Close Imitation of Expert Retouching for Black-and-White Photography

Seunghyun Shin[1] Jisu Shin[1] Jihwan Bae[2] Inwook Shim[3†] Hae-Gon Jeon[1†]
[1]GIST AI Graduate School [2]School of Medicine, Cha University [3]Inha University
{seunghyuns98, jsshin98}@gm.gist.ac.kr, haegonj@gist.ac.kr
jihwan1008@chauniv.ac.kr, iwshim@inha.ac.kr

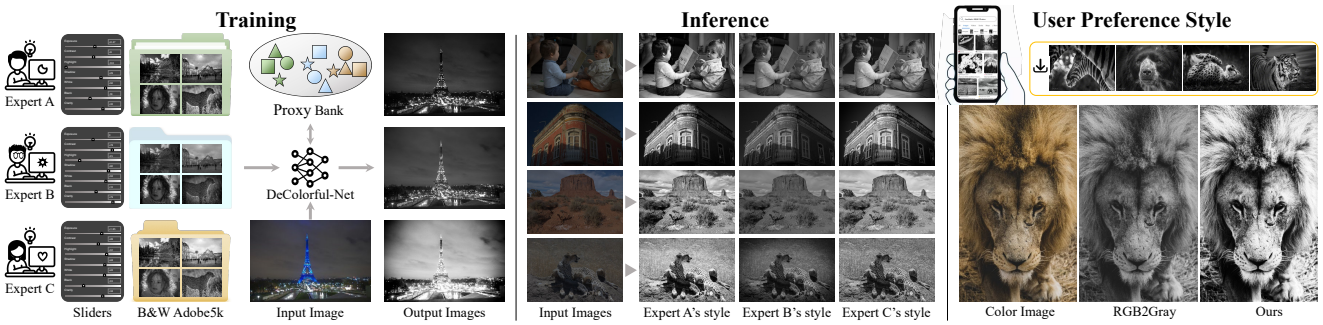


Figure 1. We train our Decolorful-Net on our new *B&W Adobe 5K* dataset, retouched by three experts, to produce expert-specific proxies and corresponding output images. Once trained, we can decolorize an image with any style based on internet downloaded photos.

Abstract

Since the widespread availability of cameras, black-and-white (BW)photography has been a popular choice for artistic and aesthetic expression. It highlights the main subject in varying tones of gray, creating various effects such as drama and contrast. However, producing BW photography often demands high-end cameras or photographic editing from experts. Even the experts prefer different styles depending on the subject or even the same subject when taking grayscale photos or converting color images to BW. It is thus questionable which approach is better. To imitate the artistic values of decolorized images, this paper introduces a deep metric learning framework with a novel subject-style specified proxy and a large-scale BW dataset. Our proxy-based decolorization utilizes a hierarchical proxy-based loss and a hierarchical bilateral grid network to mimic the experts' retouching scheme. The proxy-based loss captures both expert-discriminative and class-sharing characteristics, while the hierarchical bilateral grid network enables imitating spatially-variant retouching by considering both global and local scene contexts. Our dataset, including color and BW images edited by three experts, demonstrates the scalability of our method, which can be further enhanced by constructing additional proxies from any set of BW photos like Internet downloaded figures. Our Experiments show that our framework successfully produce visually-pleasing BW images from color ones, as evaluated by user preference with respect to artistry and aesthetics. Code and dataset are publicly available at https://github.com/seunghyuns98/Decolorization

†Corresponding Author

1. Introduction

Since the birth of camera photography, BW photography has been beloved by photographers and the public because its better dynamic range and natural-looking sharpness [26] convey an ability to enjoy the textures, lines and patterns as well as contrasts. By representing subjects in varying shades of neutral gray, people can appreciate the revealed aesthetics that cannot be enjoyed with color photos.

Barbara Davidson, a winner of the 2011 Pulitzer prize in feature photography, adds both the intimacy and emotion to the photos with the use of BW because she believes that color can sometimes become visual pollution [24, 49]. In addition, some of the BW conversions in Instagram were selected as the 2021 most popular filters in the world [7]. Recent hit movies like 'Parasite' and 'Mad-Max' are even re-opening in BW videos. This trend indicates that the artistic and aesthetic purpose of BW photography is becoming recognized.

However, artistic and aesthetic BW photography is not easy for normal people to achieve. The monochrome camera's price is burdensome, and the photo-retouching required to produce satisfying artistic effects still remains the domain of experts. Although smartphones and DSLR cameras can provide basic BW photography functionalities, they are not aesthetically delightful enough. Image decolorization in the computer vision literature has mostly delved into preprocessing for various downstream tasks, including semantic segmentation [62], stereo matching [20] and image recognition [27]. Whether local [16, 31, 37, 52] or global approaches [6, 57], they focus on minimizing the loss of textures during the color to grayscale conversion. Of

course, a data-driven manner with a convolutional neural network (CNN) in [36] is designed to take advantage of the local and global approaches but still focuses on contrast preservation during the color-to-monochrome conversion. Unfortunately, the question of whether the decolorized image is aesthetically entertaining has not been answered yet.

In this paper, we design a novel decolorization framework that allows users to easily produce quasi-professional BW photography with the subject-style aware proxy guidance in Fig. 1, called DeColorful-Net. As usual, the decolorization scheme differs by the scene and subject, and it is not easy for users to suggest their favorite options in detail, different from image editing with explicit user guidance like image [23, 25] or text [1, 33, 35]. To produce user-preferred BW images, the proxy-based learning using implicit vector representations as a user-guidance is the most probable framework.

To do this, we first construct a large-scale BW dataset to harness the network. Three professional photographers are hired to produce their own decolorized version of the MIT-Adobe 5K [5]. Considering the fact that the retouching scheme varies on the experts, and even on the subject of a scene, we design a hierarchical proxy-based deep metric learning (DML) framework to extract subject-style aware proxies. With the given style from the user and the defined proxies, we propose a multi-level bilateral grid to imitate the experts' coarse-to-fine retouching schemes along with to alleviate unwanted artifacts coming from conventional single-level bilateral grids. To further demonstrate the scalability of our DeColorful-Net, we collect public BW photos on the internet taken by professional photographers. By simply fine-tuning the proxy generation network, we construct new proxies, allowing the creation of user-specified BW images for arbitrary new styles. We show how DeColorful-Net mimics the experts' retouching in quantitative evaluations. Furthermore, extensive and meticulous user studies support our claim that DeColorful-Net produces more aesthetic BW images compared to existing decolorization methods and even commercial filters.

2. Related Work

Deep Metric Learning. A goal of metric learning is to learn a new metric to address the distances between data and to classify them. With the help of CNNs, DML [21, 34] enables non-linear data to be handled in a higher feature space. The loss functions aim to efficiently learn an embedding space where similar features are attracted, and dissimilar examples are repelled [42, 48]. One of the representative concepts in the loss functions is proxy-based losses [2, 41, 44] which assigns proxies for each class and learns correlation between data points and proxies. However, each data point is associated only with proxies which miss the data-to-data relations. Proxy-Anchor loss [30] solves this limitation by designating each proxy as an anchor and associating it with the entire data in a batch, which allows for inter-data interactions throughout the training process. In contrast to the other proxy-based methods, Hierarchical proxy-based loss [60] builds multiple levels of learnable proxies where the lowest level of proxies follows the scheme of existing proxy-based losses. Imposing the hierarchical structures acts as a regularization to prevent models from overfitting

DML-based Low-level Vision Tasks. With its powerful ability to discover embedding space, DML has been in the public eye, and is one of the main frameworks for several vision tasks such as image retrieval [30, 51, 64], person re-identification [8, 9, 47], and face recognition [12, 46]. Recently, it has been also applied in low-level vision tasks, including image enhancement and image super-resolution. A work in [11] finds matches between geometric shape descriptors under proper metrics learned by Triplet metric learning [40] to improve visual qualities in super-resolution. Another work in [29] proposes a personalized high dynamic range photography by modeling various user preferences as feature vectors. Both preferred and non-preferred images of users are fed to the CNN with a triplet loss function.

Image Decolorization. Classical decolorization focuses on converting 3-channel color images into 1-channel grayscale images while preserving the scene structures and contrasts in the original images. Methods to keep the scene configurations during the conversion can be categorized into local and global methods. Local methods [16, 31, 37, 52] minimize differences in illumination and chrominance information between color and BW images, by using different mapping functions according to local regions of the input images. Since those methods usually cause unpleasant halo artifacts [3], global methods [6, 38, 39, 57] utilize global linear mapping functions to effectively preserve a perceptual quality and spatial information. One of the global methods in [39] minimizes the contrast gap between color and the decolorized image by optimizing the Log-Euclidean distance for them. A work in [38] introduces the semi-parametric approach. This uses the second-order multivariance polynomial gradient to compare the color and the decolorized images. In [6], a perception preserving decolorization is proposed by minimizing a multi-level perceptual loss between learned features from input decolorized and target images.

Unlike the existing decolorization methods which only focus on preventing the loss of structural information, we view the decolorization task as an aesthetic realm and propose a novel framework for generating visually plausible BW images from color ones.

3. Expert-Retouched Imageset

MIT-Adobe 5K dataset [5] consists of $5,000$ colorful images and each image retouched by 5 photographers using Adobe Lightroom. The retouching style of each photographer differs significantly, as already shown in [5]. Since it is guaranteed to include a broad diversity of scenes, subjects, and lighting conditions, it is widely used in image enhance-

(a) Adjustment (b) Retouched Image

Figure 2. Illustration of three experts' retouching schemes on the proposed BW photography dataset.

ment [15, 53] and harmonization [10, 54]. To incorporate those rich scene configurations, we utilize this dataset to produce aesthetic BW images.

To represent a variety of aesthetic effects in BW images, we hire three professional photographers to retouch the $5,000$ images with the payment. Fig. 2 shows the example of their retouching schemes and the corresponding BW photo examples obtained from them, who are referred to as experts A, B, and C. Their retouching order is generally similar: (1) adjusting global parameters such as gamma curve, exposure, and contrast, (2) applying spatially varying filters like vignetting or local gamma correction.

However, we find out significant variations in the detailed retouching scheme among experts. Expert A prefers higher exposure, enhanced clarity, and shadow effects to brighten dark areas. Expert B tends to enhance scene details by adjusting texture and clarity, further highlighting the main subject like a person and a flower with a vignetting effect. Expert C controls levels of highlight and shadow to manipulate the brightest and darkest segments in images for spatial variant retouching. This retouching enables outputs to have a more dynamic range than input images.

We also observe that their retouching schemes depend on scene configurations, such as the number of subjects, types, and photographic compositions. For example, expert A tends to use edge enhancement and higher exposure for landscape photographs while increasing the contrast on buildings, and using vignetting effect to concentrate on landscapes. For portrait photographs, expert A uses a strong shadow effect to enrich the fine details of hair-like structures. Local contrast suitable to capture the skin's pores and wrinkles is then applied. Other experts also have their own retouching styles dealing with a variety of scene configurations as well.

Considering the varying retouching scheme, we categorize our dataset into four classes (human, non-human, building, and nature) based on [5], and ask them to achieve scene-aware photo retouching. In our pipeline, images edited by the experts are regarded as ground-truth BW images.

4. Methods

The main challenge in creating quasi-professional BW photography with our dataset is that the choice of retouching schemes differs depending on the expert, even by the subject of a scene. Furthermore, experts usually start with global adjustments like exposure and contrast, followed by spatially specific retouching. Therefore, it is essential to integrate both global and local features to effectively manage these spatially varying retouching techniques.

We handle these issues with a novel proxy-based image decolorization network, named *DeColorful-Net*, that brings up the concept of DML into the image decolorization task. Our DeColorful-Net consists of two stages, which first yields style-scene aware proxies based on DML, and then produces a decolorized image from a hierarchical bilateral grid network, as illustrated in Fig. 3

4.1. Proxy Generation

Let us consider the embedding space, where proxies represent subject-aware preferred styles for various users. Based on the proxy-based metric learning [41], we aim to learn the proxy generation network which produces the subject-style aware proxies in the embedding space. For the input $\mathbf{I} \in \mathbb{R}^{h \times w \times 1}$, a proxy is annotated as $\{p_{ij} | i \in 0, \cdots, N-1, j \in 0, \cdots, M-1\}$, where i and j denote an object and a style index, respectively. Here, we set both N and M to 4 because our dataset consists of 4 subjects (human, non-human, nature, and building) and 4 styles (RGB2Gray operator and 3 experts). These learnable 16 proxies are held in Proxy Bank $\mathcal{P}$, which is later used to provide target proxy when given a style and a subject index.

For feature extraction, input images are fed into two individual encoders: ResNet [19] and VGG [50] which are used to classify subjects and styles, respectively. A cross-entropy loss is then calculated based on the output vectors x and y from each classifier header with fully-connected layers as:

$$\mathcal{L}_{CE} = \frac{1}{N} \sum_{i=0}^{N-1} \hat{x}_i \log(x_i) + \frac{1}{M} \sum_{j=0}^{M-1} \hat{y}_j \log(y_j), \quad (1)$$

where $\hat{x}$ and $\hat{y}$ are binary indicators of the true label for the subject and the style, respectively.

To learn an embedding space where the proxy from the same class of the input image is regarded as positive and the other proxies are negative, we adopt Proxy-Anchor loss [30]. Let v denote the embedding vector obtained by an additional lightweight MLP architecture and p indicates the proxy from the proxy bank $\mathcal{P}$. Then, the Proxy-Anchor loss is given by

$$\mathcal{L}_{PL}(v, \mathcal{P}) = \frac{1}{|\mathcal{P}^+|} \sum_{p \in \mathcal{P}^+} \log \left(1 + \sum_{v \in V_p^+} e^{-\alpha(k(v,p)-\delta)}\right) + \frac{1}{|\mathcal{P}|} \sum_{p \in P} \log \left(1 + \sum_{v \in V_p^-} e^{\alpha(k(v,p)+\delta)}\right), \quad (2)$$

where δ is a margin, α is a scaling factor, $k(\cdot, \cdot)$ is the cosine similarity function between two vectors and $\mathcal{P}^+$ denotes the set of positive proxies of data in the batch. In addition, for

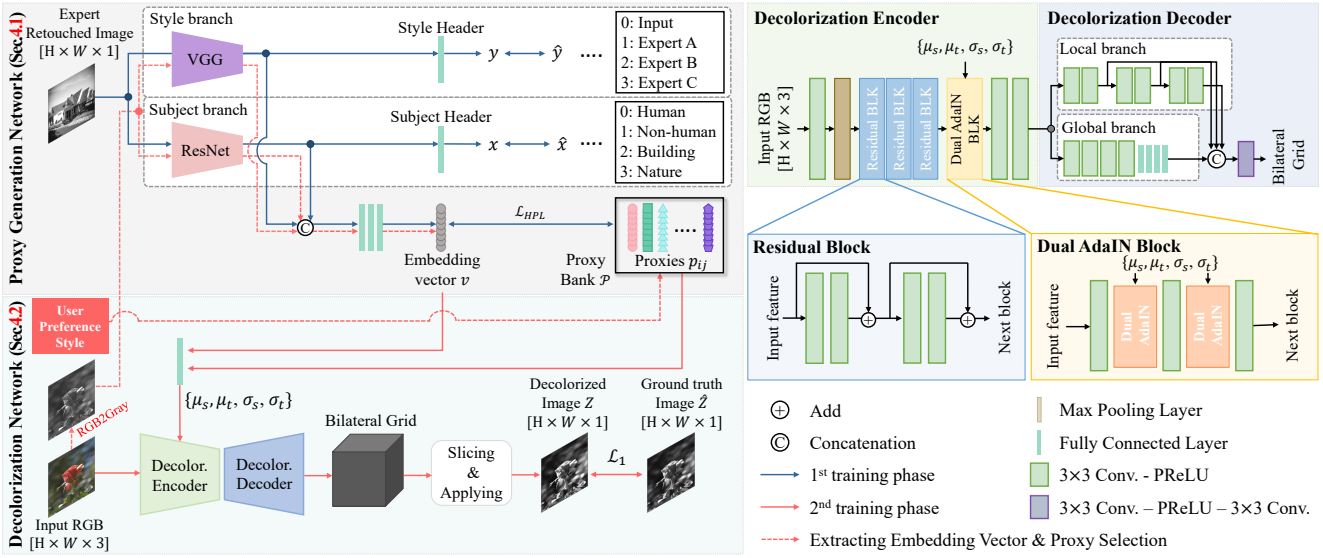


Figure 3. The overall training pipeline of our DeColorful-Net, consisting of the proxy generation network and decolorization network. PReLU denotes the PReLU activation [18].

each proxy p, a batch of embedding vectors V is divided into the set of positive vectors of V_p^+ and negative vectors V_p^-.

We note that experts first roughly manipulate the whole image, based on their artistic preferences, and then proceed to detailed retouching based on scene configurations later on. Accordingly, we adopt a hierarchical proxy-based loss (HPL) [60], which allows the network to capture expert-discriminative as well as class-sharing characteristics.

Since the proxy is learned in the subject-style dependent manner, the upper-level proxy p_j^H, where H denotes the higher level, is obtained as below:

$$p_j^H = \frac{1}{N}\sum_i^n p_{ij}, \tag{3}$$

and thus 4 proxies exist on a higher level according to the expert's style. The HPL based on the $\mathcal{L}_{PL}(v, \mathcal{P})$ is then formulated as below:

$$\mathcal{L}_{HPL} = \mathcal{L}_{PL}(v, \mathcal{P}) + \lambda_1 \mathcal{L}_{PL}(v, \mathcal{P}^H), \tag{4}$$

where λ_1 is a loss weight for a higher level of proxies and is empirically set to 0.1. $\mathcal{P}^H$ means the set of proxies on a higher level. Therefore, our HPL can seize class-shared knowledge from higher-level coarse proxies along with class-discriminative features through the lower-level proxies, just as the Proxy-Anchor loss does.

In total, the final loss function for the proxy generation in DeColorful-Net is defined as:

$$\mathcal{L}_{proxy} = \mathcal{L}_{CE} + \lambda_2 \mathcal{L}_{HPL}, \tag{5}$$

where λ_2 is the hyper-parameter, which is empirically set to 0.1. By training the proxy generation network on $\mathcal{L}_{proxy}$, we obtain subject-style aware proxies that represent each training set of our dataset retouched from the three experts.

4.2. Image Decolorization via Bilateral Grid

The final goal is to produce a quasi-professional decolorized image from single RGB images, which allows users to engage in selecting the user preference style. To transfer the preferred style with the help of the proxy defined for each subject-style from Sec. 4.1, we incorporate Dual-AdaIN [53] into the Decolorization Encoder to mitigate the discrepancy between a source (an input RGB image) and a target domain (subject-style aware proxy). Dual-AdaIN transforms the feature map from the encoder by utilizing the discrepancy between the source and the target vectors. Here, the embedding source vector v is extracted by feeding the converted BW image into the pre-trained proxy generation network, and the proxy p_{ij} (target vector) is selected from the Proxy Bank according to a user given style-index and a subject-index. Hence, the source vector v and the target vector p_{ij} are fed to a fully-connected layer that outputs the set of mean and variance vectors $\{\mu_s, \sigma_s, \mu_t, \sigma_t\}$. Transformation of the feature map is given by:

$$\mathcal{F}' = \sigma_t\left(\frac{\mathcal{F} - \mu_s}{\sigma_s}\right) + \mu_t, \tag{6}$$

where $\mathcal{F}$ is an intermediate feature map from the decolorization encoder and $\mathcal{F}'$ denotes a transformed feature map.

Along with the Dual-AdaIN embedded encoder, we leverage the idea of bilateral grid processing [15] to implement the edge-aware spatially varying retouching scheme. Additionally, to mimic the coarse-to-fine retouching of experts, we propose a multi-level bilateral grid framework that extracts different levels of feature maps from a local decoder of a bilateral grid network. By using the hierarchical feature maps of the Decolorization Decoder, we can gather information from the scene, which facilitates to generate a decolorized image with spatially-variant effects.

Figure 4. Visualization of embedding vectors using t-SNE [56] and examples of proxy modification. The color bars represent the proxies used for producing the decolorized images adjacent to them. The predominant color indicates the style class, while the position of the stripes denotes the channels altered from the original proxy. Here, Multi-Slider contains channels related to 'Exposure', 'Clarity', and 'Shadow'

Next, we fuse multi-level local feature maps using an additional convolution layer to construct the bilateral grid, where each grid cell contains four weights, $\{w^r, w^g, w^b, w^d\}$. Compared to conventional decolorization methods [37, 52] that restrict the sum of the channel weights $\{w^r, w^g, w^b\}$ to 1, we do not limit the sum of the weight values and use an additional bias term w^d to capture the local changes in brightness and contrast similar to an affine matrix used in image enhancement tasks [14, 15, 17]. Then, the final grayscale image Z is produced as below:

$$Z_d = w_c^r R_d + w_c^g G_d + w_c^b B_d + w_c^d, \tag{7}$$

where $\{R_d, G_d, B_d\}$ means color channels at a pixel d on input image, respectively. c is the corresponding location on the grid of a pixel d.

To enforce the style consistency between the decolorized image Z and the expert retouched ground-truth image $\hat{Z}$, we use the cross entropy loss between the style indicator vector y and the binary indicator vector $\hat{y}$ of the user preference style index. The style indicator vector y is extracted from the proxy generation network's style header. Along with the L_1 loss, the final loss for decolorization is defined as:

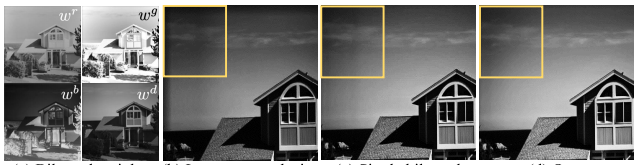
$$\mathcal{L}_{decolor} = \underbrace{\sum_{d=1}^{D}(Z_d - \hat{Z}_d)}_{L_1 \text{ loss}} + \lambda_3 \underbrace{\frac{1}{M}\sum_{j=0}^{M-1}\hat{y}_j \log(y_j)}_{\text{Cross Entropy loss}}, \tag{8}$$

where D and $\hat{Z}$ denote the total number of pixels in the image and the expert retouched ground-truth image, respectively. λ_3 is a hyper-parameter, which is empirically set to 0.01.

4.3. Analysis

To the best of our knowledge, our DeColorful-Net is the first to unfold the decolorization problem into an aesthetic aspect. To validate its effectiveness, we provide a series of analyses on DeColorful-Net.

Hierachical proxy-based loss. First, we observe a distribution of the embedding vectors in the learned space based on a hierarchical proxy-based loss (Eq. (4).) Fig. 4 visualizes the embedding vectors projected onto 2D space using t-SNE [56] where the same styles and subjects are categorized by the same color and shape, respectively. The clusters on the projection surface indicates that the proxies are well learned through the $\mathcal{L}_{HPL}$ in that the proxies are attracted and repelled according to their styles.



(a) Bilateral weights (b) Improper patch size (c) Single bilateral (d) Ours

Figure 5. Visualization of the weights and results of bilateral grid. Note that the patch size in (b) is 64.

Deep Metric Learning. We analyze the efficiency of a DML framework in decolorization task. Since the retouching parameters for BW photography are more limited than in color photo, experts should meticulously adjust them to produce visually pleasing results. According to [28], 9 sliders in retouching tools: 'Exposure', 'Contrast', 'Blacks', 'Whites', 'Shadows', 'Highlights', 'Clarity', 'Texture', and 'Vignetting' are keys for professional BW photos. By leveraging DML, our proxy generation network captures the differences of each expert's adjustment and encapsulates them in an output vector. To validate this, we create image sets with variations in each slider and mapped them onto the latent space. As illustrated in Fig. 4, these sets are distinctly clustered into separate groups. Also, by modifying 50 channels of RGB2Gray proxy that have the biggest discrepancies with the central values of these clusters, we can produce decolorized images that contain the corresponding effects. Additionally, we found that by manipulating these channels, we can handle multiple effects at once, and even can control the expert's proxy to better align with individual preferences, thereby generating outputs that more closely match a user's desired aesthetic.

Hierarchical bilateral grid. Lastly, we analyze the effectiveness of a hierarchical bilateral grid by comparing it with a single bilateral grid with various patch sizes. Fig. 5 (a) visualize the weight values, $\{w^r, w^g, w^b, w^d\}$, which are applied to each color channel with the additional bias term. Here, we found that the retouching schemes such as exposure, contrast, and local gamma correction, are controlled by the color channel weights. The vignetting effect, which is applied independently to the pixel values in the post-processing stage, is represented by the bias term.

To acquire the weights that produce the best result, we first tune the spatial range of the bilateral grid from 4 to 64. Fig. 5 (b) shows an example result with an improper spatial range. As the spatial range widens, the bigger segment of the image shares the same weight parameters, which leads to the poor local adjustment. On the other hand, as the range becomes narrow, the local regions are precisely delineated, but emboss the halo effects around the edge of the objects. To balance between them, we choose the proper range of the patch size of the single bilateral grid as 8 in our experiment.

However, an optimal spatial range can only partially mimic the experts' coarse-to-fine retouching. For example, some retouching schemes, such as vignetting effect, often cause visually unpleasing grid artifacts. It is obvious that a fixed single patch size has a limitation in generating the fading, which is dependent on the spatial location of the image. Our multi-level bilateral grid alleviates these artifacts by capturing the diverse receptive fields of the image, thus effectively implementing the spatially varying retouching. The grid artifacts revealed in the single bilateral grid from Fig. 5 (c) are removed in Fig. 5 (d).

4.4. Constructing Proxies from Internet Photos

We further demonstrate the scalability of our DeColorful-Net, enabling users to create quasi-professional BW photographs in their preferred style. For this, we additionally collect a variety of BW photos from professional photographers available online with their permission to use. We find out that photographers often hone their craft by specializing in a singular subject area. We thus choose two photographers for each subject (8 photographers in total). Note that we use 100 and 20 images for training and validation purposes, respectively.

Since it is impossible to obtain color images corresponding to the downloaded BW images, we devise a method to construct additional proxies using only them. As our Style branch is originally trained to classify 4 styles, we first fine-tune the proxy generation network on the newly collected unpaired dataset to classify 12 styles (RGB2Gray, expert A, B, C, and additional 8 styles). Next, we simply use our pre-trained decolorization network in Sec. 4.2. With the source vector and target proxy from the re-trained proxy generation network embedded to the encoder of decolorization network, we can obtain user-specified BW images for new styles. Through this experiment, we show the scalability of the proposed network, even the possibility of unpaired learning.

5. Experiments

In this section, we conduct quantitative evaluations to demonstrate the effectiveness of our DeColorful-Net, both with and without ground truth image pairs. We also compare its performance against state-of-the-art methods and popular commercial filters through a fair and extensive user-study.

Implementation Details. We train and evaluate DeColorful-Net on our dataset. We implement DeColorful-Net using Pytorch framework [43], utilize the Adam [32] optimizer with β_1=0.9 and β_2=0.999 and set a learning rate to 0.0001. On both the proxy generation and the decolorization stage, the number of learnable parameters is 31M and 15M, respectively. We split our dataset into training and test set with a ratio of 4 : 1. We train on images with a 512×512 resolution, and the training and inference take about 8 hours and 0.03 seconds on four NVIDIA RTX 3090 GPUs, respectively.

Model	PSNR↑	SSIM↑	LPIPS↓	Ablation	PSNR↑	SSIM↑	LPIPS↓
HDRNet [15]	23.93	0.909	0.072	$\mathcal{L}_1$	<u>27.34</u>	<u>0.947</u>	<u>0.059</u>
StarEnhancer [53]	25.50	0.924	0.104	$\mathcal{L}_2$	27.16	0.946	0.062
NeurOp [58]	25.08	0.925	0.067	$\mathcal{L}_2 + \mathcal{L}_{CE}$	27.21	**0.948**	0.062
MAXIM [55]	25.65	0.890	0.184	w/ $\mathcal{L}_{Triplet}$	26.35	0.943	0.063
RSFNet [45]	25.95	0.937	0.075	w/ $\mathcal{L}_{PL}$	26.00	0.939	0.065
CSNorm [61]	26.62	0.931	0.158	single-stage	25.14	0.931	0.067
Ours($\mathcal{L}_1 + \mathcal{L}_{CE}$)	**27.50**	<u>0.947</u>	**0.057**	w/o hierarchy	26.79	0.941	0.074

Table 1. Quantitative evaluation with other baselines(left) and ablation study for the effects of loss functions, hierarchical proxy-based loss, and multi-level bilateral grid(right). **Bold**: Best and <u>underlined</u>: the second best.

5.1. Quantitative Evaluation

In this evaluation, we verify how well our DeColorful-Net imitates the experts' retouching schemes. We thus use common quantitativ emeasures of image quality: PSNR, SSIM [59] and LPIPS [63], whose results are reported in Tab. 1.

We compare our DeColorful-Net with state-of-the-art image enhancement models: HDRNet [15], StarEnhancer [53], NeurOp [58], MAXIM [55], RSFNet [45] and CSNorm [61].

HDRNet directly predicts affine color transform coefficients in a bilateral grid, enabling both global and local adjustments. StarEnhancer presents a curve-based tone mapping technique that leverages embedding vectors, which serve as representatives of different styles, allowing users to transform the retouched images into their preferred style. NeurOp conducts color operators in the embedding space by estimating parameters to modify the embedding vectors. MAXIM employs multi-axis MLP structures to address the limitations of CNNs and transformers which are the absence of global receptive fields and suffer from quadratic complexity, respectively. RSFNet uses a white-box framework for image retouching by estimating a set of region masks and their parameters for pre-defined retouching functions. These parameters are able to modify attributes of images, such as temperature, hue, and exposure for each region. CSNorm proposes a channel selective normalization for lighting components of images. The normalization ensures the reconstruction quality because it prevents an information loss of learned features in the auto-encoder structure from a centralization process based on their mean and variance values in conventional image retouching.

We train them on our dataset from scratch with the authors' provided codes for user-preferred BW images. Note that since the methods except for StarEnhancer cannot train various styles at once, we separately train each expert's style.

Despite their impressive performance on color image

Age (Gender	PDecolor [6]	LeDecolor [39]	SPDecolor [38]	Style A	Style B	Style C	iPhone 14 Pro	Galaxy S21 Ultra	Instagram	Tiktok	Photoshop	BW Mode	Ours
20 (M)	2.07 ± 0.59	2.37 ± 0.76	2.82 ± 0.75	5.16 ± 0.60	3.48 ± 0.75	5.11 ± 0.54	4.18 ± 1.82	3.77 ± 1.74	3.25 ± 2.01	3.63 ± 1.97	4.34 ± 2.11	4.12 ± 2.08	4.71 ± 1.87
20 (F)	2.02 ± 0.44	2.59 ± 0.71	2.93 ± 0.65	5.12 ± 0.42	3.50 ± 0.71	4.83 ± 0.48	4.28 ± 1.86	4.30 ± 1.62	3.41 ± 2.90	3.35 ± 1.81	3.72 ± 2.08	4.15 ± 2.02	4.79 ± 2.04
30 (M)	2.79 ± 0.60	2.72 ± 0.66	2.95 ± 0.60	4.57 ± 0.55	3.45 ± 0.54	4.53 ± 0.53	3.98 ± 2.10	4.12± 1.68	3.63 ± 2.16	3.73 ± 1.85	3.80 ± 2.07	4.33 ± 2.04	4.41 ± 1.93
30 (F)	2.10 ± 0.48	2.53 ± 0.79	2.80 ± 0.74	5.09 ± 0.42	3.59 ± 0.73	4.88 ± 0.46	4.33 ± 1.78	4.33 ± 1.97	3.33 ± 1.89	3.89 ± 1.84	3.63 ± 2.19	3.56 ± 1.89	4.94 ± 1.92
40 (M)	2.54 ± 0.51	2.55 ± 0.70	2.84 ± 0.71	4.88 ± 0.55	3.46 ± 0.67	4.73 ± 0.66	3.86 ± 1.94	4.17 ± 2.04	3.91 ± 1.98	3.87 ± 1.88	3.72 ± 2.02	4.07 ± 2.04	4.40 ± 2.02
40 (F)	2.48 ± 0.50	2.65 ± 0.63	2.91 ± 0.69	4.81 ± 0.54	3.50 ± 0.58	4.65 ± 0.52	3.82 ± 1.99	4.14 ± 1.98	3.82 ± 1.98	3.93 ± 1.90	3.97 ± 2.11	3.98 ± 2.06	4.36 ± 1.91
50 (M)	2.53 ± 0.48	2.71 ± 0.58	3.02 ± 0.56	4.66 ± 0.54	3.45 ± 0.58	4.64 ± 0.49	4.31 ± 1.87	4.17 ± 1.91	3.30 ± 2.01	4.18 ± 1.82	3.58 ± 2.05	3.75 ± 2.04	4.72 ± 1.85
50 (F)	2.01 ± 0.46	2.36 ± 0.73	2.87 ± 0.76	5.22 ± 0.53	3.48 ± 0.76	5.06 ± 0.50	4.21 ± 1.92	3.95 ± 1.83	3.63 ± 2.05	3.92 ± 1.90	3.81 ± 2.08	4.15 ± 2.07	4.33 ± 2.03
Total	2.32 ± **0.38**	2.56 ± 0.62	2.89 ± 0.60	**4.94** ± 0.39	3.49 ± 0.58	4.80 ± 0.41	4.12 ± 1.92	4.11 ± **1.86**	3.54 ± 2.05	3.81 ± 1.89	3.82 ± 2.10	4.01 ± 2.04	**4.58** ± 1.96

Table 2. User Study 1 (Left) and Study 2 (Right). We report the mean and standard deviation of each result.

Figure 6. An example of the result in Study 1.

enhancement, the performance on BW images are unsatisfactory as depicted in Tab. 1. The reasons are mainly two folds: (1) The image-to-image translations, trained independently for each expert's style, focus on learning their global retouching skills, thereby overlooking intra-style variations based on the subject. Additionally, when they are trained separately for each subject and style, they show the worse results due to the insufficient number of training images. StarEnhancer, in contrast, which uses a simple ResNet [19] architecture and cross-entropy loss to train a style classifier, fails to account for the hierarchical relation of subject and style, which leads to the sub-optimal performances. (2) Some methods, specialized to the color image retouching, face challenges for BW images. RSFNet, for example, relies on color-based operations like temperature and hue adjustments. StarEnhancer heavily depends on the global manipulation by only tuning RGB curves. To produce aesthetic BW images, more local retouching schemes are required as discussed in [4, 13, 22].

5.2. Ablation Study

Loss functions. We use common distance metrics: $\mathcal{L}_1$ and $\mathcal{L}_2$. As widely known, since the $\mathcal{L}_1$ is helpful to yield sharp images than $\mathcal{L}_2$, which leads to higher PSNR and SSIM and lower LPIPS values. In addition, $\mathcal{L}_{CE}$ is beneficial to classify each expert's retouching scheme, which initially determines a luminance intensity level and local effects like edge enhancement of output images. Thus, we use both $\mathcal{L}_1$ and $\mathcal{L}_{CE}$ to train our DeColorful-Net. Furthermore, we train an embedding space using a triplet loss denoted as w/ $\mathcal{L}_{Triplet}$ in Tab. 1. Since our pipeline needs a proxy bank, we assign a random vector for each class and use it as an anchor, similar to PIE-Net [29]. Along with the lower performance, in a model design step, we decide not to use the triplet loss for two reasons: (1) Scalability: Due to the fact that the triplet loss compares each potential pair within the mini-batch, its time complexity increases at a rate of O(n^2), hindering the learning of diverse styles like internet photos Sec. 4.4. (2) Initialization Sensitivity: The model's performance depends heavily on the triplet samples of earlier steps. We train the model 4 times and could confirm the PSNR variances, from 26.01 dB to 26.35 dB.

Hierarchical proxy-based loss. When we use the $\mathcal{L}_{HPL}$, our DeColorful-Net produces subject-aware visually-pleasing results, even in the same style. This is because the loss term aggregates proxies with the same style and enforces to form a hierarchy between the style and subjects. On the other hand, to check whether the $\mathcal{L}_{HPL}$ is replaceable or not, we test a conventional $\mathcal{L}_{PL}$ [30]. As expected, we observe that the $\mathcal{L}_{PL}$ sometimes fails to imitate experts' style-aware retouching because it just spreads both the subject-style aware proxies into the embedding space without considering their hierarchical relation.

Two-stage training. We evaluate our two-stage training scheme and a single-stage training which skips the proxy generation step and directly trains the decolorization network in Sec. 4.2 for each expert's style. While our two-stage learning effectively differentiates styles in the embedding space with DML framework as shown in Sec. 4.3, the singe-stage learning suffers from a limited ability to extract the style characteristics, similar to the image-to-image translation models that can only learn one style at a time. Therefore, we believe that our two-stage training scheme is the effective way for capturing the style features as a whole.

Multi-level bilateral grid. We compare the single and multi-level bilateral grids. As mentioned in Fig. 5, the multi-level bilateral grid is necessary to implement the experts' coarse-to-fine retouching schemes. We verify the positive effect in this quantitative evaluation again. With only the single-level bilateral grid, posterization artifacts appear when we add spatial-variant effects like vignetting. In contrast, the multi-level bilateral grid alleviates the artifact by aggregating multi-scale local information of image patches.

5.3. User Study

To examine the aesthetic value of BW photos produced by our DeColorful-Net, we conduct a user study via Amazon MTurk. Considering that user preferences can greatly differ by gender and generation, we gather 20 users from each decade group - 20s, 30s, 40s, and 50s - with an equal number of male and female participants in each, 80 users in total. The questionnaire for the study consists of 3 sets: comparison with (1) state-of-the-art decolorization methods, (2) commercial BW filters with user preference (3) validation of additional proxies from internet photos.

Study 1: Comparison with State-of-the-art decoloriza-

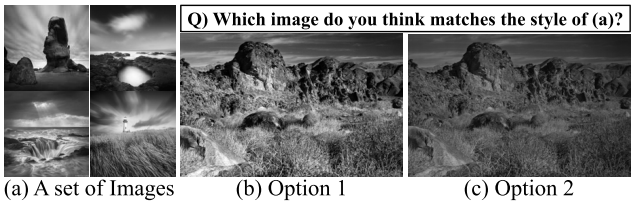
Figure 7. Comparison of ours with commercial filters used in Study 2.

tions. This study 1 is designed for the comparison of ours with the previous decolorization methods[1]. We randomly choose 20 image pairs from each class in our dataset, and a total of 80 image pairs are then presented to the users. The participants are shown in random order of 6 images, which are PDecolor [6], LeDecolor [39], SPDecolor [38], and ours with the expert-style A, B and C. According to the rankings given by the users, scores are assigned from 6 to 1 in a reciprocal order. As shown in Tab. 2, the results from our DeColorful-Net with the experts' styles are preferred in every group. In particular, the DeColorful-Net with the style A and C achieves the outstanding result on this user-study, whose example is displayed in Fig. 6.

Study 2: User preference. Next, we consider personal preferences of BW images by comparing ours with commercial BW filters in smartphones including Galaxy S21 Ultra, iPhone 14 Pro, Instagram, and Tiktok. Since each of them provides three types of BW filters, users can choose one of each that they prefer. Additionally, we use the neural filter in Photoshop, which can stylize images with a reference, to construct 3 additional sets based on each expert's style in our dataset. For a fair comparison, the experts A, B and C directly take both real-world color and BW images at the same spot using their own cameras. Note that they just use BW modes of their cameras to obtain the BW images. Each expert takes 12 photos such that 3 photos are captured for each subject class (a total of 36 photos used in this study). Before asking the user about their preferred image, we sort filtered images from the commercial filters and captured images from the experts by grouping them with similar styles together. In the beginning stage, the users first observe 3 sets of the sorted images and have to choose the most preferred style of them, similar to a user-study protocol in [29]. After that, the users watch 7 BW images in arbitrary order, consisting of ours, BW photography taken by BW modes of the experts' cameras, and the 5 filtered images (see Fig. 7).

Tab. 2 shows that our DeColorful-Net exhibits the highest preferences over those of the commercial filters and the experts' images. Based on this, we claim that DeColorful-Net produces personalized and visually-pleasing BW images. Surprisingly, the experts' photos fail to have a good score over the comparison methods. This is because there is a fundamental limitation of the color DSLR cameras which have narrower spectral ranges than monochrome cameras.

[1]The reason choosing them is their source codes are available in public.



(a) A set of Images (b) Option 1 (c) Option 2

Figure 8. An example of the questionnaire in Study 3. Obviously, the answer is "Option 1."

Age(Gender)	20 (M)	20 (F)	30 (M)	30 (F)	40 (M)	40 (F)	50 (M)	50 (F)	Total
Human	75.50	83.00	78.50	66.00	81.50	71.00	79.50	60.00	74.38
Non-human	80.50	69.00	73.50	72.00	81.00	66.00	70.00	76.00	73.50
Nature	84.50	81.50	85.00	65.50	78.50	80.00	75.00	66.50	77.06
Building	78.50	62.50	70.00	68.00	80.50	72.50	70.50	64.00	70.81
Total	79.75	74.00	76.75	67.88	80.38	72.38	73.75	66.63	73.94

Table 3. User Study 3. The percentage of voting result on each subject.

Study 3: Additional proxies from internet photos. We evaluate the scalability of our DeColorful-Net, described in Sec. 4.4. Let us denote two photographers for each subject as P1 and P2, as an example. The participants observe a set of images from P1, and then are asked to choose one of two retouched images. They have to find the most matched style of the given images (see Fig. 8). Note that these two image options are retouched from our DeColorful-Net using proxies learned with images from P1 and P2. We randomly pick 10 images from each photographer, a total of 80 image pairs presented to each user. Tab. 3 indicates that our DeColorful-Net remarkably mimics the distinct styles of professional photographers, even though the proxies are made using only internet downloaded images.

6. Conclusion

We present a novel decolorization DML framework to produce a visually-pleasing BW photography. We view decolorization as an aesthetic realm and handle the varying user preference issue. To do this, we collect large-scale black-and-white images, retouched by three professional photographers, from public colorful image datasets. Using our dataset, we train a DML framework with a hierarchical proxy-based loss to extract subject-style aware proxies. It enables us to imitate the experts' retouching schemes.

Acknowledgement This research was supported by 'Project for Science and Technology Opens the Future of the Region' program through the INNOPOLIS FOUNDATION funded by Ministry of Science and ICT (Project Number: 2022-DD-UP-0312), Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2021-0-02068, Artificial Intelligence Innovation Hub and No.2019-0-01842, Artificial Intelligence Graduate School Program (GIST)) and the Ministry of Trade, Industry and Energy (MOTIE) and Korea Institute for Advancement of Technology (KIAT) through the International Cooperative R&D program in part (P0019797).

References

[1] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *CVPR*, 2022. 2

[2] Nicolas Aziere and Sinisa Todorovic. Ensemble deep manifold similarity learning using hard proxies. In *CVPR*, 2019. 2

[3] Raja Bala and Reiner Eschbach. Spatial color-to-grayscale transform preserving chrominance edge information. In *CIC*, 2004. 2

[4] John Batdorff. *Black and White: From Snapshots to Great Shots*. Peachpit Pr, 2011. 7

[5] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input/output image pairs. In *CVPR*, 2011. 2, 3

[6] Bolun Cai, Xiangmin Xu, and Xiaofen Xing. Perception preserving decolorization. In *ICIP*, 2018. 1, 2, 7, 8

[7] Canva. Study: The most popular Instagram filters from around the world. https://url.kr/6tvkjf, 2021. 1

[8] Guangyi Chen, Yongming Rao, Jiwen Lu, and Jie Zhou. Temporal coherence or temporal motion: Which is more critical for video-based person re-identification? In *ECCV*, 2020. 2

[9] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *CVPR*, 2017. 2

[10] Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. Dovenet: Deep image harmonization via domain verification. In *CVPR*, 2020. 3

[11] Mengyu Dai and Haibin Hang. Manifold matching via deep metric learning for generative modeling. In *ICCV*, 2021. 2

[12] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 2

[13] Michael Freeman. *Mastering Black and White Digital Photography: A Lark Photography Book*. Union Square & Co, 2005. 7

[14] Michaël Gharbi, YiChang Shih, Gaurav Chaurasia, Jonathan Ragan-Kelley, Sylvain Paris, and Frédo Durand. Transform recipes for efficient cloud photo enhancement. *ACM TOG*, 34 (6):1–12, 2015. 5

[15] Michaël Gharbi, Jiawen Chen, Jonathan T Barron, Samuel W Hasinoff, and Frédo Durand. Deep bilateral learning for real-time image enhancement. *ACM TOG*, 36(4):1–12, 2017. 3, 4, 5, 6

[16] Amy A Gooch, Sven C Olsen, Jack Tumblin, and Bruce Gooch. Color2gray: salience-preserving color removal. *ACM TOG*, 24(3):634–639, 2005. 1, 2

[17] Kaiming He, Jian Sun, and Xiaoou Tang. Guided image filtering. *IEEE TPAMI*, 35(6):1397–1409, 2012. 5

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, 2015. 4

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3, 7

[20] Heiko Hirschmuller and Daniel Scharstein. Evaluation of stereo matching costs on images with radiometric differences. *IEEE TPAMI*, 31(9):1582–1599, 2008. 1

[21] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *SIMBAD*, 2015. 2

[22] Torsten Andreas Hoffmann. *The Art of Black and White Photography: Techniques for Creating Superb Images in a Digital Workflow*. Rocky Nook, 2012. 7

[23] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 2

[24] Kenneth Irby. Pulitzer Prize-winning photographer captures emotional, physical wounds from gang violence. https://url.kr/u91hzv, 2011. [Online; accessed 15-June-2011]. 1

[25] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 2

[26] Hae-Gon Jeon, Joon-Young Lee, Sunghoon Im, Hyowon Ha, and In So Kweon. Stereo matching with color and monochrome cameras in low-light conditions. In *CVPR*, 2016. 1

[27] Christopher Kanan and Garrison W Cottrell. Color-to-grayscale: does the method matter in image recognition? *PloS one*, 7(1):e29740, 2012. 1

[28] Scott Kelby. *Scott Kelby's Lightroom 7-Point System*. Rocky Nook, 2021. 5

[29] Han-Ul Kim, Young Jun Koh, and Chang-Su Kim. Pienet: Personalized image enhancement network. In *ECCV*, 2020. 2, 7, 8

[30] Sungyeon Kim, Dongwon Kim, Minsu Cho, and Suha Kwak. Proxy anchor loss for deep metric learning. In *CVPR*, 2020. 2, 3, 7

[31] Yongjin Kim, Cheolhun Jang, Julien Demouth, and Seungyong Lee. Robust color-to-gray via nonlinear global mapping. In *SIGGRAPH*, 2009. 1, 2

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

[33] Umut Kocasari, Alara Dirik, Mert Tiftikci, and Pinar Yanardag. Stylemc: Multi-channel based fast text-guided image generation and manipulation. In *WACV*, 2022. 2

[34] Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. Siamese neural networks for one-shot image recognition. In *ICMLW*, 2015. 2

[35] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip H. S. Torr. Manigan: Text-guided image manipulation. In *CVPR*, 2020. 2

[36] Qiegen Liu and Henry Leung. Variable augmented neural network for decolorization and multi-exposure fusion. *Information Fusion*, 46:114–127, 2019. 2

[37] Qiegen Liu, Peter X Liu, Weisi Xie, Yuhao Wang, and Dong Liang. Gcsdecolor: gradient correlation similarity for efficient contrast preserving decolorization. *IEEE TIP*, 24(9):2889–2904, 2015. 1, 2, 5

[38] Qiegen Liu, Peter Xiaoping Liu, Yuhao Wang, and Henry Leung. Semiparametric decolorization with laplacian-based

perceptual quality metric. *TCSVT*, 27(9):1856–1868, 2016. 2, 7, 8

[39] Qiegen Liu, Guangpu Shao, Yuhao Wang, Junbin Gao, and Henry Leung. Log-euclidean metrics for contrast preserving decolorization. *IEEE TIP*, 26(12):5772–5783, 2017. 2, 7, 8

[40] Deen Dayal Mohan, Nishant Sankaran, Dennis Fedorishin, Srirangaraj Setlur, and Venu Govindaraju. Moving in the right direction: A regularization for deep metric learning. In *CVPR*, 2020. 2

[41] Yair Movshovitz-Attias, Alexander Toshev, Thomas K Leung, Sergey Ioffe, and Saurabh Singh. No fuss distance metric learning using proxies. In *ICCV*, 2017. 2, 3

[42] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *CVPR*, 2016. 2

[43] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 6

[44] Qi Qian, Lei Shang, Baigui Sun, Juhua Hu, Hao Li, and Rong Jin. Softtriple loss: Deep metric learning without triplet sampling. In *ICCV*, 2019. 2

[45] Wenqi Quyang, Yi Dong, Xiaoyang Kang, Peiran Ren, Xin Xu, and Xuansong Xie. Rsfnet: A white-box image retouching approach using region-specific color filters. In *ICCV*, 2023. 6

[46] Yongming Rao, Jiwen Lu, and Jie Zhou. Attention-aware deep reinforcement learning for video face recognition. In *ICCV*, 2017. 2

[47] Yongming Rao, Jiwen Lu, and Jie Zhou. Learning discriminative aggregation network for video-based face recognition and person re-identification. *IJCV*, 127(6):701–718, 2019. 2

[48] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 2

[49] Michael Shaw. A Photo's Reach: Barbara Davidson Revisits the Navajo, Eddie Adams and Capa, Too. https://url.kr/bxwrmd, 2015. [Online; accessed 14-June-2015]. 1

[50] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 3

[51] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *NeurIPS*, 2016. 2

[52] Yibing Song, Linchao Bao, Xiaobin Xu, and Qingxiong Yang. Decolorization: Is rgb2gray () out? In *SIGGRAPH*, 2013. 1, 2, 5

[53] Yuda Song, Hui Qian, and Xin Du. Starenhancer: Learning real-time and style-aware image enhancement. In *ICCV*, 2021. 3, 4, 6

[54] Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. Deep image harmonization. In *CVPR*, 2017. 3

[55] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxim: Multi-axis mlp for image processing. In *CVPR*, 2022. 6

[56] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(11), 2008. 5

[57] Wei Wang, Zhengguo Li, and Shiqian Wu. Color contrast-preserving decolorization. *IEEE TIP*, 27(11):5464–5474, 2018. 1, 2

[58] Yili Wang, Xin Li, Kun Xu, Dongliang He, Qi Zhang, Fu Li, and Errui Ding. Neural color operators for sequential image retouching. In *ECCV*, 2022. 6

[59] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004. 6

[60] Zhibo Yang, Muhammet Bastan, Xinliang Zhu, Douglas Gray, and Dimitris Samaras. Hierarchical proxy-based loss for deep metric learning. In *WACV*, 2022. 2, 4

[61] Mingde Yao, Jie Huang, Xin Jin, Ruikang Xu, Shenglong Zhou, Man Zhou, and Zhiwei Xiong. Generalized lightness adaptation with channel selective normalization. In *ICCV*, 2023. 6

[62] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Learning a discriminative feature network for semantic segmentation. In *CVPR*, 2018. 1

[63] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6

[64] Wenliang Zhao, Yongming Rao, Ziyi Wang, Jiwen Lu, and Jie Zhou. Towards interpretable deep metric learning with structural matching. In *ICCV*, 2021. 2