + OPR	18.9767	0.7220	0.2748
MAST [8] + OPR	18.1063	0.7042	0.2855
PCAPST [5] + OPR	17.8949	0.7061	0.2874
Average	18.4733	0.7117	0.2880
ExColTran [22] + OPR-R	19.5796	0.7885	0.2563
ReHistoGAN [1] + OPR-R	20.0458	0.7987	0.2109
MAST [8] + OPR-R	19.0148	0.7853	0.2270
PCAPST [5] + OPR-R	19.1731	0.7908	0.2197
Average	19.4533	0.7908	0.2285
OPR-R + ExColTran [22]	20.1565	0.7989	0.2400
OPR-R + ReHistoGAN [1]	19.8990	0.7932	0.2115
OPR-R + MAST [8]	17.6050	0.7591	0.2374
OPR-R + PCAPST [5]	17.7387	0.7702	0.2317
Average	18.8498	0.7803	0.2302
Ours	**21.2212**	0.7919	**0.2027**

Table 2. Quantitative results of modernization on synthetic dataset.

Method	NIQE↓	BRISQUE↓
OPR [18]	4.8705	21.4588
OPR-R	3.8616	25.2025
ExColTran [22]	3.3852	28.5359
ReHistoGAN [1]	**3.2115**	32.4907
MAST [8]	3.4060	26.6633
PCAPST [5]	3.2264	24.8812
Average	3.3073	28.1428
ExColTran [22] + OPR	4.9415	18.8971
ReHistoGAN [1] + OPR	4.8051	26.2557
MAST [8] + OPR	4.8111	18.9555
PCAPST [5] + OPR	4.7094	18.9860
Average	4.8168	20.7736
OPR + ExColTran [22]	5.1461	22.7619
OPR + ReHistoGAN [1]	3.3192	33.7882
OPR + MAST [8]	4.7573	22.5228
OPR + PCAPST [5]	4.7087	22.7718
Average	4.4829	25.4612
Ours - Single	3.4737	15.5152
Ours - Multiple	3.4487	**15.4180**

Table 3. Quantitative results of modernization on real old photos.

average, the ‘stylization + enhancement’ baselines achieve better results than other baselines’ variations. Even though ‘ExColTran’ [22] achieves higher PSNR and SSIM than other baselines, we still choose ‘stylization + enhancement’

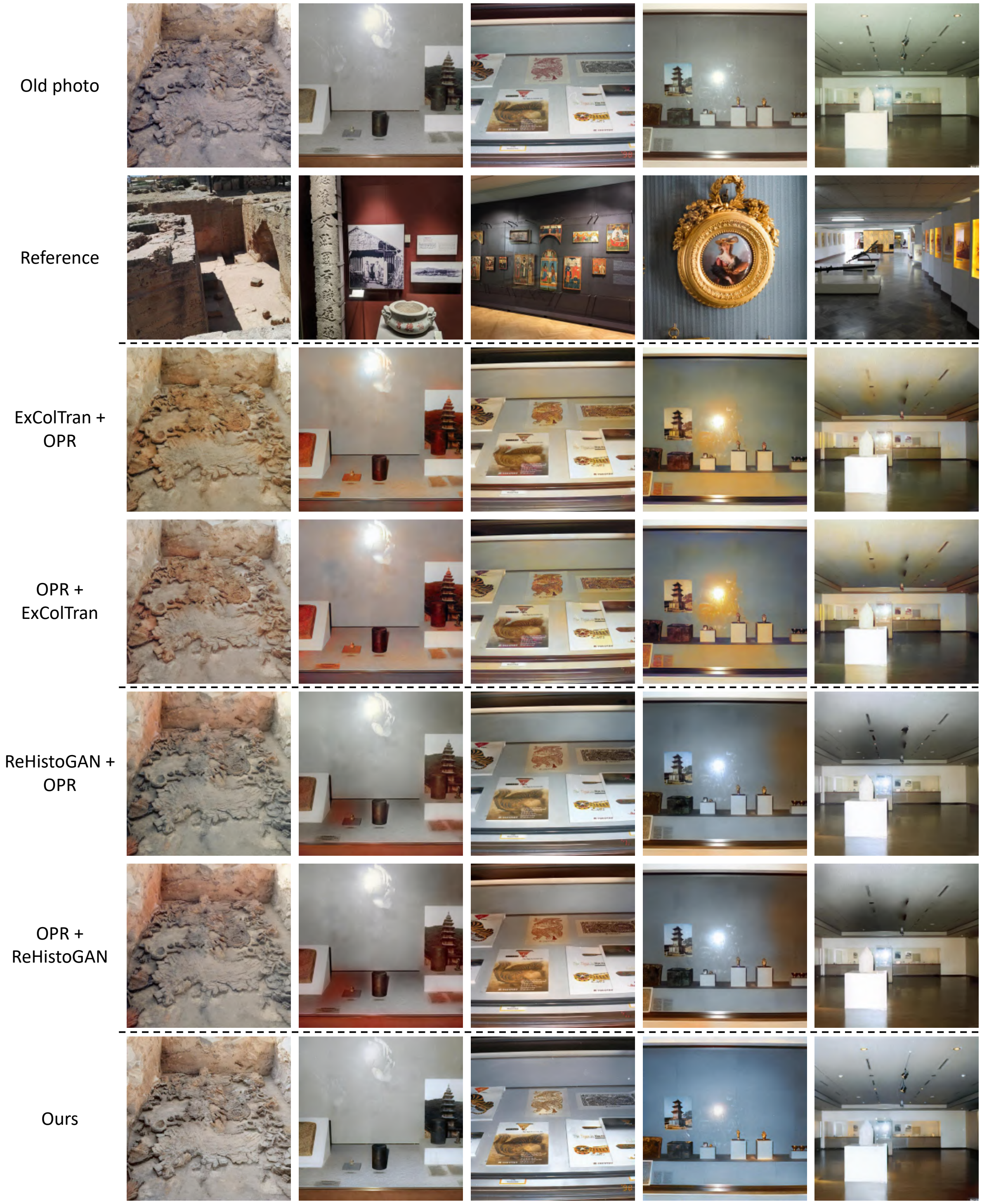


Figure 14. Comparison between 'stylization + enhancement' and 'enhancement + stylization' for ExColTran [22] and ReHistoGAN [1].

as the main baselines since this sequence provides the most stable results for all of the baselines. In addition, Table. 2 also shows that pre-trained OPR [18] 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.

The same observation can also be seen in the quantitative results of modernization on real old photos shown in Table. 3. On average, other variations: 'enhancement + stylization' and 'stylization' achieves better (lower) aver-

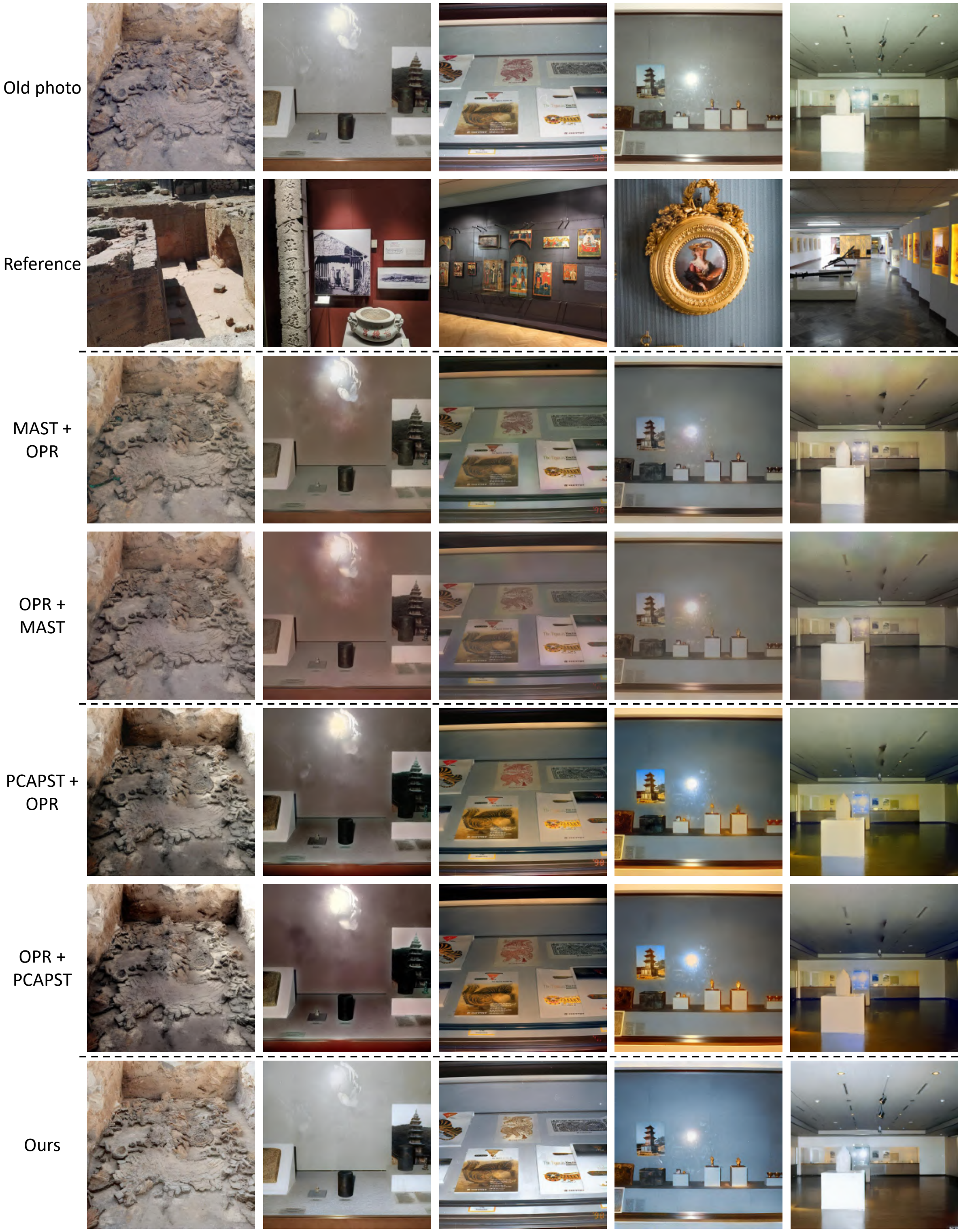


Figure 15. Comparison between 'stylization + enhancement' and 'enhancement + stylization' for MAST [8] and PCAPST [5].

age NIQE [14] scores and worse (higher) BRISQUE [13] 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. 13, where OPR achieves better results. However, the NIQE performance of OPR-R is better than OPR, even though the

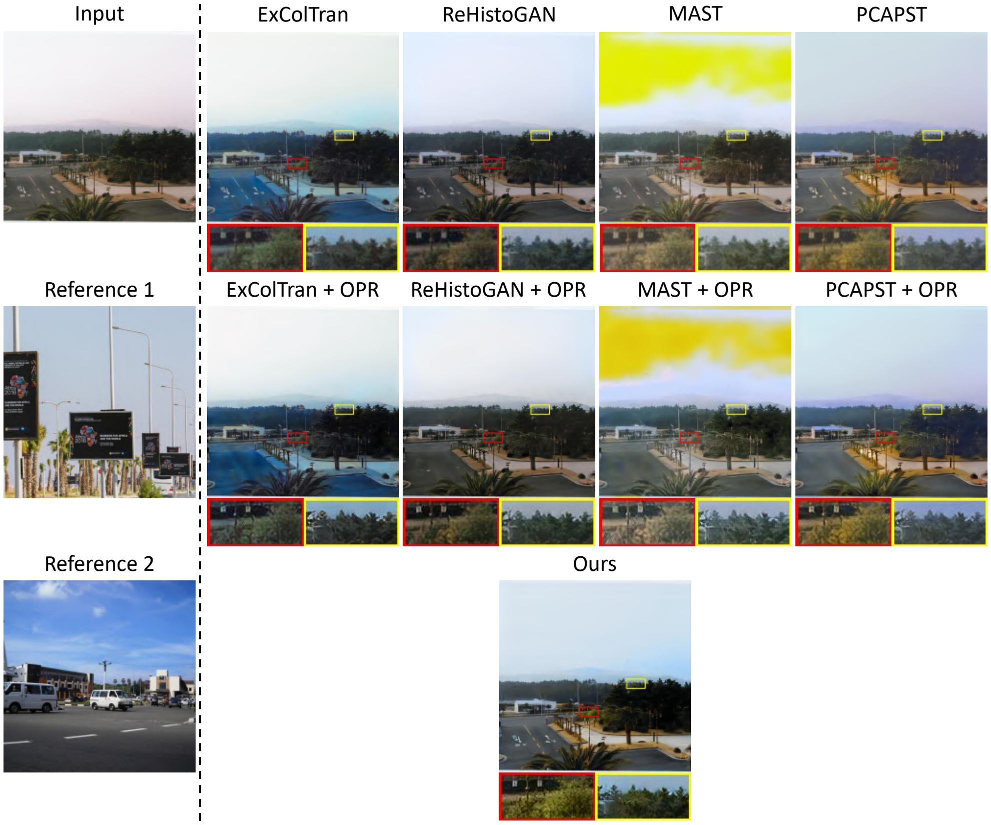


Figure 16. Comparison between only 'stylization' and 'stylization + enhancement' for all of the stylization baselines: ExColTran [22], ReHistoGAN [1], MAST [8], and PCAPST [5]. OPR [18] is used for the enhancement method.

qualitative results show the opposite. In addition, the qualitative results of 'ReHistoGAN [1] + OPR' are also better than 'OPR + ReHistoGAN [1]' shown in Fig. 14. All in all, we show the qualitative results of both 'stylization + enhancement' and 'enhancement + stylization' for every baseline in Fig. 14 and Fig. 15, 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. 16. The results show that using 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. 17. 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.

7. Additional Ablation Studies

Ablation study on loss functions for single stylization subnet. Fig. 18 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. 18. 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 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. 19 shows the visual results of the ablation study on loss functions for the merging-refinement subnet.

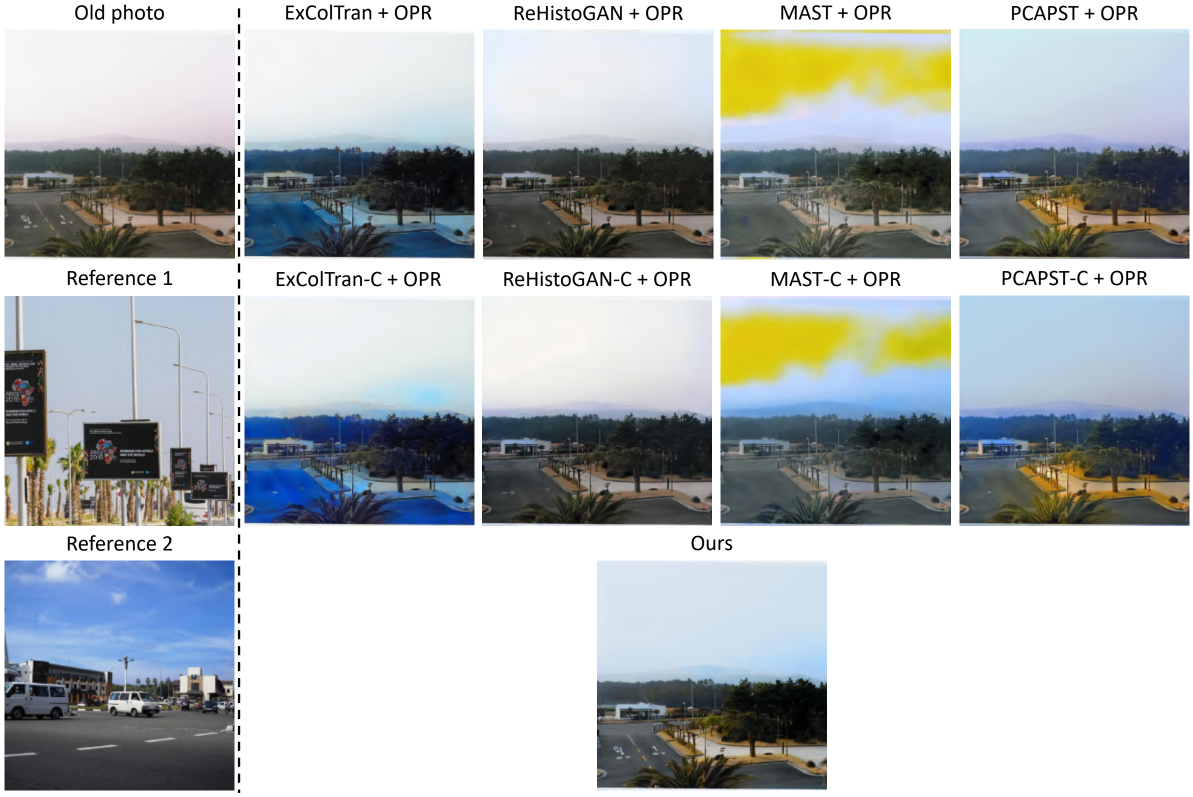


Figure 17. The results of multi-reference stylization using spatial concatenation and single-reference stylization. Baseline, e.g., ReHistoGAN denotes the result of performing single-reference stylization using reference 1. Meanwhile, Baseline-C, e.g., ReHistoGAN-C denotes the result of performing multi-reference stylization by spatially concatenating references 1 and 2 into a single reference.

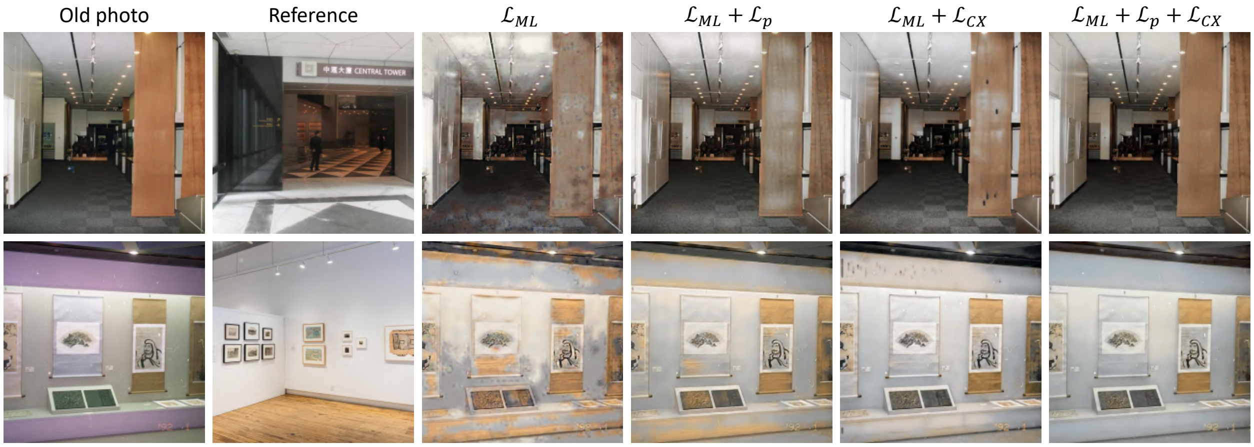

Old photo
Reference
$\mathcal{L}_{ML}$
$\mathcal{L}_{ML} + \mathcal{L}_{p}$
$\mathcal{L}_{ML} + \mathcal{L}_{CX}$
$\mathcal{L}_{ML} + \mathcal{L}_{p} + \mathcal{L}_{CX}$

Figure 18. Ablation study on loss functions for single stylization 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. 19. 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 study, we explore the capability of the merging-

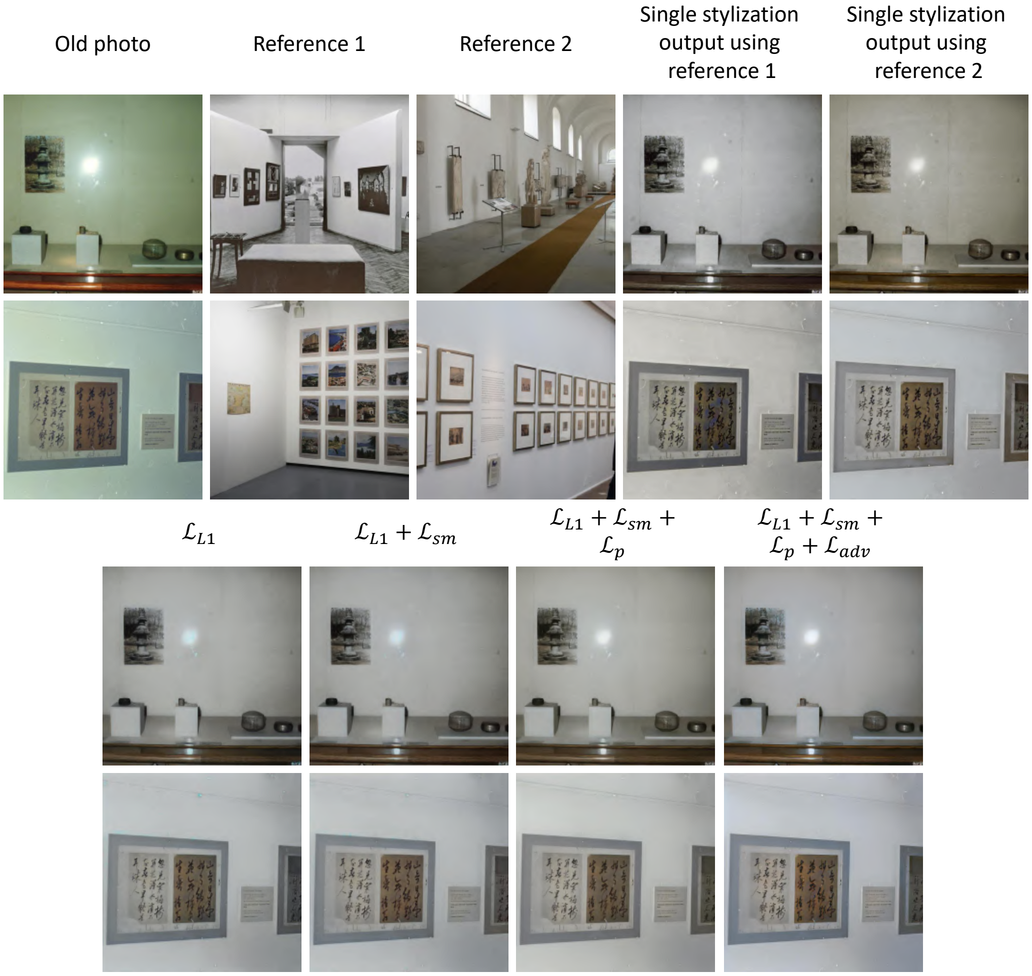

Old photo
Reference 1
Reference 2
Single stylization output using reference 1
Single stylization output using reference 2
$\mathcal{L}_{L1}$
$\mathcal{L}_{L1} + \mathcal{L}_{sm}$
$\mathcal{L}_{L1} + \mathcal{L}_{sm} + \mathcal{L}_{p}$
$\mathcal{L}_{L1} + \mathcal{L}_{sm} + \mathcal{L}_{p} + \mathcal{L}_{adv}$

Figure 19. Ablation study on loss functions for merging-refinement subnet.

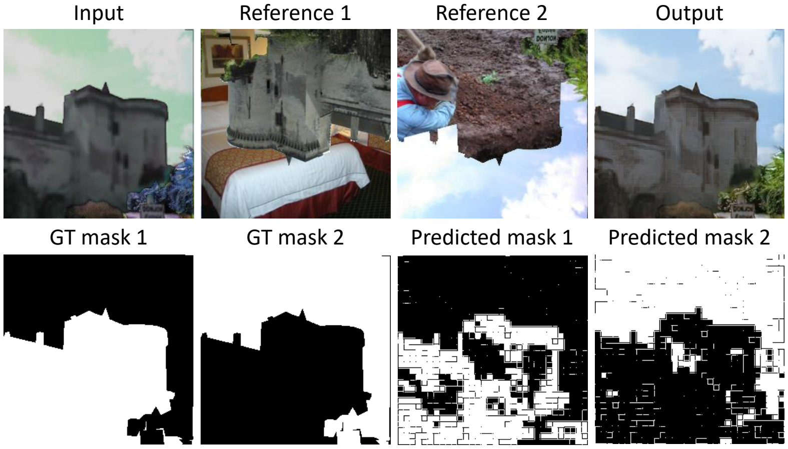


Figure 20. Study on the capability of merging-refinement subnet $\mathcal{M}$ to select relevant regions from multiple references to transfer their styles to the corresponding regions in the input.

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. 20, our $\mathcal{M}$ can select relevant regions in the input where it achieves an average of 70.70% mIoU for both predictions.

8. 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. 16, 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 [21] with severe SD and UD in Fig. 21 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. 22, 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. 22 and is on par with the SOTA PST network (PCAPST [5]) results.

Modernization results using unrelated references. Fig. 23 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. 24. 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. 25, Fig. 26, Fig. 27, and Fig. 28. 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. 29, Fig. 30, Fig. 31, and Fig. 32. In most cases, the results produced by our method are more preferably selected by the users compared to other baselines.

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. 5, 6, 9, 10, 11, 13, 18, 19, 25, 26, 27, 28

[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. 3

[3] 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. 3

[4] 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. 4

[5] 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. 3, 5, 6, 9, 10, 12, 13, 16, 18, 19, 25, 26, 27, 28

[6] 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

[7] 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. 3, 7

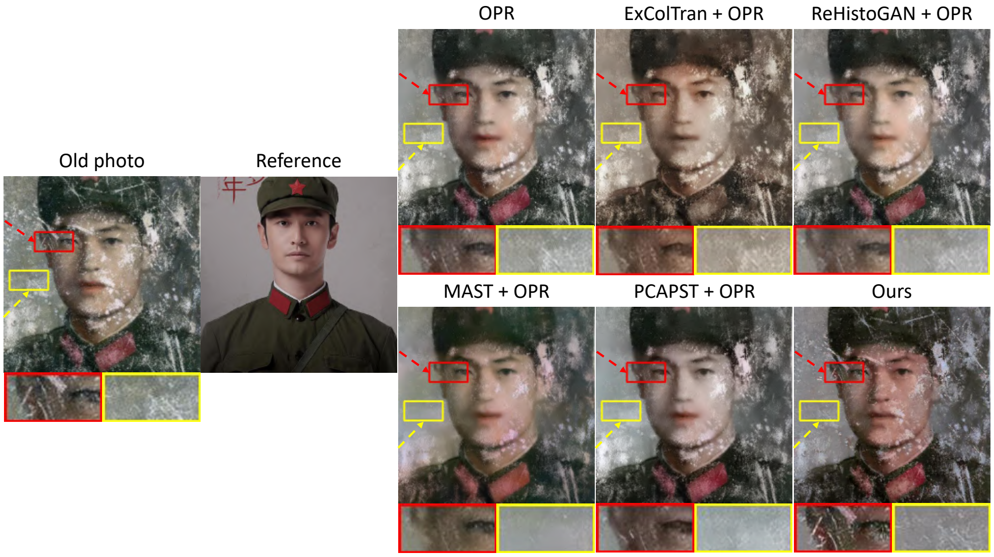


Figure 21. Results of our method compared to other baselines on an old photo from RealOld dataset [21] with severe structured and unstructured degradations.

[8] 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. 5, 6, 9, 10, 12, 13, 18, 19, 25, 26, 27, 28

[9] 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. 7

[10] 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. 3, 7

[11] 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. 3, 4

[12] 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] 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. 2, 12

[14] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 12

[15] 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. 7

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

[17] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 7

[18] 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. 5, 6, 9, 10, 11, 13, 19, 25, 26, 27, 28

[19] 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. 6

[20] 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. 7

[21] 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. 2, 3, 4, 16, 17

[22] 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. 3, 5, 6, 9, 10, 11, 13, 18, 19, 25, 26, 27, 28

[23] 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. 3, 4, 6, 7

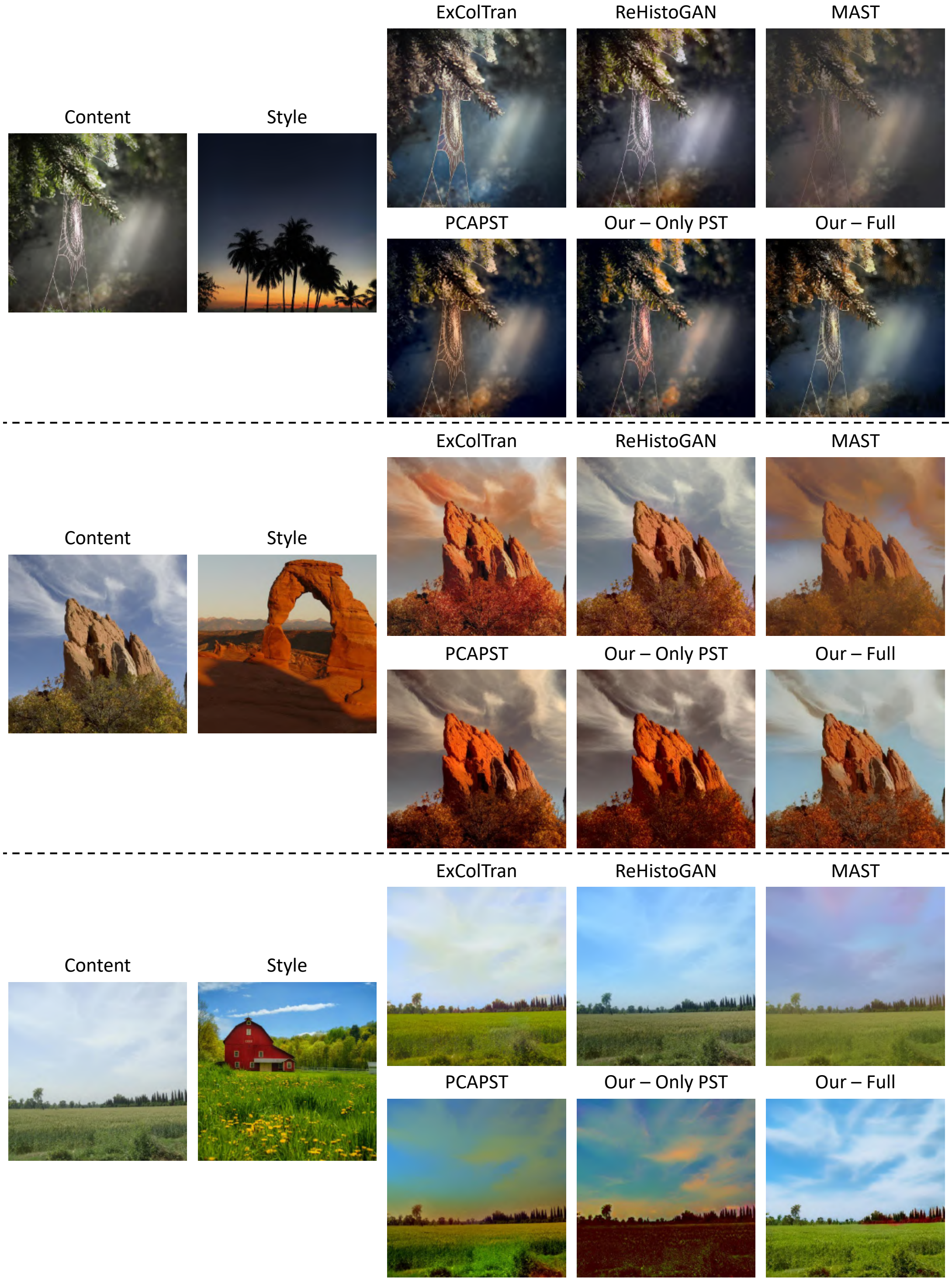


Figure 22. Comparison of stylization results on modern images. Our MROPM-Net, denoted as Our – Full, achieves the best local PST results on all examples compared to other PST baselines such as MAST [8] and PCAPST [5], and other baselines such as ExColTran [22] and ReHistoGAN [1]. Our – Only PST denotes the results of PST using our PST network (without style code predictor and merging-refinement subnet).

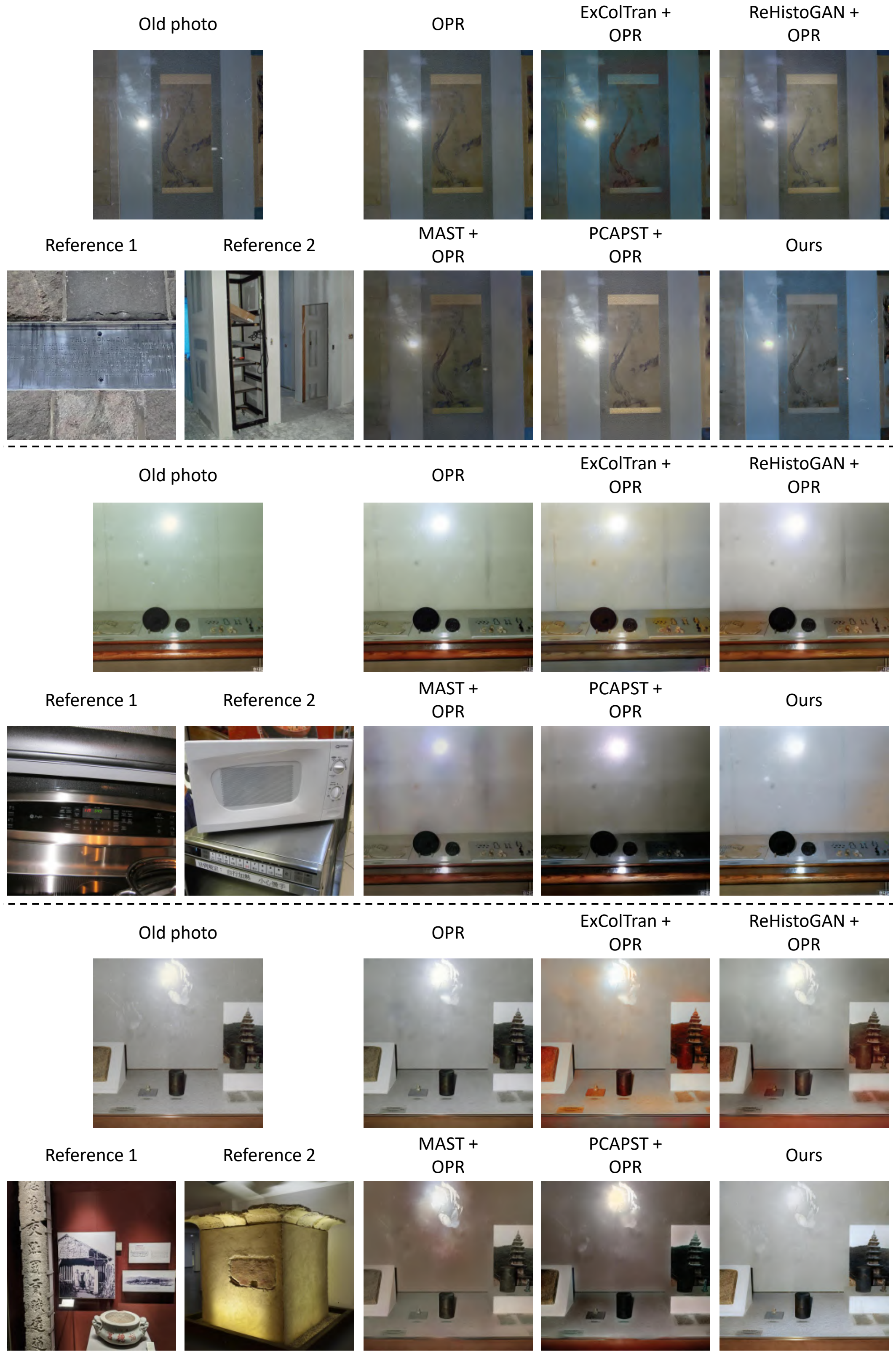


Figure 23. Old photo modernization results using unrelated references. In most cases, our method outperforms baselines (OPR [18], ExColTran [22] + OPR, ReHistoGAN [1] + OPR, MAST [8] + OPR, and PACPST [5] + OPR) even though the references are unrelated with the old photo. Reference-based baselines use reference 1 as their reference.

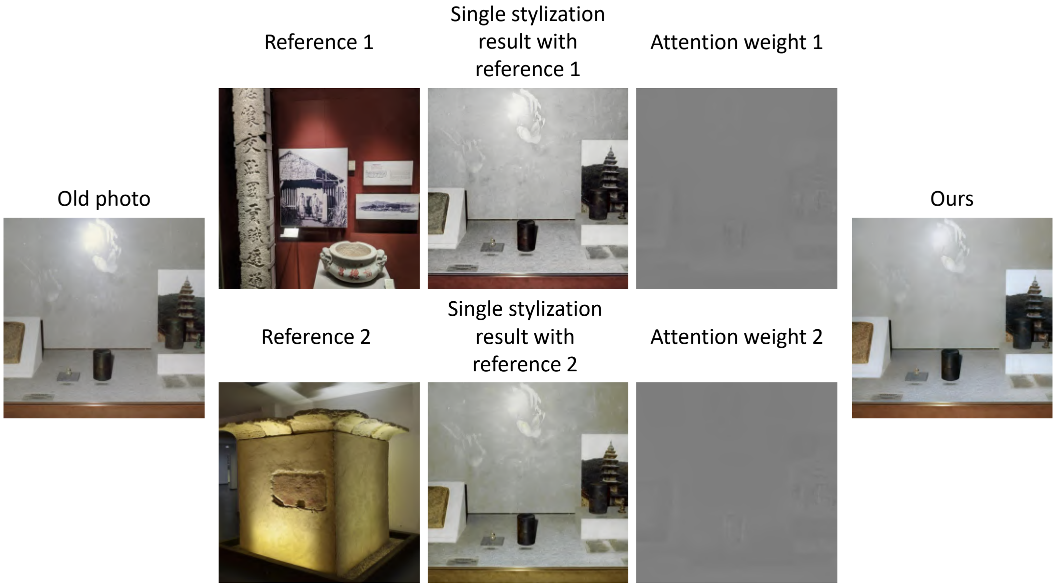

Reference 1
Single stylization result with reference 1
Attention weight 1
Old photo
Ours
Reference 2
Single stylization result with reference 2
Attention weight 2

Figure 24. The internal working of our MROPM-Net when handling unrelated references.

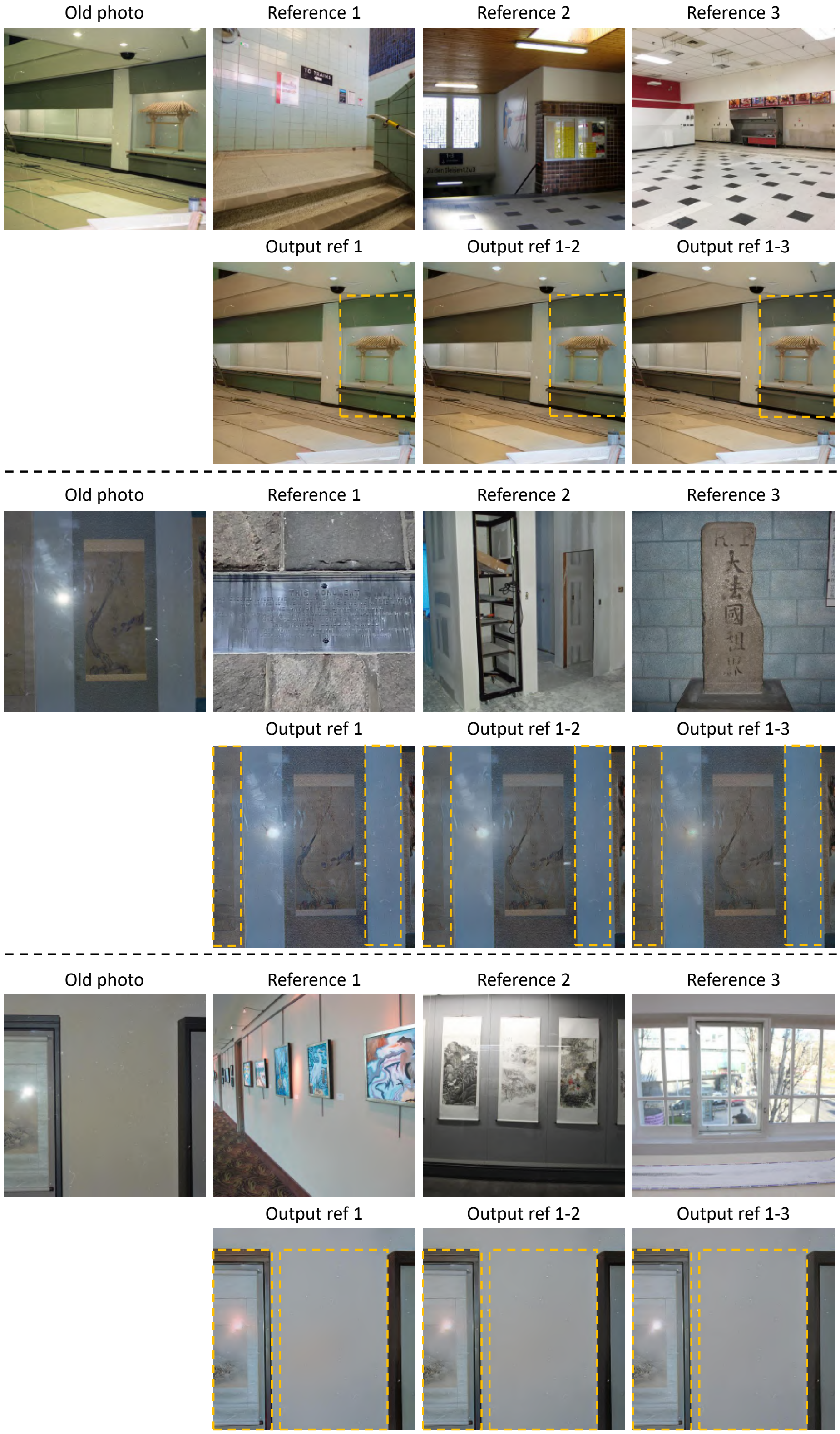


Figure 25. Progressive old photo modernization results using three references. Some regions with distinctive improvements are shown inside yellow boxes.

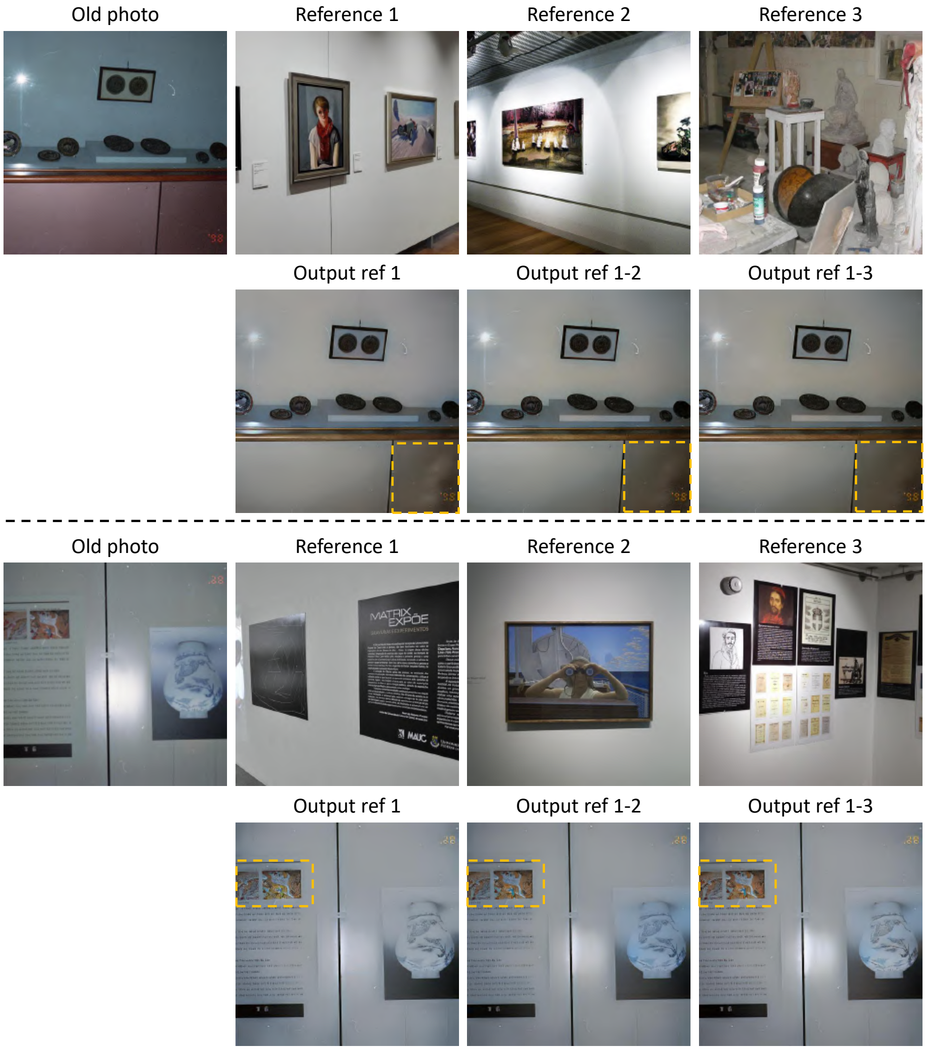


Figure 26. Progressive old photo modernization results using three references. Some regions with distinctive improvements are shown inside yellow boxes.

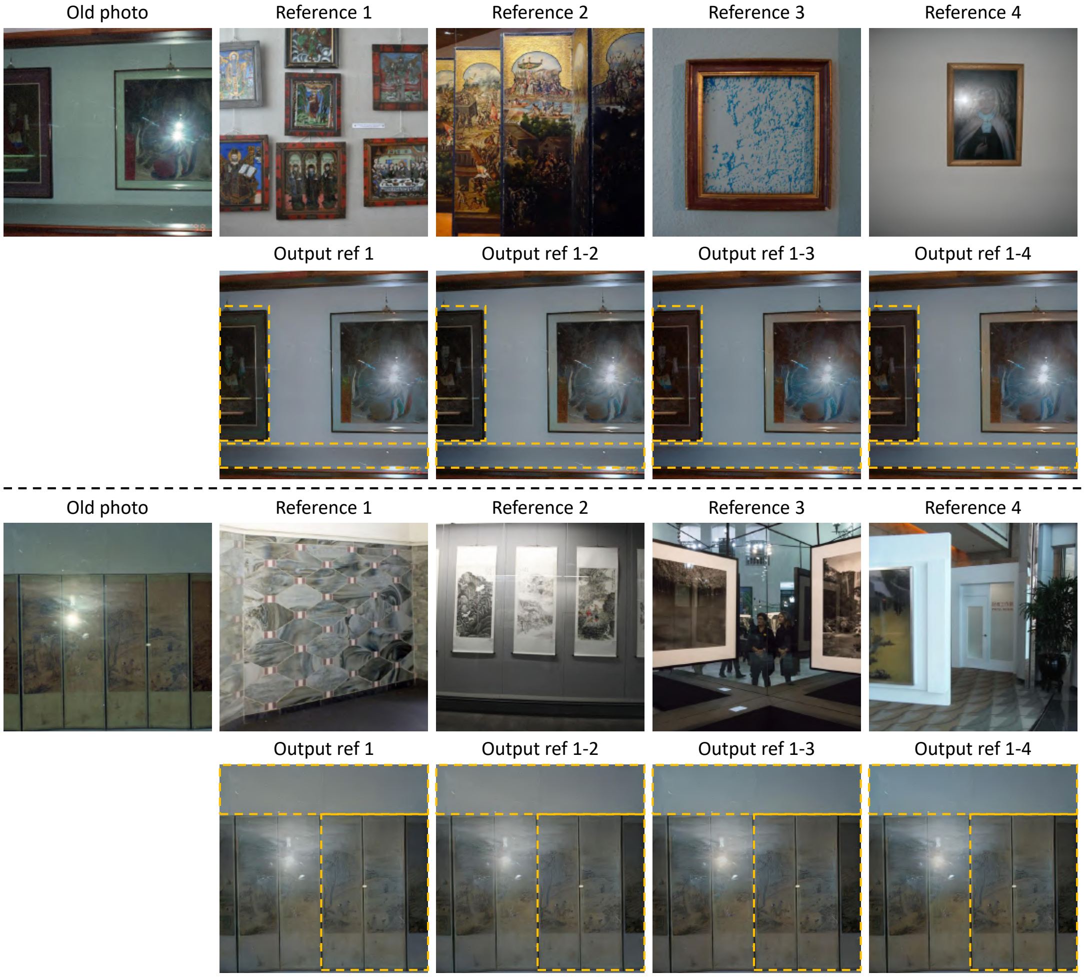


Figure 27. Progressive old photo modernization results using four references. Some regions with distinctive improvements are shown inside yellow boxes.

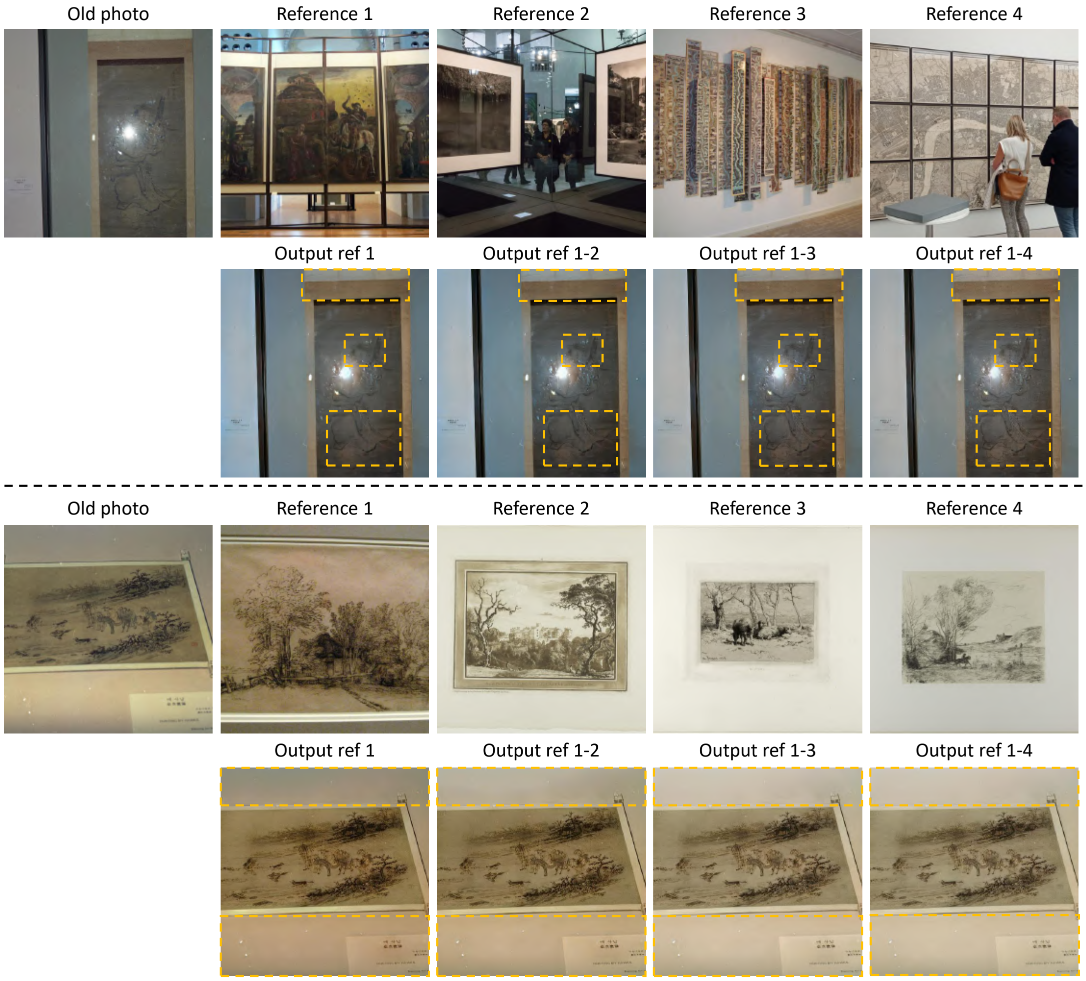


Figure 28. Progressive old photo modernization results using four references. Some regions with distinctive improvements are shown inside yellow boxes.

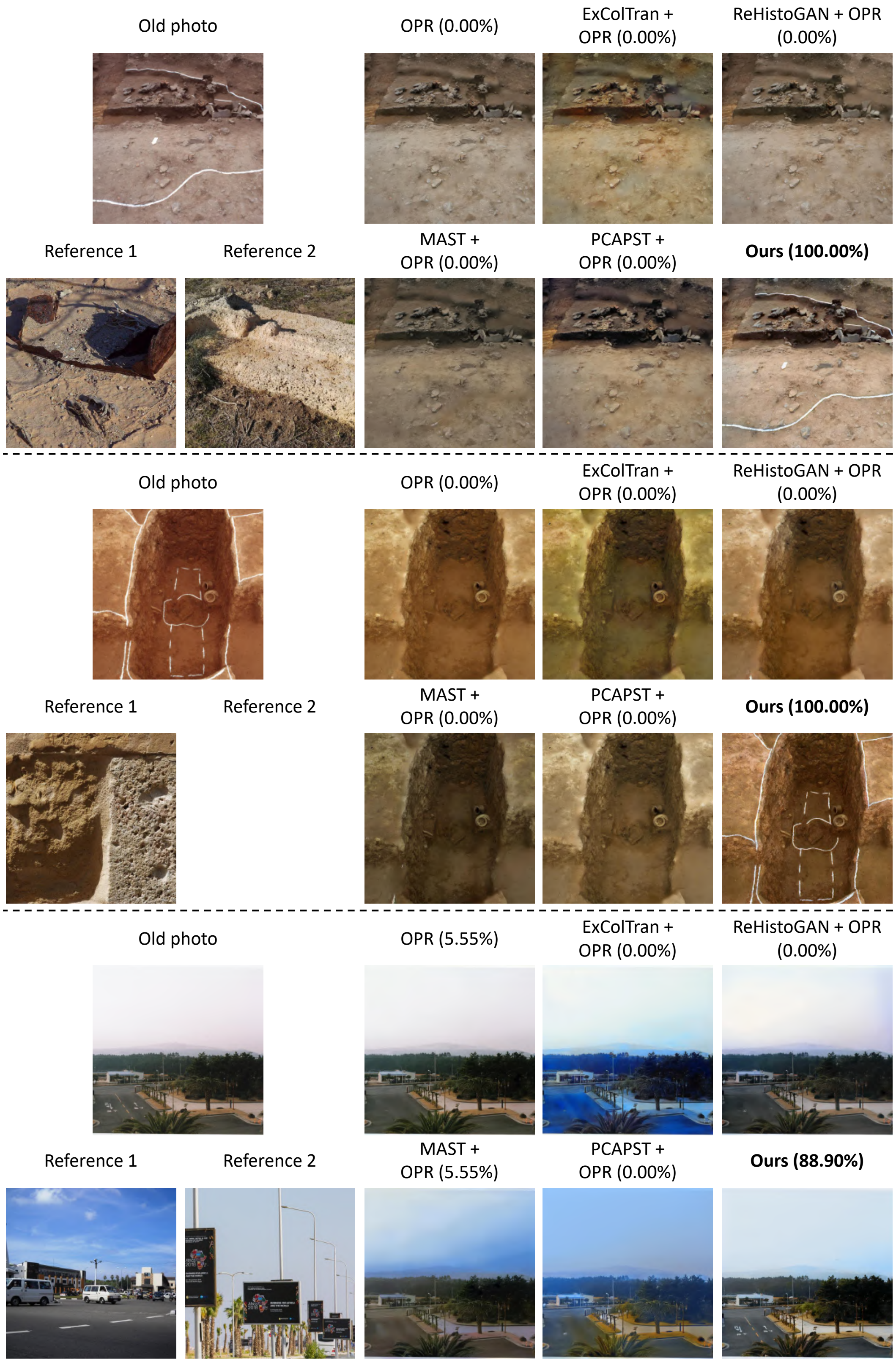


Figure 29. User study results with the percentage of user voting. Our method compares favorably against baselines (OPR [18], ExColTran [22] + OPR, ReHistoGAN [1] + OPR, MAST [8] + OPR, and PCAPST [5] + OPR). Reference-based baselines use reference 1 as their reference.

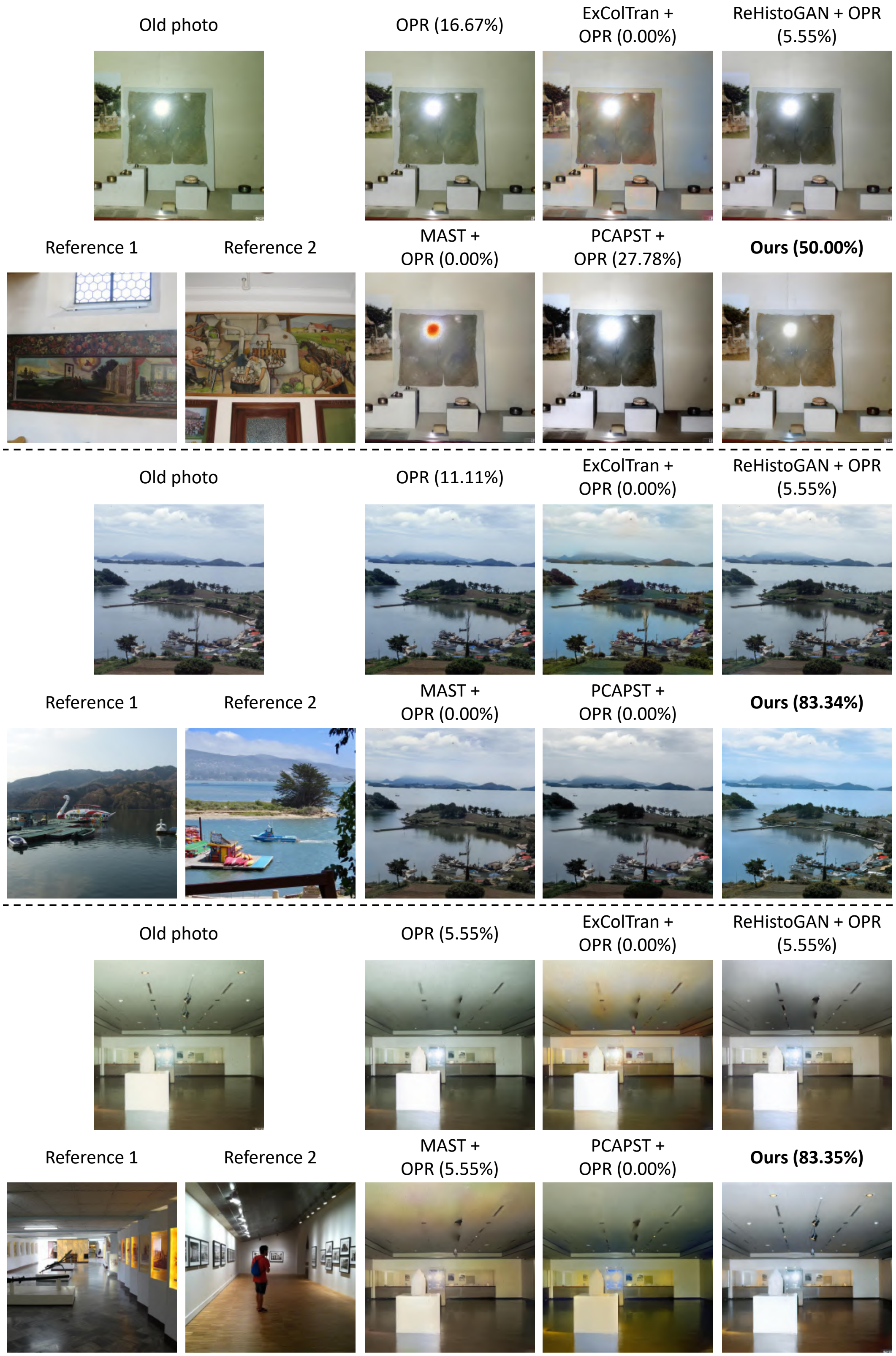


Figure 30. User study results with the percentage of user voting. Our method compares favorably against baselines (OPR [18], ExColTran [22] + OPR, ReHistoGAN [1] + OPR, MAST [8] + OPR, and PCAPST [5] + OPR). Reference-based baselines use reference 1 as their reference.

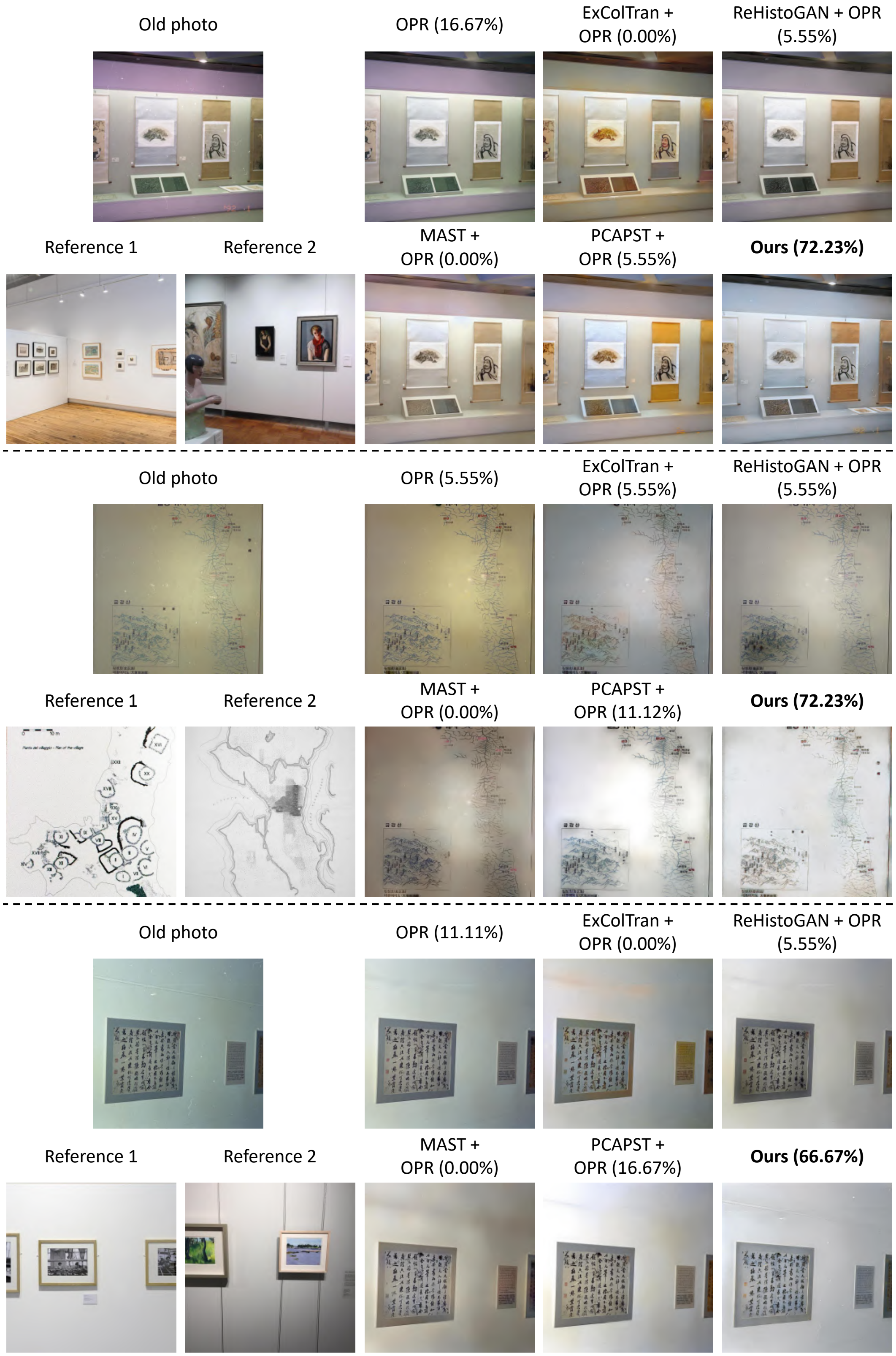


Figure 31. User study results with the percentage of user voting. Our method compares favorably against baselines (OPR [18], ExColTran [22] + OPR, ReHistoGAN [1] + OPR, MAST [8] + OPR, and PCAPST [5] + OPR). Reference-based baselines use reference 1 as their reference.

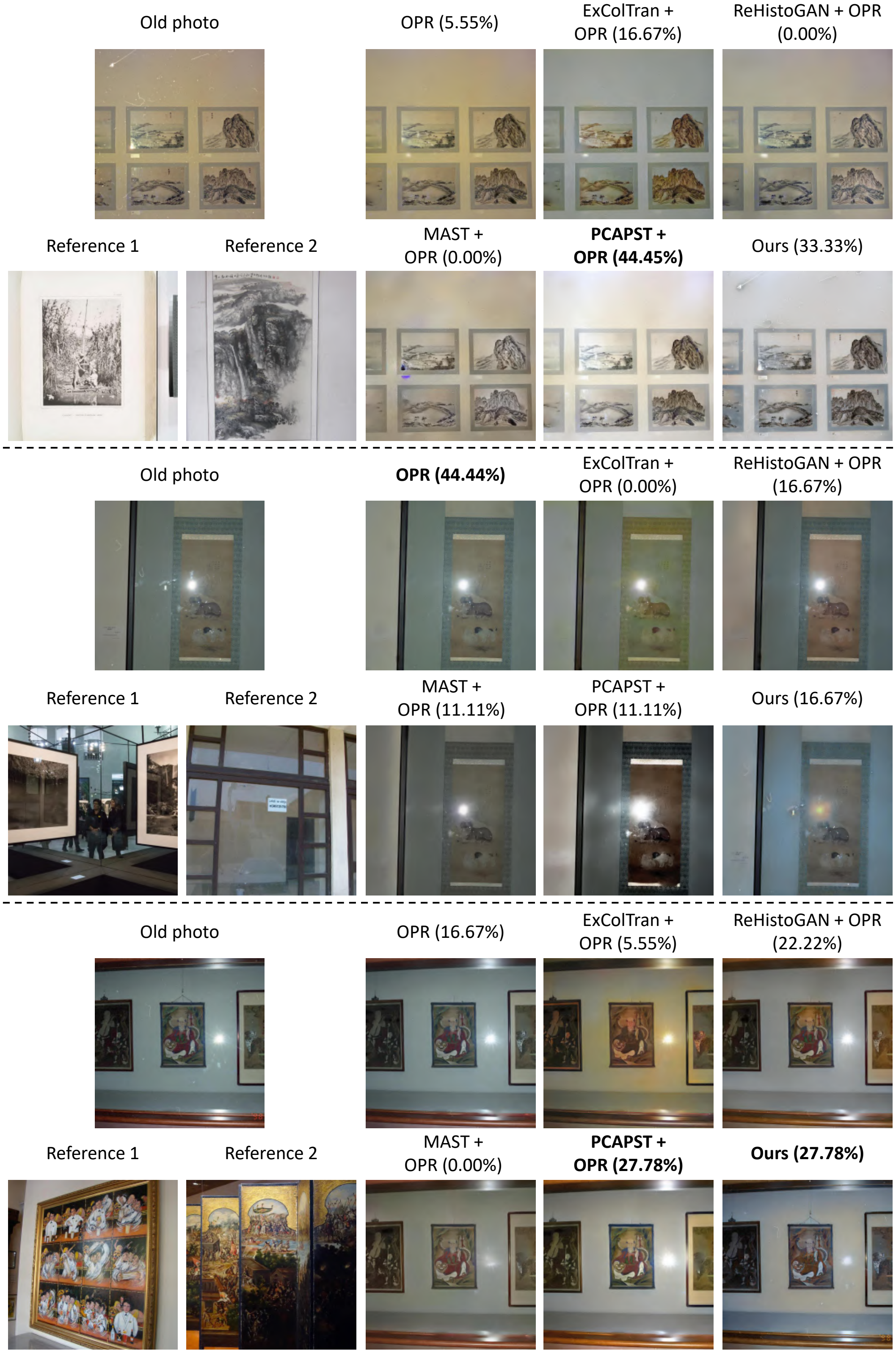


Figure 32. User study results with the percentage of user voting. Our method compares favorably against baselines (OPR [18], ExColTran [22] + OPR, ReHistoGAN [1] + OPR, MAST [8] + OPR, and PCAPST [5] + OPR). Reference-based baselines use reference 1 as their reference.