One-shot Detail Retouching with Patch Space Neural Transformation Blending

Fazilet Gokbudak
fg405@cam.ac.uk
University of Cambridge
Cambridge, UK

Cengiz Oztireli
aco41@cam.ac.uk
University of Cambridge
Cambridge, UK

Figure 1: Our technique automatically transfers retouching edits to new images by learning the desired edits from one example *before-after* pair (insets). The transferred edits accurately capture intricate details such as wrinkles, dark spots, strands of hair, or eyelashes, as shown in the input (top) and retouched (bottom) pairs. Image courtesy of Jenavieve (top-left), Logan ProPro (top-left, inset), Marissa Oosterlee (top-middle). (CC-BY).

ABSTRACT

Photo retouching is a difficult task for novice users as it requires expert knowledge and advanced tools. Photographers often spend a great deal of time generating high-quality retouched photos with intricate details. In this paper, we introduce a one-shot learning based technique to automatically retouch details of an input image based on just a single pair of before and after example images. Our approach provides accurate and generalizable detail edit transfer to new images. We achieve these by proposing a new representation for image to image maps. Specifically, we propose neural field based transformation blending in the patch space for defining patch to patch transformations for each frequency band. This parametrization of the map with anchor transformations and associated weights, and spatio-spectral localized patches, allows us to capture details well while staying generalizable. We evaluate our technique both on known ground truth filters and artist retouching edits. Our method accurately transfers complex detail retouching edits.

CCS CONCEPTS

• **Computing methodologies** → **Computational photography**; **Image processing**.

KEYWORDS

detail retouching, image-to-image translation, context-aware image enhancement, one-shot learning, neural networks

ACM Reference Format:
Fazilet Gokbudak and Cengiz Oztireli. 2023. One-shot Detail Retouching with Patch Space Neural Transformation Blending. In *European Conference on Visual Media Production (CVMP '23), November 30–December 01, 2023, London, United Kingdom.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3626495.3626499

CVMP '23, November 30–December 01, 2023, London, United Kingdom

ACM ISBN 979-8-4007-0426-0/23/11.
https://doi.org/10.1145/3626495.3626499

1 INTRODUCTION

Photo retouching is often desirable as it improves the aesthetic quality of photographs by eliminating imperfections and highlighting subjects of interest. Even with significant progress in digital photography owing to advancements in camera sensors and image processing algorithms, professional retouches via manual adjustments are still needed to achieve a desired look. These artistic edits require considerable manual effort as they consist of global adjustments, such as brightening and contrast enhancement, as well as fine edits applied to local regions. Professionals spend a great deal of time to generate such edits, which motivates us to automatically mimic a specific style or type of retouch.

The development of automatic photo retouching tools can be helpful for both novice users and experts as it offers a basis for a professional retouching style. However, automating detailed edits of professionals is challenging as their editing pipelines are spatially varying, context-aware, and highly nonlinear, containing per-pixel adjustments. Recent learning-based methods address this complexity in image-to-image translation by proposing local context-aware methods, such as pixel-adaptive neural network architectures [Li et al. 2020; Shaham et al. 2021], learning parameters of local filters [Moran et al. 2020], or multi-stream models to extract global and local features separately [Gharbi et al. 2017]. However, these data-driven methods require a large dataset of matching example image pairs. Even then, the mappings are sensitive to segmentation errors, unseen semantic regions, and image content [Yan et al. 2016a].

Motivated by the gap between manual and automatic enhancement, we propose a novel photo retouching technique that can learn global and local adjustments from just *a single example image pair*. Our method thus sidesteps the need for large datasets, which are very difficult to obtain for the detail retouching task. We allow users to choose one example *before-after* pair from which our technique learns the underlying retouching style. Subsequently, we can apply the retouching edit to a different input image.

We assume that example and input images share similar local content. The user can thus decide on the semantics of the example and input photos and the structural changes to be transferred. This is easy for humans and practical for many scenarios, e.g. face edits transferred to faces. Our method then handles the difficult part for humans: capturing how fine details change in an edit and applying those automatically to a new image. The method can further be combined with brushes if fully automatic transfers are not desired.

We achieve these by defining the retouching problem as a map that is given by a *spatio-spectral patch-space neural field based transformation blending*. This representation is primarily inspired by professional detail retouching pipelines as we elaborate on in Section 3. Our map representation is composed of learned patch maps at multiple scales, i.e. frequency bands. Each of these maps is represented by a number of *transformation matrices* blended with *patch-adaptive weights* that are represented as neural fields. We jointly optimize the transformation matrices and corresponding weights for each band. This representation captures edits to details better than any previous techniques while staying generalizable to new images. It is also simple enough to be extended in many different ways in future works.

In summary, there are two main contributions of this work:

- **A novel patch-space image map representation** as a blending of transformation matrices with neural fields.
- **A one-shot detail retouching algorithm** that allows transfer of edits to details to new images based on a single before-after image pair.

2 RELATED WORK

Photo retouching has been explored in image processing and computer vision communities under different domains, such as photo enhancement and image-to-image translation. Below, we first discuss recent methods on photo enhancement and then image to image map definitions with the main focus on learning-based methods.

2.1 Digital Photo Enhancement

Global image enhancement. Color and tone transfer has been considered a very effective technique to improve the perceptual quality of photos with pre-defined rules or examples [Faridul et al. 2014; Mustafa et al. 2022]. Earlier methods typically apply global changes and adjust image statistics [Bae et al. 2006; Bychkovsky et al. 2011; He et al. 2020; Park et al. 2018; Pitie et al. 2005; Pitié et al. 2007; Reinhard et al. 2001; Sunkavalli et al. 2010], e.g., mean and standard deviation, without considering image content and local variations [Cohen-Or et al. 2006]. These methods generally transfer color changes, ignoring edits in fine details. On the other hand, our method learns a mapping per frequency band, capturing transfers even in high frequencies. Bychkovsky et al. [Bychkovsky et al. 2011] collected the MIT-Adobe FiveK dataset of 5,000 photographs and their retouched versions by five artists. The authors propose a regression model to learn artists' retouching styles from before-retouched pairs. Chen et al. [Chen et al. 2017] introduce a fully-convolutional neural network model to learn global image processing operators, such as photographic style, nonlocal dehazing, and pencil drawing. In [Hu et al. 2018], a photo retouching pipeline for various post-processing operations is presented, where global adjustment curves are approximated. The authors suggest a deep reinforcement learning approach to model users' edit preferences from a given photo collection.

Nevertheless, global transfers cannot capture local and regional variations in a photo [Cohen-Or et al. 2006].They may result in artifacts when the local target regions of the example and input images do not match. We adapt our mappings to each image patch separately, thus accurately capturing local edits in intricate details.

Local context-aware image enhancement. To capture local variations, different methods have been proposed, such as learning local representative color transform [Kim et al. 2021], estimating an image-to-illumination mapping with a local feature extractor [Wang et al. 2019], local histogram matching [Shapira et al. 2013], segmentation [Laffont et al. 2014; Tai et al. 2007], combining and learning pre-defined filters [Berthouzoz et al. 2011; Chen et al. 2018a; Huang et al. 2014; Omiya et al. 2018; Saeedi et al. 2018] or with further user guidance [An and Pellacini 2010; Pouli and Reinhard 2011; Tai et al. 2005], detection or learning of image semantics and context [Gharbi et al. 2017; Hwang et al. 2012; Kaufman et al. 2012; Nam and Kim 2017; Yan et al. 2014; Zhu and Yu 2018], matching

[HaCohen et al. 2011], or precise alignment [Kagarlitsky et al. 2009; Shih et al. 2013].

Furthermore, recent work has focused on learning global and local adjustments via spatially-varying filters [Chen et al. 2018b; Gharbi et al. 2017; Li et al. 2020; Moran et al. 2020; Shaham et al. 2021]. Chen et al. [Chen et al. 2018b] introduce a global feature extraction layer along with per-pixel adjustments to enhance photos. Bilateral guided joint upsampling [Chen et al. 2016] also allows for local and global image processing with an encoder-decoder approach. HDRNet [Gharbi et al. 2017] learn content-aware, global, and local adjustments via a two-stream convolutional architecture, which extracts local and global features separately to fit local affine transformations and encode the high-level description of images, respectively. Also, Moran et al. [Moran et al. 2020] propose to learn the parameters of three different spatially local filters to automatically enhance photos.

Local color and tone adjustments might still be insufficient to capture intricate details [Bae et al. 2006]. Transfer of such details, in general, requires a dense matching [HaCohen et al. 2011] or alignment between example and input images [Shih et al. 2014]. To achieve either dense matching or alignment, methods constrain their datasets to contain very similar example and input images, for example, faces with similar characteristics and views [Shih et al. 2014]. On the other hand, our method does not require dense correspondences between input and example images but still transfers intricate details. It accurately represents such complex mappings with an operator summing the effects of various transformations multiplied with corresponding patch-adaptive weights, applied at multiple frequency bands.

Differentiable image processing pipelines. To have more flexibility and control over the rendering process, methods based on image signal processors (ISP)s have been proposed to enhance photos. In both [Tseng et al. 2019; Yu et al. 2021], hyperparameters of an ISP are optimized. Different from [Tseng et al. 2019], which only applies to a fixed pipeline, Yu et al. [Yu et al. 2021] can explore different ISP architectures. Furthermore, Tseng et al. [Tseng et al. 2022] model a commercial raw processing pipeline with a series of neural networks to render sRGB images from raw inputs. As we assume example and input images to be processed RGB images rather than raw data, we refrain from comparing our method with such ISP-based methods.

2.2 Defining Maps between Images

Unsupervised methods. Some learning-based techniques only require one or more examples of retouched photos without their before examples to learn the transfer. Such unsupervised methods capture a certain style by decomposing images into a reflection map and an illumination map [Ma et al. 2021], extracting and recomposing band representations of training images [Yang et al. 2020], regularizing unpaired training using information extracted from the input [Jiang et al. 2021], segmenting the image into semantic regions [Liu et al. 2016], adaptive image regions [Frigo et al. 2016], learning semantic and global features [Chen et al. 2018a], progressively translating image from coarse to fine via pyramids of generative models [Lin et al. 2020], or utilizing artistic principles and pre-defined filters [Hu et al. 2018; Zhang et al. 2013]. These methods transfer pre-defined elements of the desired style, or global color and tone. Defining the desired style and the content of the input image that is to remain is challenging. Hence, these methods typically assume prior knowledge of the type of desired adjustments. Even then, capturing the retouching edits in details remains out of scope since these methods are typically designed for domain transfer, working on high level features of images.

As a weakly supervised method, Liao et al. [Liao et al. 2017] propose a technique based on image analogy [Hertzmann et al. 2001] to transfer the visual attributes, such as color, tone, texture, and style, across images that look very different but share similar semantic structures. Similar to our method, they also work with one example pair (A and B') to transfer the attributes. However, they rather focus on high-level features, disregarding the edits in intricate details.

Supervised methods. For a conceivable representation, many supervised transfer methods require a large dataset of well-aligned example image pairs whose contents are very similar [Kim et al. 2021; Wang et al. 2019]. However, finding or generating such a dataset is difficult as the content of images can change dramatically. Even with such a dataset, segmentation errors, unseen semantic regions, or image content can still change the results significantly [Yan et al. 2016b]. In contrast, our method allows users to choose the example pairs from which the desired style is learned, hence sidestepping the challenging semantics problem. Similar content and structures between example and input images lead to more natural transfers.

Convolutional neural networks (CNNs) are the de-facto model for image processing with supervised learning methods. While CNNs present state-of-the-art results in computer vision tasks, they are not required [Tolstikhin et al. 2021]. MLP-based architectures have recently gained popularity in image classification and image-to-image translation. Cazenavette and De Guevara [Cazenavette and De Guevara 2021] propose the MLP-Mixer architecture that only uses simple MLP blocks to learn image classification. Cazenavette and De Guevara [Cazenavette and De Guevara 2021] also show an application of an MLP-based architecture for image synthesis. They adapt the MLP-Mixer architecture [Tolstikhin et al. 2021] to perform unpaired image-to-image translation. Our observation that MLP-based architectures attain competitive results in challenging vision tasks motivated us to explore the use of an MLP block as an alternative to CNNs in the context of photo retouching.

3 OVERVIEW AND MOTIVATIONS

Given a pair of example images X and Y, we aim to learn a map M such that $Y = M(X)$. The learned map can then be applied to a new input image I to obtain the retouched output $O = M(I)$.

To define this map, we first decompose the example images into multiple feature maps X_l, Y_l capturing details at different scales, such as coefficients at different bands of a Laplacian pyramid. We then define a separate mapping M_l for each X_l, Y_l pair in the patch space as a blending of transformation matrices with neural field based weights, all learned jointly. We illustrate the overall map representation in Figure 2.

For transfer of edits, the M_l are computed and applied to each patch of the decomposition I_l of an input image I to obtain the

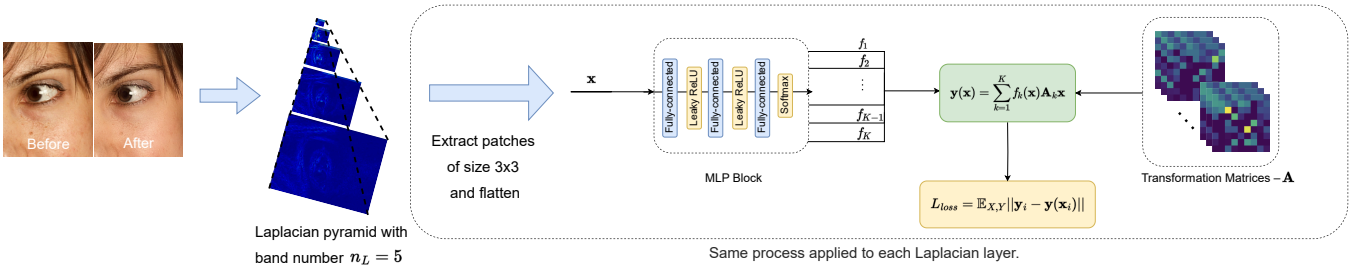

Figure 2: Our technique learns a separate mapping per frequency band by decomposing images into five different bands with a Laplacian pyramid. At each Laplacian band l, we define a mapping between flattened patches $\mathbf{x}_i$, $\mathbf{y}_i$ extracted from before-after bands X_l, Y_l. Our field based method (MLP block) adapts transformations to input patches, providing local context-aware adjustments. The transformation matrices and MLP parameters are learned jointly from scratch for each Laplacian band of the before-after pair.

corresponding output patch of O_l. The patches are finally placed at their spatial locations and averaged to reconstruct each image O_l , which are summed to get the output image O.

Our motivation behind designing such a map representation with frequency decomposition and transformation blending comes from studying the nature of retouching edits. First, artists often decompose images into different frequency bands to have better control over structural and textural edits to details. Second, image patches of similar content, e.g. skin or hair, are retouched similarly. This means similar patches in the patch space translate into similar edits. Our representation leads to a different transformation for patches of differing content. Third, these edits are typically applied via brushes for smooth transitions. Our neural field based blending allows for such smooth interpolation, mimicking such brush strokes.

Although we are inspired by professional artist pipelines, we illustrate in the next sections that this new image to image mapping representation can replicate the effect of and transfer edits for many filters.

4 ONE-SHOT RETOUCHING

4.1 Frequency Decomposition

We first decompose example and input images into different frequency bands by constructing a Laplacian pyramid to capture details at multiple scales. In principle, it is possible to utilize any multiscale image decomposition method. However, we observed that a basic Laplacian pyramid helped us capture more accurate and generalizable results compared to a guided or bilateral pyramid. Therefore, we decompose images by

$$X_l = L_l(X) = \begin{cases} X - G(2) * X & l = 0 \\ G(2^l) * X - G(2^{l+1}) * X & l > 0, \end{cases} \quad (1)$$

where $G(\sigma)$ is the normalized Gaussian kernel, and $*$ denotes convolution. We also store the low-pass filtered image $S(X)$ such that $X = S(X) + \sum_{l=0}^{n_L} L_l(X)$. We then downsample each $L_l(X)$ and $S(X)$ according to the maximum frequency present at that band. This allows us to use small 3×3 patches at each band. In our experiments, we used $n_L = 5$ bands for the Laplacian pyramid.

Since each band is processed independently, we explain the steps of our technique below for two generic images X and Y.

4.2 Transformation Blending

The mapping is defined between patches $\mathbf{x} \in \mathbb{R}^{d_X}$ to $\mathbf{y} \in \mathbb{R}^{d_Y}$ extracted from X and Y, respectively, where we denote the patches with vectors stacking the pixel values and define the patch spaces as $\mathbb{R}^{d_X}$ and $\mathbb{R}^{d_Y}$. For all results in this work, we work with 3×3 patches and thus $d_X = d_Y = 9$.

Our mapping takes the form of a weighted average of learned transformation matrices $\mathbf{A}$, where each transformation matrix $\mathbf{A}_k$ is first multiplied with its corresponding blending weight:

$$\mathbf{y}(\mathbf{x}) = \sum_{k=1}^{K} f_k(\mathbf{x})\mathbf{A}_k\mathbf{x}, \quad (2)$$

Here, K is the number of transformation matrices, and f_k are the blending weights, learned by an MLP block of output size K. The $\mathbf{A}_k$'s and f_k's are jointly learned by minimizing the following loss on patches extracted from the before and after images.

$$L_{loss} = \mathbb{E}_{X,Y}||\mathbf{y}_i - \mathbf{y}(\mathbf{x}_i)|| \quad (3)$$

Each $\mathbf{A}_k$ corresponds to a different type of transformation and the $f_k(\mathbf{x})$'s, represented with the MLP, allow for a smooth transition between different transformations. The form of f_k's is relatively simple with three fully-connected layers and nonlinear activation functions applied after each layer. This blending forms a simple but expressive transformation as we illustrate in the Section 5.

4.3 Retouching an Input Image

We process the input image I the same way as the before-after pair. First, we decompose the input into its Laplacian layers and then extract its patches per layer. After applying the learned mappings M_l to the patches of the corresponding layers $L_l(I)$ independently, we then reconstruct the Laplacian bands of the output image O_l by placing the patches at their spatial locations and averaging over the overlapping regions. Later, we obtain the final output image O

by summing the outputs O_l and the residual of the input image:

$$O = S(I) + \sum_{l=0}^{n_L} M_l(L_l(I)) \,. \tag{4}$$

4.4 Implementation

Patch size and stride. In order to capture each frequency band at the right level of detail, we do not upsample the images $L_l(X)$ and use a small 3×3 patch size (with stride 1). We experimented with larger patch sizes. However, this turned out to be counterproductive for the detail level we target, as details are blurred in larger patches. They also lead to overfitting and are harder to optimize for in general. We used a stride of 1, and hence patches overlap on the image plane. The overlapping patches are averaged while reconstructing the image.

Detail and color modifications. We aim to capture intricate details present in highly detailed retouches and a wide range of image processing operators. Based on the observation that various operators can edit materials in the image space using the luminance component [Boyadzhiev et al. 2015], we focus on learning changes in luminance while preserving the input chrominance channels.

Evaluation metrics. To quantitatively compare our method with state-of-the-art methods, we used PSNR and SSIM metrics. This is only possible if the before-after image pair was processed with a known, reproducible operator (see Section 5.3 for details).

Training details. We train different mappings with the same structure, defined in Equation 2, for each frequency band of the Laplacian pyramid. Each mapping consists of one MLP block and K number of transformation matrices, which are learned jointly per frequency band from scratch for each before-after pair. The MLP block employed in our experiments consists of three fully-connected layers and non-linearities applied after each layer. The output size of the last layer is the same as the number of transformation matrices.

To normalize the weights, we chose the last activation function to be Softmax, while for the first two layers, we apply Leaky ReLU. Each transformation matrix is randomly initialized with uniform distribution in the range $[0, 1]$. All experiments use the Adam optimizer with a learning rate of 10^{-2}, which exponentially decays with a decay rate of 0.96. We use l_1 loss function in all our experiments. Through backpropagation, both the MLP parameters and the entries of the transformation matrices are learned at the same time.

5 RESULTS

5.1 Ablation Study

The success of our learned mappings relies on two key components: patch-adaptive retouching and transformation blending. We thus conduct experiments to illustrate the significance of these.

Transformation Matrices. We compared transformation matrices of size 9×9 with scalar values. The method still remained spatially-varying, since we left the MLP the same, and used $K = 256$ scalar weights. We tested both methods on 100 images and computed average PSNR values. We observed that our technique with matrices performed better than scalar values even in simple algorithmic filters, such as Gaussian and Unsharping Masking (around 2 dB and 3 dB higher PSNRs, respectively).

As the complexity of a retouching style depends on multiple factors, such as artists' design choices, user preferences, or the artist toolbox, it is challenging to analyze such effects on retouching examples quantitatively. For simple algorithmic filters, such as a Gaussian filter or unsharp masking, $K = 1$ can sufficiently reproduce the filter. In contrast, more complex algorithms, such as a bilateral filter, require more matrices to capture the algorithmic edits accurately (Figure 3). Since retouching edits combine the effect of multiple operators and are highly non-linear, we empirically chose $K = 256$ for our retouching examples.

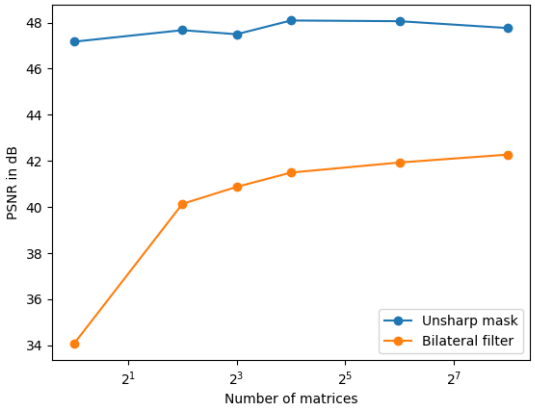

Figure 3: The higher the complexity of the learned algorithm, the more transformation matrices our technique requires to capture the effects on local regions accurately. While $K = 1$ can be sufficient for our model to capture unsharp masking, it requires more matrices to represent bilateral filtering precisely.

Patch-adaptive Transformation Blending. We also compared our patch-adaptive mapping to an MLP regressor on the extracted patches. This directly learns the mapping from the decomposition of example before-after images instead of utilizing blended transformations. The MLP regressor follows a similar architecture as our MLP block (Figure 2), with the only difference being the last activation function. We used Leaky ReLU here, since the Softmax function outputs pseudo-probabilities and is unsuitable for regression. Not explicitly handling the spatially-varying structure of the mapping and directly regressing limits the expressiveness of the model. This results in blurry results as shown in Figure 4 because such a model cannot capture edits in intricate details, such as highlights around eyes and hair or brightening of the skin. We also tried increasing the capacity of the MLP regressor but did not observe much improvement in performance.

5.2 Qualitative Results

We tested our technique on a diverse range of before-after pairs, including face images from the **FFHQ** dataset [Karras et al. 2021]. We focus on human portraits and face retouching in our experiments as they are arguably the most common and prioritized types

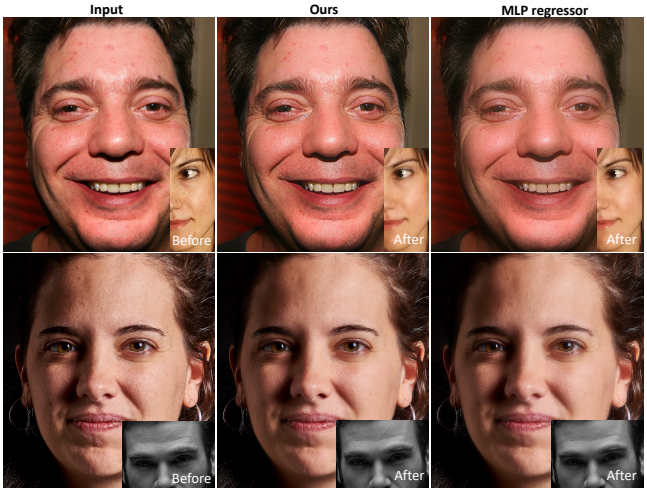

Figure 4: An MLP regressor cannot capture local edits, resulting in inaccurate retouching edits, such as blurring on the skin or around the eyes.

of photos for retouching. We also illustrate that our technique provides visually pleasing results in different types of images, such as materials or rooms, and accurately captures image processing filters.

Figure 5: The reproduced retouching style from the example pair (inset) improves skin texture without affecting fine details, such as eyes and hair, for a visually improved portrait. Moreover, our technique generalizes well to faces with different lighting conditions and accurately reproduces the example retouching style.

Human faces pose a particular challenge for our technique. However, our model can still capture highly nonlinear retouching edits and generalizes well to different types of faces, view directions, and lighting conditions, as illustrated in Figures 1, 5, and 7.

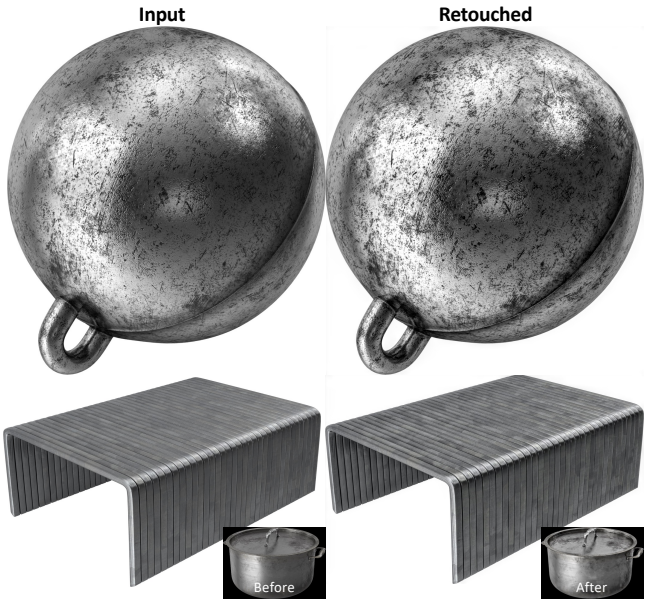

Figure 6: Material editing results on photos (left), and rendered images (right), based on the before-after pair (inset). The details, such as scratches or lines are emphasized, and materials became shinier. Image courtesy of royalmix (top and bottom-inset), tsmdunn (bottom). (PixelSquid).

The example pairs in Figures 1, 5 and 7 were generated by brushing onto the skin with artist created brushes, eye sharpening (sharpening example in Figure 1), and further brightness/contrast adjustments. These brushes first decompose the skin into a detail and base layer, typically with frequency decomposition, alter the detail layer and blend it with the base layer. They differ in how (1) they decompose the skin into the layers, i.e., what frequencies are in each layer, and (2) they edit and blend each layer with different opacity values. This variation creates retouching nuances, as shown in Figure 7. Our method can still accurately capture such slight differences in styles.

In all our experiments, intricate details of the desired retouching, such as small-scale texture, eye, facial hair or material details, and global features, such as overall lighting and tone, are accurately reproduced. It is interesting to observe that the *glamour* implied by, e.g., the example retouching in Figure 5 is transferred from the example pair very accurately without causing an artificial look. Zooming into the skin reveals that pores and wrinkles are minimized, and the blemishes and discoloring of the skin are eliminated. At the same time, depending on the retouching edit, details, such as eyes or material texture, are more highlighted or preserved, and delicate features such as hair are preserved well (Figures 5, 6 and 7).

In summary, our technique efficiently edits such intricate details, due to the significantly distinct local statistics of the texture at

Figure 7: Our patch-adaptive technique can accurately capture the nuances between different retouching styles as given by the examples (top row).

multiple scales, without affecting overlaying structures thanks to its spatially-varying nature and frequency decomposition.

5.3 Comparison with the state-of-the-art

Although there are various works related to automatic photo enhancement, to the best of our knowledge, none of them works with a single example pair for detail retouching. We thus compare our results with closely related automatic image-to-image translation methods, namely U-Net [Ronneberger et al. 2015], ASAPNet generator [Shaham et al. 2021], and Deep edge-aware filters [Xu et al. 2015].

We trained each network from scratch with one *before-after* pair. To train the U-Net architecture, we changed the activation function of its last layer to ReLU and used l_1 loss function with Adam optimizer (same as ours). Similar to our method, ASAPNet is also a spatially-adaptive network. However, it is instead designed to hallucinate new details. Therefore, we similarly trained their generator model to ours with $l1$ loss, removing the discriminator. We observed that bilinear downsampling in their model causes checkerboard artifacts. Hence, we also removed this operator and learned an MLP per pixel, which caused the model to be highly complex with too many parameters.

For a fair comparison with contemporary methods, we trained each network with the same example pair processed by four algorithmic filters: Gaussian, unsharp masking, Bilateral, and local Laplacian filters (LLF). As LLFs can perform a wide range of edge-aware operations, we apply two different versions of the filter, one for smoothing ($\alpha = 2, \sigma = 0.2$) and one for enhancing details ($\alpha = 0.5, \sigma = 0.1$). Each network is trained from scratch with the same example pair resized to 256×256 for the corresponding filter.

To prove the generalizability of our technique, we tested the models on different types of images, namely face images (100 images that are randomly sampled from MIT-Adobe FiveK [Bychkovsky et al. 2011]), material images (22 images), room images (30), and landscape images (30). Each type was trained separately with its corresponding example pair. For instance, we trained an example pair of landscape images to test our model on landscape images. We evaluated the models using average PSNR and SSIM values. To generate the ground truths of the input images, we applied the same filter as applied to the before example image to obtain the after image. We trained each model in Y-channel after converting RGB images to their YCbCr versions and evaluated the results for Y-channel images. We duplicated the Y-channel in case the model requires three-channel images.

To obtain the UNet results for each type of images, we ran an additional experiment in which we changed the number of trainable parameters by removing some layers and trained the network from scratch for unsharp masking and bilateral filtering. The number of parameters we chose were 0.1M (with a few convolutional layers), 1.8M, 10M and 30M. For material images, we observed that 10M performed the best in terms of PSNR and SSIM values, while for other types of images 30M performed best. We tested the trained models on the images of the corresponding types and computed average PSNR values. Later, we chose the model with the best-performing parameters for each type of image for the quantitative comparison (Table 1).

Overall, our method can outperform all architectures for each considered filter in terms of PSNR values. UNet shows the closest performance to our method, but their network capacity is significantly higher than ours (0.16M). As the filter becomes more complex, the performance gap increases. For instance, other methods perform fairly well in simple algorithmic filters, such as Gaussian or unsharp masking. However, in spatially-varying filters, namely Bilateral filter or LLFs, our method proves more generalizable thanks to its path-adaptive structure. Figure 8 further demonstrates that compared to the state-of-the-art methods, our algorithm can preserve and edit intricate details, such as text, texture, or leaves, more effectively without causing much distortion.

5.4 Limitations and Future Work

A primary limitation of our work is its dependence on local patches at different scales, disregarding their spatial location. Hence, our method is most useful when details are retouched based on local and repeated characteristics of an image. Non-repeating spatially-dependent strong effects, e.g., tattoos or portrait stylizations with spatially varying lighting [Shih et al. 2014], cannot be handled by the current technique (see Figure 9). We leave this as future work.

Since we rely on a single example image pair, transferring filters applied to arbitrary images [Yan et al. 2014] is out of the scope of our current work. We require example and input images to have similar semantics for predictable transfer. Extending the technique to more than one pair of example images will require us to have consistently retouched details on all those example images. Finally, we require the example before and after images to be perfectly aligned. This requirement can be alleviated by incorporating an

Table 1: Quantitative performance comparison for the reproduction of various image processing filters. Average PSNR and SSIM values are computed over 182 images of different types of images including faces, landscapes, materials, and rooms. Qualitative results can be found in the supplementary material.

Comparison results (PSNR in dB / SSIM)				
Filter Type	**ASAPNet Generator**	**Deep Edge-aware**	**UNet**	**Ours**
Gaussian	39.36 / **0.983**	36.52 / 0.979	40.52 / 0.979	**40.67 / 0.983**
Unsharp Mask	29.77 / 0.889	32.62 / **0.959**	32.05 / 0.919	**33.88** / 0.931
Bilateral Filter	33.65 / 0.936	33.56 / 0.958	34.00 / 0.939	**38.16 / 0.965**
Local Laplacian ($\alpha = 2$, $\sigma = 0.2$)	30.85 / 0.913	30.69 / 0.943	31.53 / 0.925	**33.50 / 0.950**
Local Laplacian ($\alpha = 0.5$, $\sigma = 0.1$)	31.98 / 0.909	31.62 / 0.931	33.08 / 0.929	**35.72 / 0.940**

ICP [Besl and McKay 1992]-like approach into the optimization in Section 4.

Although our main focus in this paper is on artist-driven subjective retouching edits, the proposed technique is general. It can be applied to summarize and transfer arbitrary image transformations, significantly where details are modified. We are thus planning to investigate our technique further as a general transfer method for image-to-image translation. The patch-adaptive nature of our mappings makes them amenable to analysis.

6 CONCLUSIONS

We presented a neural field based technique for example-based automatic retouching of images. By formulating the transfer problem in the patch space, we showed that blending multiple transformation matrices with patch-adaptive weights can be utilized to learn an accurate and generalizable map. This allowed us to use images of different scenes, people, views, and environmental conditions as the example pair and input. We illustrated the technique's utility on various retouching examples. We believe that our image map representation can be helpful in many other image processing tasks.

ACKNOWLEDGMENTS

This work was supported by a UKRI Future Leaders Fellowship [grant number G104084].

REFERENCES

Xiaobo An and Fabio Pellacini. 2010. User-Controllable Color Transfer. *Computer Graphics Forum* 29, 2 (2010), 263–271. https://doi.org/10.1111/j.1467-8659.2009.01595.x

Soonmin Bae, Sylvain Paris, and Frédo Durand. 2006. Two-scale Tone Management for Photographic Look. *ACM Trans. Graph.* 25, 3 (July 2006), 637–645. https://doi.org/10.1145/1141911.1141935

Floraine Berthouzoz, Wilmot Li, Mira Dontcheva, and Maneesh Agrawala. 2011. A Framework for Content-adaptive Photo Manipulation Macros: Application to Face, Landscape, and Global Manipulations. *ACM Trans. Graph.* 30, 5, Article 120 (Oct. 2011), 14 pages. https://doi.org/10.1145/2019627.2019639

Paul J. Besl and Neil D. McKay. 1992. A Method for Registration of 3-D Shapes. *IEEE Trans. Pattern Anal. Mach. Intell.* 14, 2 (Feb. 1992), 239–256. https://doi.org/10.1109/34.121791

Ivaylo Boyadzhiev, Kavita Bala, Sylvain Paris, and Edward Adelson. 2015. Band-Sifting Decomposition for Image-Based Material Editing. *ACM Trans. Graph.* 34, 5, Article 163 (Nov. 2015), 16 pages. https://doi.org/10.1145/2809796

V. Bychkovsky, S. Paris, E. Chan, and F. Durand. 2011. Learning Photographic Global Tonal Adjustment with a Database of Input/Output Image Pairs. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '11)*. IEEE Computer Society, Washington, DC, USA, 97–104. https://doi.org/10.1109/CVPR.2011.5995413

George Cazenavette and Manuel Ladron De Guevara. 2021. MixerGAN: An MLP-Based Architecture for Unpaired Image-to-Image Translation. *arXiv preprint arXiv:2105.14110* (2021).

Jiawen Chen, Andrew Adams, Neal Wadhwa, and Samuel W Hasinoff. 2016. Bilateral guided upsampling. *ACM Transactions on Graphics (TOG)* 35, 6 (2016), 1–8.

Qifeng Chen, Jia Xu, and Vladlen Koltun. 2017. Fast image processing with fully-convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*. 2497–2506.

Yu-Sheng Chen, Yu-Ching Wang, Man-Hsin Kao, and Yung-Yu Chuang. 2018a. Deep Photo Enhancer: Unpaired Learning for Image Enhancement From Photographs With GANs. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yu-Sheng Chen, Yu-Ching Wang, Man-Hsin Kao, and Yung-Yu Chuang. 2018b. Deep photo enhancer: Unpaired learning for image enhancement from photographs with gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6306–6314.

Daniel Cohen-Or, Olga Sorkine, Ran Gal, Tommer Leyvand, and Ying-Qing Xu. 2006. Color Harmonization. *ACM Trans. Graph.* 25, 3 (July 2006), 624–630. https://doi.org/10.1145/1141911.1141933

H. S. Faridul, T. Pouli, C. Chamaret, J. Stauder, A. Tremeau, and E. Reinhard. 2014. A Survey of Color Mapping and its Applications. In *Eurographics 2014 - State of the Art Reports*, Sylvain Lefebvre and Michela Spagnuolo (Eds.). The Eurographics Association. https://doi.org/10.2312/egst.20141035

Oriel Frigo, Neus Sabater, Julie Delon, and Pierre Hellier. 2016. Split and Match: Example-based Adaptive Patch Sampling for Unsupervised Style Transfer. (March 2016). Peer-reviewed paper accepted to be presented at IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), Jun 2016, Las Vegas, United States..

Michaël Gharbi, Jiawen Chen, Jonathan T. Barron, Samuel W. Hasinoff, and Frédo Durand. 2017. Deep Bilateral Learning for Real-time Image Enhancement. *ACM Trans. Graph.* 36, 4, Article 118 (July 2017), 12 pages. https://doi.org/10.1145/3072959.3073592

Yoav HaCohen, Eli Shechtman, Dan B. Goldman, and Dani Lischinski. 2011. Non-rigid Dense Correspondence with Applications for Image Enhancement. *ACM Trans. Graph.* 30, 4, Article 70 (July 2011), 10 pages. https://doi.org/10.1145/2010324.1964965

Jingwen He, Yihao Liu, Yu Qiao, and Chao Dong. 2020. Conditional sequential modulation for efficient global image retouching. In *European Conference on Computer Vision*. Springer, 679–695.

Aaron Hertzmann, Charles E. Jacobs, Nuria Oliver, Brian Curless, and David H. Salesin. 2001. Image Analogies. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '01)*. ACM, New York, NY, USA, 327–340. https://doi.org/10.1145/383259.383295

Yuanming Hu, Hao He, Chenxi Xu, Baoyuan Wang, and Stephen Lin. 2018. Exposure: A White-Box Photo Post-Processing Framework. *ACM Trans. Graph.* 37, 2, Article 26 (May 2018), 17 pages. https://doi.org/10.1145/3181974

Shi-Sheng Huang, Guo-Xin Zhang, Yu-Kun Lai, Johannes Kopf, Daniel Cohen-Or, and Shi-Min Hu. 2014. Parametric Meta-filter Modeling from a Single Example Pair. *Vis. Comput.* 30, 6-8 (June 2014), 673–684. https://doi.org/10.1007/s00371-014-0973-y

Sung Ju Hwang, Ashish Kapoor, and Sing Bing Kang. 2012. Context-based Automatic Local Image Enhancement. In *Proceedings of the 12th European Conference on Computer Vision - Volume Part I* (Florence, Italy) *(ECCV 2012)*. Springer-Verlag, Berlin, Heidelberg, 569–582.

Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. 2021. EnlightenGAN: Deep Light Enhancement Without Paired Supervision. *IEEE Transactions on Image Processing* 30 (2021), 2340–2349. https://doi.org/10.1109/TIP.2021.3051462

S. Kagarlitsky, Y. Moses, and Y. Hel-Or. 2009. Piecewise-consistent color mappings of images acquired under various conditions. In *Computer Vision, 2009 IEEE 12th*

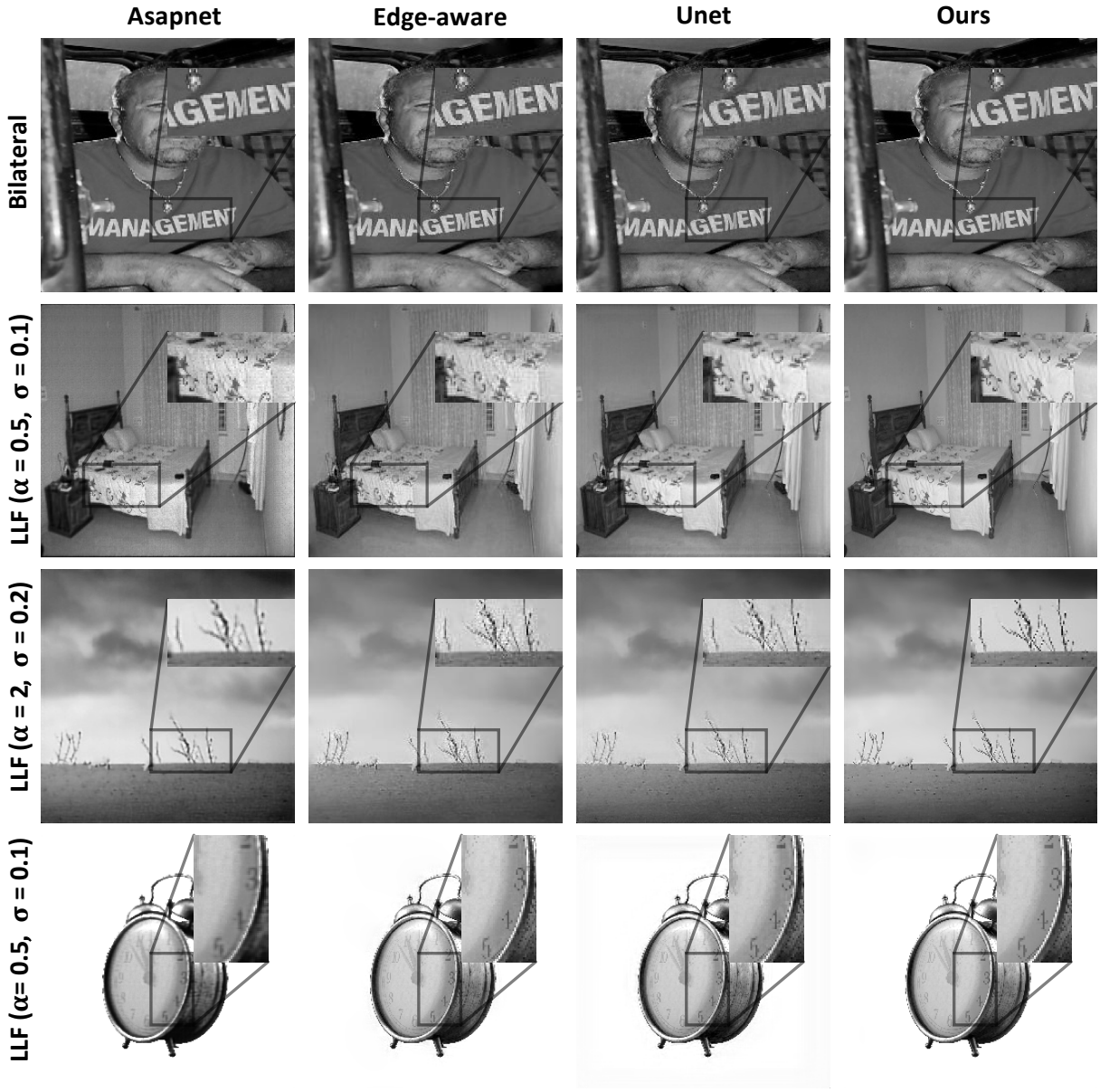

Figure 8: Qualitative comparisons with state-of-the-art methods on different types of images. These results are included in Table 1 for the filter types indicated in rows. The training strategy for the state-of-the-art methods is summarized in Section 5.3. Before-after pairs along with additional results can be found in the supplementary material. Image courtesy of Arnaud Rougetet (landscape) and virtualhorizonstudio (alarm clock). (CC-BY and PixelSquid).

International Conference on. 2311–2318. https://doi.org/10.1109/ICCV.2009.5459437

T. Karras, S. Laine, and T. Aila. 2021. A Style-Based Generator Architecture for Generative Adversarial Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 12 (dec 2021), 4217–4228. https://doi.org/10.1109/TPAMI.2020.2970919

Liad Kaufman, Dani Lischinski, and Michael Werman. 2012. Content-Aware Automatic Photo Enhancement. *Comput. Graph. Forum* 31, 8 (Dec. 2012), 2528–2540. https://doi.org/10.1111/j.1467-8659.2012.03225.x

Hanul Kim, Su-Min Choi, Chang-Su Kim, and Yeong Jun Koh. 2021. Representative color transform for image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 4459–4468.

Pierre-Yves Laffont, Zhile Ren, Xiaofeng Tao, Chao Qian, and James Hays. 2014. Transient Attributes for High-level Understanding and Editing of Outdoor Scenes. *ACM Trans. Graph.* 33, 4, Article 149 (July 2014), 11 pages. https://doi.org/10.1145/2601097.2601101

Wenbo Li, Kun Zhou, Lu Qi, Nianjuan Jiang, Jiangbo Lu, and Jiaya Jia. 2020. Lapar: Linearly-assembled pixel-adaptive regression network for single image super-resolution and beyond. *Advances in Neural Information Processing Systems* 33 (2020), 20343–20355.

Jing Liao, Yuan Yao, Lu Yuan, Gang Hua, and Sing Bing Kang. 2017. Visual Attribute Transfer through Deep Image Analogy. *ACM Trans. Graph.* 36, 4, Article 120 (jul 2017), 15 pages. https://doi.org/10.1145/3072959.3073683

Jianxin Lin, Yingxue Pang, Yingce Xia, Zhibo Chen, and Jiebo Luo. 2020. Tuigan: Learning versatile image-to-image translation with two unpaired images. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020,*

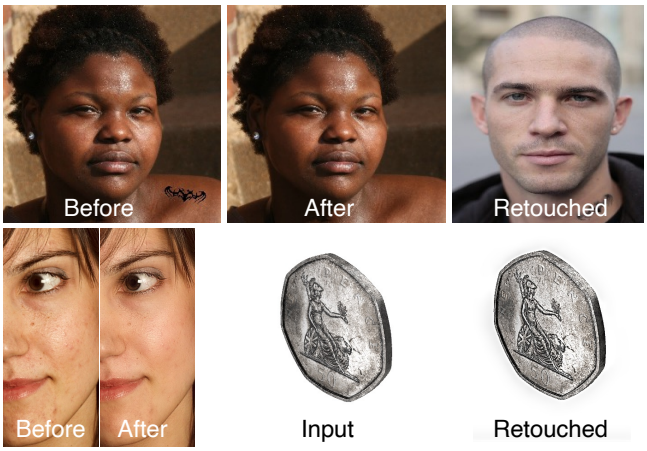

Figure 9: Our technique cannot accurately handle extreme non-repeating local effects such as tattoos (top), and when example and input images are of very different semantics (bottom). Image courtesy of bgaj23 (coin). (PixelSquid).

Proceedings, Part IV 16. Springer, 18–35.

Si Liu, Xinyu Ou, Ruihe Qian, Wei Wang, and Xiaochun Cao. 2016. Makeup like a Superstar: Deep Localized Makeup Transfer Network. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (New York, New York, USA) *(IJCAI'16)*. AAAI Press, 2568–2575.

Tian Ma, Ming Guo, Zhenhua Yu, Yanping Chen, Xincheng Ren, Runtao Xi, Yuancheng Li, and Xinlei Zhou. 2021. RetinexGAN: Unsupervised low-light enhancement with two-layer convolutional decomposition networks. *IEEE Access* 9 (2021), 56539–56550.

Sean Moran, Pierre Marza, Steven McDonagh, Sarah Parisot, and Gregory Slabaugh. 2020. Deeplpf: Deep local parametric filters for image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12826–12835.

Aamir Mustafa, Param Hanji, and Rafal Mantiuk. 2022. Distilling style from image pairs for global forward and inverse tone mapping. In *Proceedings of the 19th ACM SIGGRAPH European Conference on Visual Media Production*. 1–10.

Seonghyeon Nam and Seon Joo Kim. 2017. Deep Semantics-Aware Photo Adjustment. *CoRR* abs/1706.08260 (2017).

Mayu Omiya, Edgar Simo-Serra, Satoshi Iizuka, and Hiroshi Ishikawa. 2018. Learning Photo Enhancement by Black-Box Model Optimization Data Generation. In *SIGGRAPH Asia 2018 Technical Briefs*.

Jongchan Park, Joon-Young Lee, Donggeun Yoo, and In So Kweon. 2018. Distort-and-recover: Color enhancement using deep reinforcement learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5928–5936.

F. Pitie, A.C. Kokaram, and R. Dahyot. 2005. N-dimensional probability density function transfer and its application to color transfer. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, Vol. 2. 1434–1439 Vol. 2. https://doi.org/10.1109/ICCV.2005.166

François Pitié, Anil C. Kokaram, and Rozenn Dahyot. 2007. Automated Colour Grading Using Colour Distribution Transfer. *Comput. Vis. Image Underst.* 107, 1-2 (July 2007), 123–137. https://doi.org/10.1016/j.cviu.2006.11.011

Tania Pouli and Erik Reinhard. 2011. Progressive color transfer for images of arbitrary dynamic range. *Computers & Graphics* 35, 1 (2011), 67 – 80. https://doi.org/10.1016/j.cag.2010.11.003 Extended Papers from Non-Photorealistic Animation and Rendering (NPAR) 2010.

E. Reinhard, M. Ashikhmin, B. Gooch, and P. Shirley. 2001. Color transfer between images. *Computer Graphics and Applications, IEEE* 21, 5 (Sept. 2001), 34–41. https://doi.org/10.1109/38.946629

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 234–241.

Ardavan Saeedi, Matthew D. Hoffman, Stephen J. DiVerdi, Asma Ghandeharioun, Matthew J. Johnson, and Ryan P. Adams. 2018. Multimodal Prediction and Personalization of Photo Edits with Deep Generative Models. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS)* (2018-01-01). http://www.cs.princeton.edu/~rpa/pubs/saeedi2018multimodal.pdf arXiv:1704.04997 [stat.ML].

Tamar Rott Shaham, Michaël Gharbi, Richard Zhang, Eli Shechtman, and Tomer Michaeli. 2021. Spatially-adaptive pixelwise networks for fast image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14882–14891.

D. Shapira, S. Avidan, and Y. Hel-Or. 2013. Multiple histogram matching. In *Image Processing (ICIP), 2013 20th IEEE International Conference on*. 2269–2273. https://doi.org/10.1109/ICIP.2013.6738468

YiChang Shih, Sylvain Paris, Connelly Barnes, William T. Freeman, and Frédo Durand. 2014. Style Transfer for Headshot Portraits. *ACM Trans. Graph.* 33, 4, Article 148 (July 2014), 14 pages. https://doi.org/10.1145/2601097.2601137

Yichang Shih, Sylvain Paris, Frédo Durand, and William T. Freeman. 2013. Data-driven Hallucination of Different Times of Day from a Single Outdoor Photo. *ACM Trans. Graph.* 32, 6, Article 200 (Nov. 2013), 11 pages. https://doi.org/10.1145/2508363.2508419

Kalyan Sunkavalli, Micah K. Johnson, Wojciech Matusik, and Hanspeter Pfister. 2010. Multi-scale Image Harmonization. *ACM Trans. Graph.* 29, 4, Article 125 (July 2010), 10 pages. https://doi.org/10.1145/1778765.1778862

Yu-Wing Tai, Jiaya Jia, and Chi-Keung Tang. 2005. Local color transfer via probabilistic segmentation by expectation-maximization. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, Vol. 1. 747–754 vol. 1. https://doi.org/10.1109/CVPR.2005.215

Y. W. Tai, J. Jia, and C. K. Tang. 2007. Soft Color Segmentation and Its Applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29, 9 (Sept. 2007), 1520–1537. https://doi.org/10.1109/TPAMI.2007.1168

Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. 2021. Mlp-mixer: An all-mlp architecture for vision. *Advances in Neural Information Processing Systems* 34 (2021), 24261–24272.

Ethan Tseng, Felix Yu, Yuting Yang, Fahim Mannan, Karl ST Arnaud, Derek Nowrouzezahrai, Jean-François Lalonde, and Felix Heide. 2019. Hyperparameter optimization in black-box image processing using differentiable proxies. *ACM Trans. Graph.* 38, 4 (2019), 27–1.

Ethan Tseng, Yuxuan Zhang, Lars Jebe, Xuaner Zhang, Zhihao Xia, Yifei Fan, Felix Heide, and Jiawen Chen. 2022. Neural Photo-Finishing. *ACM Transactions on Graphics (TOG)* 41, 6 (2022), 1–15.

Ruixing Wang, Qing Zhang, Chi-Wing Fu, Xiaoyong Shen, Wei-Shi Zheng, and Jiaya Jia. 2019. Underexposed photo enhancement using deep illumination estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6849–6857.

Li Xu, Jimmy Ren, Qiong Yan, Renjie Liao, and Jiaya Jia. 2015. Deep edge-aware filters. In *International Conference on Machine Learning*. PMLR, 1669–1678.

Zhicheng Yan, Hao Zhang, Baoyuan Wang, Sylvain Paris, and Yizhou Yu. 2014. Automatic Photo Adjustment Using Deep Learning. *CoRR* abs/1412.7725 (2014).

Zhicheng Yan, Hao Zhang, Baoyuan Wang, Sylvain Paris, and Yizhou Yu. 2016a. Automatic photo adjustment using deep neural networks. *ACM Transactions on Graphics (TOG)* 35, 2 (2016), 1–15.

Zhicheng Yan, Hao Zhang, Baoyuan Wang, Sylvain Paris, and Yizhou Yu. 2016b. Automatic Photo Adjustment Using Deep Neural Networks. *ACM Trans. Graph.* 35, 2, Article 11 (Feb. 2016), 15 pages. https://doi.org/10.1145/2790296

Wenhan Yang, Shiqi Wang, Yuming Fang, Yue Wang, and Jiaying Liu. 2020. From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3063–3072.

Ke Yu, Zexian Li, Yue Peng, Chen Change Loy, and Jinwei Gu. 2021. ReconfigISP: Reconfigurable Camera Image Processing Pipeline. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4248–4257.

W. Zhang, C. Cao, S. Chen, J. Liu, and X. Tang. 2013. Style Transfer Via Image Component Analysis. *IEEE Transactions on Multimedia* 15, 7 (Nov. 2013), 1594–1601. https://doi.org/10.1109/TMM.2013.2265675

Feida Zhu and Yizhou Yu. 2018. Automatic Image Stylization Using Deep Fully Convolutional Networks. *CoRR* abs/1811.10872 (2018).