Pik-Fix: Restoring and Colorizing Old Photos

Runsheng Xu[1]*, Zhengzhong Tu[2]*, Yuanqi Du[3]*, Xiaoyu Dong[4], Jinlong Li[5], Zibo Meng[6]
Jiaqi Ma[1], Alan Bovik[2], Hongkai Yu[5]†
[1] University of California, Los Angeles, [2] University of Texas at Austin, [3] Cornell University
[4] Northwestern University, [5] Cleveland State University, [6] Innopeak Technology Inc.
rxx3386@ucla.edu, hongkaiyu2012@gmail.com

Abstract

Restoring and inpainting the visual memories that are present, but often impaired, in old photos remains an intriguing but unsolved research topic. Decades-old photos often suffer from severe and commingled degradation such as cracks, defocus, and color-fading, which are difficult to treat individually and harder to repair when they interact. Deep learning presents a plausible avenue, but the lack of large-scale datasets of old photos makes addressing this restoration task very challenging. Here we present a novel reference-based end-to-end learning framework that is able to both repair and colorize old, degraded pictures. Our proposed framework consists of three modules: a restoration sub-network that conducts restoration from degradations, a similarity network that performs color histogram matching and color transfer, and a colorization subnet that learns to predict the chroma elements of images conditioned on chromatic reference signals. The overall system makes uses of color histogram priors from reference images, which greatly reduces the need for large-scale training data. We have also created a first-of-a-kind public dataset of real old photos that are paired with ground truth "pristine" photos that have been manually restored by PhotoShop experts. We conducted extensive experiments on this dataset and synthetic datasets, and found that our method significantly outperforms previous state-of-the-art models using both qualitative comparisons and quantitative measurements. The code is available at `https://github.com/DerrickXuNu/Pik-Fix`.

1. Introduction

While our experience of the visual world are colorful, in earlier days of photography pictures were usually captured as "black and white", i.e. as gray-scale. As time elapses, they suffer other degradation as well. While consumer service are available for restoring and colorizing old photos, these require significant expertise in image manipulation, which is labour intensive, costly, and time-consuming. Thus, developing automated systems that can rapidly and accurately colorize and restore old photos is of interest.

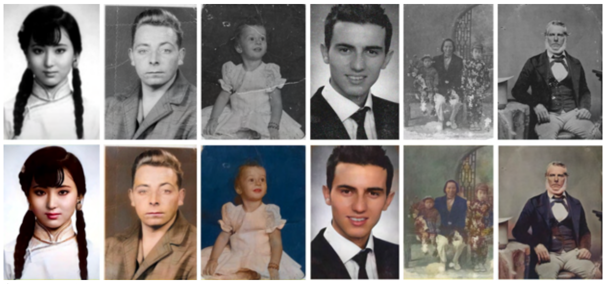
Figure 1: Examples of old photo repair (restoration and colorization) generated by Pik-Fix. Pik-Fix is able of simultaneously repair multiple image degradations of a photograph while also colorizing it.

Recently, deep learning based techniques have achieved high performance levels on a broad range of computer vision problems [9, 18, 21, 35, 36, 44, 52, 54–57, 64, 70]. They have also been successfully applied to image restoration such as image denoising [7, 33], super-resolution [30, 32], deblurring [34, 43], colorization [21, 64], and compression [3, 8]. However, learning-based colorization models generally require large-scale training datasets [64] to obtain favorable performance, which is energy-inefficient, labor-intensive, and time-consuming. Towards reducing large data requirements, the authors of [19, 21, 23, 60] proposed to employ reference/example images to assist colorization of gray-scale images. He *et al.* [21] uses separate similarity and colorization networks. However, since there are inherent ambiguities of the colors of natural objects because of the effects of ambient lighting. Better results than pixel-level color matching may be obtained by deriving features that describe the statistical color distributions of the reference pictures. In this direction, Yoo [59] deploy the means and variances of deep color features, but do not utilize second-order (spatial) distribution models, thereby discarding information descriptive of

*Equal contribution. † Corresponding author.

correlations that exist within image textures and their colors.

Since the spatial statistical color distribution is a very likely a useful source of colorization features, we have developed a reference-based, multi-scale spatial color histogram fusion method of image colorization. Using reference pictures to guide the colorization of gray-scale photographs relieves the need for large-scale training data. Precisely, we devised a novel end-to-end deep learning framework for old photo restoration which we dub Pik-Fix, which is composed of 1) a convolutional sub-network that is trained to conduct degradation restoration, 2) a similarity sub-network that performs reference color matching, and 3) a colorization sub-network that learns to render the final colorful image. As illustrated in Fig. 1, Pik-fix can restore and colorize the old degraded photos using only limited training data, making it attractive for data-efficient applications. Previous methods [46] mainly use the quantitative results on synthetic data with the restoration ground truth and qualitative results on collected real data without the restoration ground truth for experimental evaluations. To the best of our knowledge, there exists no similar public dataset of authentic, real-world degraded and gray-scale photos that are associated with pristine reference versions of the same photos. Towards advancing research in this direction, we designed and built *a first-of-a-kind real-world old photo dataset* consisting of 200 authentic old grayscale photos, where each old photo is paired with a 'pristine' version of it that was manually restored and colorized by Adobe Photoshop editors. Our experimental results show that Pik-Fix can outperform state-of-the-art methods on both existing public synthetic datasets and on our real-world old photo datasets, even though it requires much less training data. Our major contributions are summarized as follows:

- We propose *the first end-to-end deep learning framework* (Pik-Fix) that learns to simultaneously restore and colorize old photos, only requiring a small amount of training data.
- A *reference-based multi-scale color histogram fusion method* for image colorization that learns the content-aware transfer functions between the input and reference.
- *The first publicly available dataset of authentic, real-world degraded old photographs*. Each of these 200 authentic contents is paired with a 'pristine' version that were manually restored and colorized by Photoshop editors.
- Our experimental results show that the model, called Pik-Fix, achieves better visual and numerical performance than state-of-the-art methods on existing synthetic data and on our new real-world dataset.

2. Related Work

2.1. Image Colorization

Driven by deep neural networks, automatic image colorization have made great progress recently [10, 12, 67, 69]. Semantics analysis has been identified for successful colorization. For example, [22] and [69] design two-branch architectures that explicitly learn to fuse the local image features with global semantic predictions. The authors of [40] argue that pixel-level analysis is insufficient to learn subtle variations of object appearance and color, and shows that incorporating object-level analysis into that regression architecture yields better performance. Some works also try to employ reference images to help colorization and use a variety of ways to compute correspondence between the input pictures and the reference data, including pixel comparison [31, 51], semantic matching [5, 23], and super-pixel level [11, 19] similarities.

2.2. Image Restoration

There is a wide array of degradations that can affect older photographs, including some that occurred during capture, such as film grain and blur, and others that occur over time, like stains, fading color, and cracks. Traditional computational approaches to restore photos that have been digitized usually involve the application of prior constraints such as non-local self-similarity [4], sparsity [16], or local smoothness [50]. More recently, deep learning-based methods have proved efficacious on many picture restoration tasks, such as image denoising [61–63], super-resolution [15, 27, 30], and deblurring [34, 41, 53]. The success of these methods derives from the ability to simultaneously learn smooth semantics, and perceptual and local image representations.

2.3. Old Photo Restoration.

Old Photo Restoration aims at removing the degradations of old photos and colorizing them with natural colors. However, most of the existing models only address one particular aspect of old photo restoration, color restoration, or degradation restoration. The authors [47] designed an image-level pixel-to-pixel image translation framework using paired synthetic and real images. A model called Deoldify [2] also implements a pixel-to-pixel translation using a GAN. [45] learns to conduct single-degradation image restoration in an unsupervised manner. [46] first encodes image data into latent representations that separate old photos, ground truth, and synthetic images. It learns image restoration by producing the latent translation.

Although previous have been able to deliver perceptual equality by solely conducting colorization or restoration, in most instances old photo restoration requires both colorization and distortion restoration. Our work leverages both learning-based restoration and example-based color restoration methods to obtain old photo restoration that addresses both aspects. Importantly, our example-based colorization technique requires much less training data.

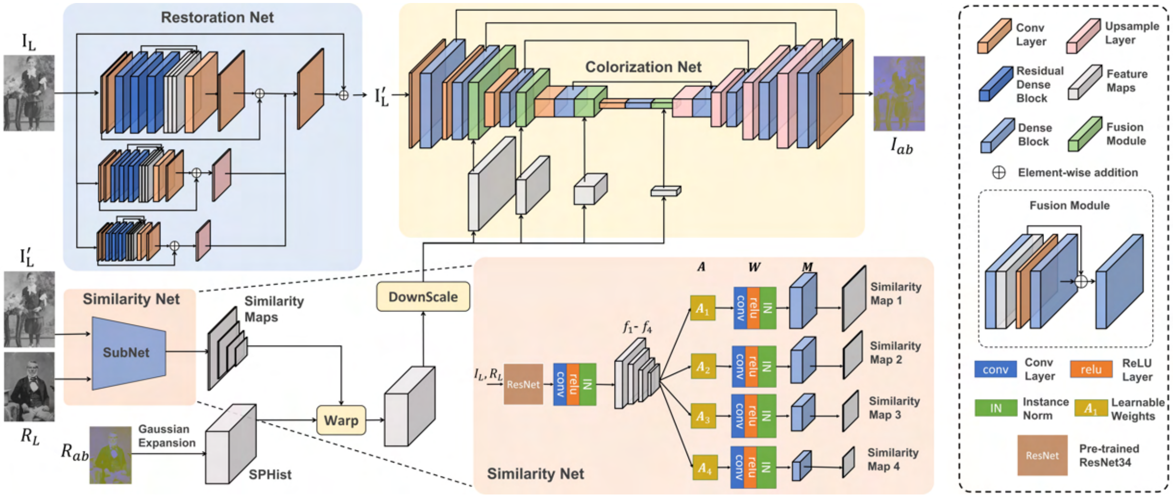

Figure 2: Flow diagram of the triplet networks (restoration, similarity and colorization) that define the flow of visual information processing in Pik-Fix.

3. Methodology

There are several major challenges that need to be addressed to further advance old photo restoration. Complex, commingled degradations are often observed in real-world old photos, which are impossible to model analytically, and difficult to gather into large amounts of representative training data. Further, colorization is an ill-posed, ambiguous problem [6], hence existing models require very large training datasets. The presence of complex distortions can make colorization harder. While this has not been deeply studied, the loss of real information likely impedes inferencing and regression. Conversely, restoring degraded gray-scale photos may be harder without clues supplied by color, which tends to be smooth and regional. Solving both problems together has the potential to improve the overall solution.

Towards overcoming these challenges, we propose an end-to-end framework, as depicted in Fig. 2. Denoting the input grayscale photo as $I_L \in \mathbb{R}^{H\times W\times 1}$, the restoration sub-net attempts to reverse any degradations to produce a restored gray-scale image $I'_L \in \mathbb{R}^{H\times W\times 1}$. Then, I'_L and the luminance channel of an associated reference picture $R_L \in \mathbb{R}^{H\times W\times 1}$ are both fed into the similarity sub-net, which produces a similarity map. Then, the chromatic features from the ab channels $R_{ab} \in \mathbb{R}^{H\times W\times 2}$ of the reference image are projected onto the input image space. The colorization sub-net accepts I'_L and the projected reference color features together as inputs, processing them to generate ab channels $I_{ab} \in \mathbb{R}^{H\times W\times 2}$, finally concatenating it with I'_L to obtain a restored and colorized result $I_{Lab} \in \mathbb{R}^{H\times W\times 3}$.

While previous methods operate by directly feeding the raw ab channels of the reference image into a colorization network [21, 60], or utilize low-order statistics (*e.g.*, mean and variance) of adaptive instance normalization of the reference image [58, 59], we instead employ a multi-scale fusion method that combines a spatial-preserving color histogram with deep features. The spatial-preserving color histogram contains useful prior information regarding the spatial relationships of color. The color features and deep features are aggregated over multiple scales, enabling the learning of the colorization process without a large number of training samples. In the following sections, we detail the restoration sub-net, similarity sub-net, colorization sub-net, and reference selection algorithms.

3.1. Restoration Sub-Net

Broadly, the types of degradations that affect old photos can be divided into two categories: physical defects (*e.g.*, cracks, tears, smudges) and capture defects (*e.g.*, blur, exposure) [46]. Correcting physical defects typically requires that the receptive fields of the analyzing neural network be large enough to capture impairments that span much of the photo dimensions. Yet it is also important that the network accesses local information since capture distortions usually manifect locally, even when globally present.

Here we address the bifurcated nature of old photo distortions by developing a multi-level Residual Dense Network (RDN [68]) that serves as the restortion sub-net. RDN models have previously demonstrated outstanding performance on common image restoration tasks like super-resolution, denoising, and deblurring, mainly facilitated by a core module called the residual dense block. The residual dense block is able to extract abundant information via the use of dense connections and contiguous memory mechanisms. While the RDN architecture has been shown to be suitable for handling capture defects, it processes images at a single resolution, restricting the sizes of the filter receptive fields and weakening its ability to correct physical flaws. To enable RDN to handle the broader range of distortions, we have formulated a multi-level RDN that is able to analyze distorted pictures over an enlarged span of receptive field sizes.

As shown in Fig. 2, an original picture, along with 4× and 8× downsampled versions of it are fed into the top, second, and third levels of the RDN, respectively. Each level consists of three residual dense blocks, each composed of 4 identical residual dense units. The outputs of the lower levels are upsampled via bilinear interpolation and fused via concatenation, then passed through another convolution layer to generate the restored luminance $I_L^{'}$.

3.2. Similarity Sub-Net

After the refined luminance map $I_L^{'}$ is obtained from the restoration sub-net, it is passed to the similarity sub-net along with the reference image's luminance channel R_L. The similarity sub-net is designed to project the reference image features onto the feature space of the input picture. As illustrated in Fig. 2, a pre-trained ResNet34 [20] is employed to retrieve layer1, layer2, layer3, layer4 feature maps from the input and reference pictures, respectively. Note that these feature maps have progressively smaller spatial resolutions and a larger number of feature channels with increased network depth. Then, four convolution layers are applied to these intermediate features, yielding feature maps having the same channel dimensions $f_i \in \mathbb{R}^{H_i \times W_i \times C}$ $(i = 1, 2, 3, 4)$. We utilize similarity maps at multiple scales to later allow for multi-level feature fusion in the colorization sub-net. Rather than simply resizing and concatenating the four feature maps, we propose to construct a learnable coefficient $A_i \in \mathbb{R}^{1\times 4}$, where $i =$1 to 4, that assigns different weights to the feature maps depending on the target similarity map size. These weighted feature maps are then concatenated together to obtain a feature tensor. For instance, the concatenated feature $M_i \in \mathbb{R}^{H_i \times W_i \times C}$ at scale i would be:

$$\begin{aligned} M_i = W \circledast [\mathrm{g}(A_{i1} * f_1) \oplus \mathrm{g}(A_{i2} * f_2) \\ \oplus \mathrm{g}(A_{i3} * f_3) \oplus \mathrm{g}(A_{i4} * f_4)], \end{aligned} \tag{1}$$

where W is the shared convolution filter for the convolution operator $\circledast$, g is an up-sampling or down-sampling function that aligns the feature size to the target similarity map size, $*$ indicates element-wise multiplication, and $\oplus$ denotes the concatenation operation. Subsequently, the three-dimensional feature vector M_i will be reshaped to a two-dimension matrix $\overline{M_i} \in \mathbb{R}^{H_i W_i \times C}$. Then, the similarity map $\Phi^i_{R\leftrightarrow I} \in \mathbb{R}^{H_i W_i \times H_i W_i}$ characterizing the correlation structure between the reference picture R and the input picture I at scale level i is computed at each spatial location (u, v) as follows:

$$\Phi^i_{R\leftrightarrow I}(u,v) = \frac{(\bar{M_i}^I(u) - \mu_{\bar{M_i}^I}) \cdot (\bar{M_i}^R(v) - \mu_{\bar{M_i}^R})}{||\bar{M_i}^I(u) - \mu_{\bar{M_i}}||_2 ||\bar{M_i}^R(u) - \mu_{\bar{M_i}^R}||_2}, \tag{2}$$

where $\mu_{\bar{M_i}^I}$ and $\mu_{\bar{M_i}^R}$ are mean feature vectors. The softmax function is then applied to the elements of the similarity map along the x-axis so each mapped element lies within [0,1]. This similarity map is then passed to the colorization sub-net, whose task is simplified since the reference picture's information is aligned with that of the input image.

3.3. Colorization Sub-Net

To tackle the aforementioned colorization problem, we develop a method of guiding the process using the color prior in the reference picture. Specifically, we utilize a space-preserving color histogram (SPHist) computed on the reference pictures. Unlike the traditional color histogram, the SPHist can retain spatial picture information while modeling the probability that each pixel color falls within each bin. Importantly, *SPHist is differentiable*, and thus, it can be used in an end-to-end neural network trained using gradient back-propagation. We accomplish this by using Gaussian expansion [38] to separately approximate the SPHist $h \in \mathbb{R}^{H \times W \times K}$ of each channel, where K is the number of histogram bins. Then, the probability of a pixel at location (i, j) falling into the k-th bin is expressed as follows:

$$h(i,j,k) = \frac{\exp(-(D_{ij} - u_k)^2/2\sigma^2)}{\sum_{k=1}^{K} \exp(-(D_{ij} - u_k)^2/2\sigma^2)}, \tag{3}$$

where D_{ij} is the value of a (or b) channel of the reference picture at spatial coordinate (i, j); the spread of the Gaussian distribution is fixed at $\sigma = 0.1$; u_k is a learnable parameter representing the center of bin k, which is initialized as:

$$u_k^0 = v_{min} + (v_{max} - v_{min})/K * k, \tag{4}$$

where v_{min} and v_{max} are the minimum and maximum possible values of the ab channels (-1 and 1, respectively in our experiments). Although the bins are equally distributed at the start, after training over several iterations, their distributions become unequal, since some colors are rarer than others 'in the wild'. The extracted color histogram is reshaped to $\overline{h} \in \mathbb{R}^{HW \times K}$ and down-sampled to the available four scales to enable matrix multiplication with the corresponding scales of similarity maps, leading to a warped SPHist that contains similarity-guided space-preserved color histogram from the reference picture. The warped SPHist is then fed into different levels of the encoder in the colorization sub-network to conduct color prediction.

The backbone of our colorization network employs a global U-Net shape [37] with densenet blocks [42]. There are four dense blocks in the encoder containing 6, 12, 24, and 16 dense units. The decoder shares a similar structure as the encoder, and bi-linear interpolation is employed to upscale the forwarding features between the dense blocks. The warped SPHist extracted from the reference picture is concatenated with the intermediate features after each dense block in the encoder, yielding inputs to the fusion module. The fusion module contains a dense block with six dense units and a $3{*}3$ convolution layer, which is responsible for efficiently combining the traditional color heuristics and deep features to enable accurate colorization. Since the reference information of reference is fused during the intermediate stages instead of at the start, the model learns to deal with dissimilarities between the input and reference pictures in a multi-scale manner.

3.4. Training Objective

In order to 1) simultaneously train the restoration and colorization nets, 2) exploit the rich color information available in the reference pictures, and 3) improve the visual quality of the overall restored output, we employed a weighted sum of diverse objectives functions against which the entire Pik-Fix system can be trained end-to-end. Among these, the *luminance reconstruction loss* between the restored luminances $I_L^{'}$ and the ground truth luminances G_L are used to supervise the training of restoration subnet: $\mathcal{L}_{rec,L} = ||I_L^{'} - G_L||_1$.

However, it is well-known that relying on ℓ_p norms as loss function tends to generate blurred estimates of picture restoration [30]. Hence, we also used a measure of *perceptual loss* that has been shown to deliver better quality visual results on a variety of restoration tasks [14, 30, 48]: $\mathcal{L}_{perc,L} = \sum_j \frac{1}{C_j H_j W_j} ||\phi_j(I_L^{'}) - \phi_j(G_L)||_2^2$, where ϕ_j is a feature map of shape $C_j \times H_j \times W_j$.

The colorization subnet is intended to transfer color distributions from the reference picture to the predicted output pictures. Thus, we also use the *histogram loss* to measure the distribution distance between the color histograms of the output and reference pictures as expressed by the Earth Mover's Distance (EMD): $\mathcal{L}_{\mathrm{EMD},\hbar} = \sum_{k=1}^{K} (\mathrm{CDF}_{\hbar_{I'}}(k) - \mathrm{CDF}_{\hbar_R}(k))^2$, where $\mathrm{CDF}_p(k)$ is the k-th element of the cumulative density function of the probability mass function p. $\hbar_{I'}$ and $\hbar_R$ are one-dimensional differentiable histograms formed by globally pooling over the SPHist features $h_{I'}$ and h_R in Eq. (3), respectively.

We also use the *chroma reconstruction loss* to impose the spatial consistency between the predicted chromatic channels ab and the ground truth ab channels, supplementing the histogram loss by directly controlling the pixel-wise chromatic loss: $\mathcal{L}_{rec,ab} = ||I_{ab}^{'} - G_{ab}||_1$.

The *adversarial loss* is a recipe that is often used to enhance the visual quality of images synthesized using GANs [26, 29, 30]. We utilize a PatchGAN [24] structure to ensure that all of the local patches of the enhanced output channels are visually similar to realistic chroma maps. The adversarial loss is expressed as: $mathcalL_{adv,ab} = \mathbb{E}_{G_{ab}}[\log D(G)] + \mathbb{E}_{I'_{ab}}[\log(1 - D(I^{'}, R))]$.

Finally, we combine all of the above directed loss functions into an overall loss under which Pik-Fix is trained: $\mathcal{L} = \alpha\mathcal{L}_{rec,L} + \beta\mathcal{L}_{perc,L} + \lambda\mathcal{L}_{\mathrm{EMD},\hbar} + \gamma\mathcal{L}_{rec,ab} + \eta\mathcal{L}_{adv,ab}$.

3.5. Reference Picture Selection

As discussed earlier, Pik-Fix requires color reference pictures as additional inputs to guide the colorization process. Thus, we developed an automatic reference curation model that generates good reference pictures from a given database, given an input grayscale picture during either the training and or inferencing phases.

An ideal reference should be both visually and semantically similar to the target image to be colorized, while providing rich and appropriate color information for the colorizing process. Inspired by the deep perceptual similarity models [13, 65], we leveraged a pre-trained VGG19 net [39] as backbone to extract intermediate deep feature maps. Then, we measured the degrees of textural and the structural similarity between each given grayscale input image and each of the available color images using the global means and variance/covariances of their feature maps, respectively [13]. Finally, a weighted summation of the texture and structure similarities is used to determine which of the color reference pictures in the training set has the greatest similarity to the grayscale input picture to be repaired and colorized. When deploying the trained Pik-Fik system, user may either automatically select a recommended reference picture retrieved from an available corpus, or they may choose to manually select a reference picture, according to their preference.

4. Experiments

4.1. Experimental Setting

Dataset We trained and evaluated our method on three datasets: Div2K [1], Pascal [17], and RealOld. In our experiments, we used the Div2K training and validation sets (800/100) for model training and testing, respectively. For Pascal, we randomly selected 10,000/1000 images to serve as training data and testing data. Two different experiments were conducted: simultaneous image restoration and colorization, and only image colorization. In order to produce realistic defect pictures, similar to those used in [46], we hired Photoshop experts to mimic the degradation patterns in real old photos (but not from our newly created RealOld dataset) on images from the Pascal dataset, using alpha compositing with randomized transparency levels, thereby generating synthetic old photos. We also added Gaussian blur, and simulated severe photo damage by randomly setting polygonal picture regions to pure white. We restrict that reference images can only be retrieved from the training set.

Real-World Old Photos (RealOld): To validate the efficacy and generalizability of our model under realistic conditions, we collected digitized copies of 200 real old black & white photographs. Each of these photos were digitally manually restored and colorized by Photoshop experts. To the best of our knowledge, this is the first real-world old photo dataset that has aligned "ground truth" 'pristine' pho-

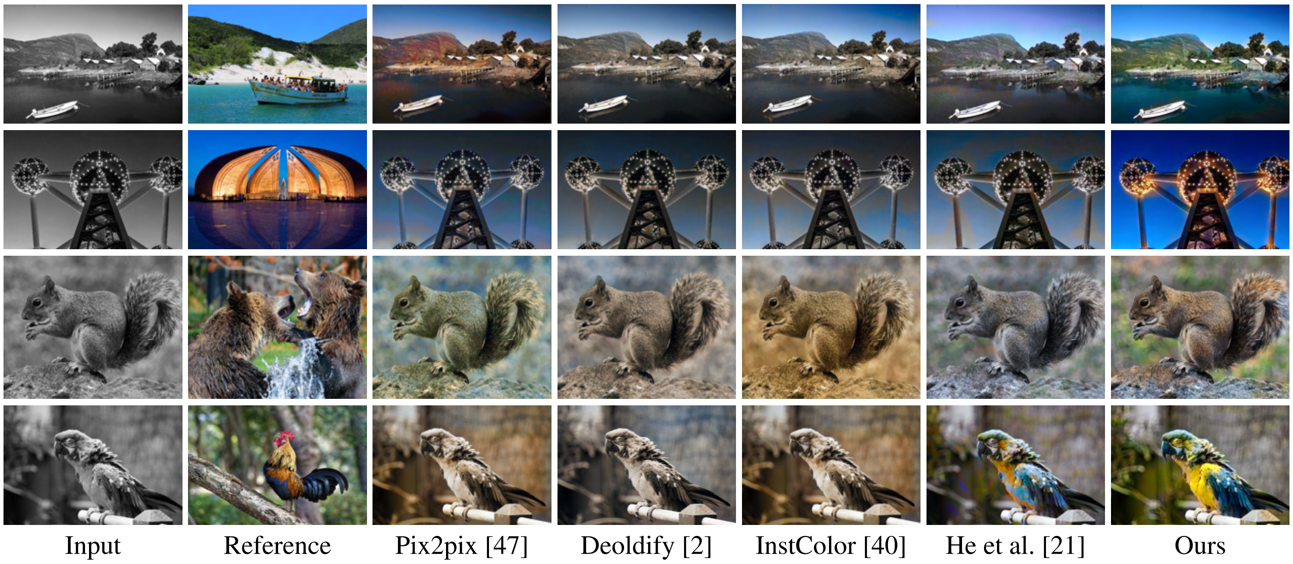

Figure 3: Visual comparisons against state-of-the-art colorization methods on DIV2K. It shows that with only 800 training images, our method is able to accomplish visually pleasant colorization and our result is significantly better than others.

Table 1: Quantitative comparison on the DIV2K and Pascal VOC validation datasets. Up-ward arrows indicate that a higher score denotes a good image quality. We highlight the best score for each measure.

Dataset	DIV2K (w/o degradation)			Pascal VOC (w/o degradation)			Pascal VOC (w/ degradation)		
Metric	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Pix2pix	21.12	0.872	0.138	20.89	0.782	0.200	20.37	0.732	0.231
DeOldify	23.65	0.913	0.128	23.96	0.873	0.117	21.45	0.789	0.192
He *et al.*	23.53	0.918	0.125	23.85	0.925	0.114	-	-	-
InstColorization	22.45	0.914	0.131	23.95	0.932	0.111	-	-	-
Wan *et al.* -	-	-	-	-	-	-	18.01	0.598	0.421
Ours	**23.95**	**0.925**	**0.120**	**24.01**	**0.940**	**0.100**	**22.22**	**0.828**	**0.186**

tos to enable pixel-to-pixel processing and comparison. We are making this dataset publicly available to allow other researchers to develop advanced algorithms that can both colorize and repair old photos impaired by scratches, blur, cracks, wear, film grain noise, and physical and capture distortions. In our experiments, RealOld is used for testing models that have been trained on Pascal. Furthermore, we randomly downloaded 2,000 RGB portraits from Google Images, and utilize our picture selection algorithm to pick the best references among them, when testing on ReadOld.

Evaluation Metrics. We report PSNR and SSIM [49] scores between the reference and restored/colorized pictures. As an alternative, we also use the learned perceptual image patch similarity (LPIPS) metric [66].

Training Details. We trained Pik-Fix in an end-to-end manner using the Adam solver [28], with $\beta_1 = 0.99$ and $\beta_2 = 0.999$. The initial learning rate was set to 0.0001 and exponentially decreased at the end of each epoch using a decay rate of 0.99. The loss balance weights were fixed as follows: $\alpha = 1.0, \beta = 0.2, \lambda = 0.5, \gamma = 1.0, \eta = 0.2$. For data augmentation, randomly cropped 256×256 patches were included in the training data. At each epoch, we selected one of the following two methods of reference patch/picture generation: 1) a patch was cropped from an RGB image at a location different from that of the patch used as input, and processed with color jittering and a small affine transformation to create a reference picture; 2) one picture was randomly selected from the training set (excluding the ground-truth) and treated as the reference. All of the compared models were trained on DIV2K and Pascal for 20 epochs, respectively, on a single GTX 3090Ti GPU.

4.2. Experimental Results

Since no existing work has explicitly considered the simultaneous correction of picture degradation and colorization, we compared Pik-Fix with models developed for image-to-image translation (denoted as Pix2pix [25]), image restoration (denoted as Wan *et al.* [46]), and colorization (denoted as Deoldify [2], He *et al.* [21] and InstColorization [40]). For fair comparisons, we do not evaluate InstColorization and He *et al.* (which do not restore degradation) on Pascal VOC with degradation. Likewise, we did not compare against Wan *et al.* (which does not colorize), on DIV2K without degradation or on Pascal VOC without degradation.

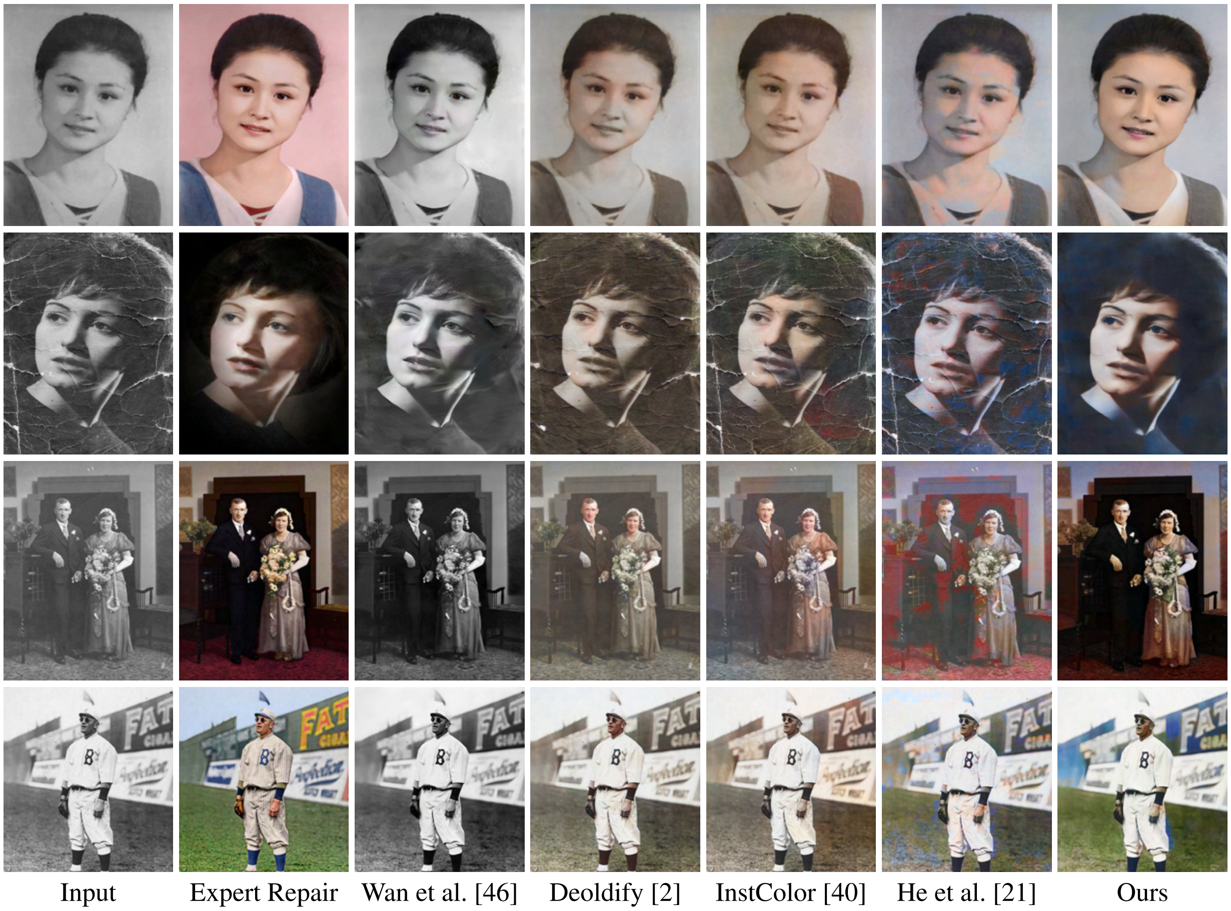

Figure 4: Visual comparisons against state-of-the-art colorization and restoration methods on RealOld. It shows that with limited synthetic training data from Pascal, our model is able to fix most of the degradation and deliver plausible colorization.

All the compared methods were trained from scratch using the training strategies and code provided by those authors.

4.2.1 Quantitative Comparison

Table 1 compares Pik-Fix and the other models' performances on two public datasets under two scenarios: DIV2K without degradation, Pascal VOC without degradation, and Pascal VOC with degradation. Pik-Fix delivered the best results against all three evaluation metrics as compared with these state-of-the-art models. For example, on the DIV2K dataset without degradation, Pik-Fix achieved much better scores of 23.95 PSNR, 0.925 SSIM, and 0.120 LPIPS, than any of the compared models.

Table 2 shows the results obtained on the RealOld dataset, where again, Pik-Fix generated the highest performance scores among all compared models. These results strongly highlight the performance resilience by Pik-Fix when transferring from synthetic dataset to real-world old photo dataset.

4.2.2 Qualitative Comparison

Figure 3 provides qualitative comparison on Divk2K. Pik-Fix produced pictures having vivid, realistic colors, while compared models delivered incomplete colorization. Fig. 4 shows results on the RealOld dataset, showing that Pik-Fix can simultaneously perform picture restoration and colorization, producing perceptually satisfying results.

4.2.3 User Study

We conducted a user study to compare the visual results of all the methods. We randomly selected 100 old photos from the RealOld dataset, and asked 15 users to rank the results based on their subjective visual impressions. We gathered reports from these 15 people with the results presented in Table 3. Pik-Fix attained a high probability of 50.6% of being selected as the single top performer, outperforming all of the other methods over all rankings, further illustrating the strong performance of the Pik-Fix picture restoration and colorization engine.

Table 2: Quantitative comparisons of restoration/colorization performance on the RealOld dataset.

Dataset	Real old photo		
Metric	PSNR↑	SSIM↑	LPIPS↓
Pix2pix	16.80	0.684	0.320
DeOldify	17.14	0.723	0.287
He *et al.*	16.72	0.707	0.314
InstColorization	16.86	0.715	0.312
Wan *et al.*	16.99	0.709	0.303
Ours	**17.20**	**0.758**	**0.258**

Table 3: User rankings of algorithm performance on the RealOld dataset. The percentage (%) of users choosing each model ranking is shown.

Method	Top 1	Top 2	Top 3	Top 4	Top 5
Pix2Pix [25]	3.1	11.8	24.3	41.6	67.7
Deoldify [2]	10.5	33.9	56.5	79.3	92.8
He *et al.* [21]	4.6	26.3	44.1	61.9	83.2
InstColorization [40]	8.0	26.3	53.6	75.5	92.2
Wan *et al.* [46]	23.1	38.6	48.9	59.1	72.0
Ours	**50.6**	**65.9**	**75.3**	**85.1**	**94.4**

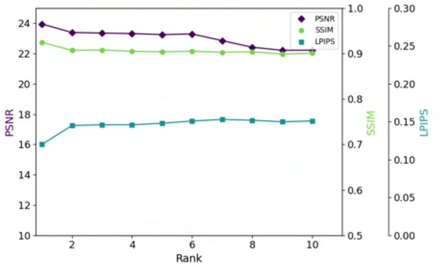

Figure 5: Sensitivity analysis of reference image selection.

4.3. Ablation Studies

Multi-scale SPHist. We conducted three experiments on the Div2k dataset to evaluate the effectiveness of multi-scale SPHist: 1) following [21], the transferred ab channels of the reference picture and the L channel of the input are concatenated and fed into the backbone of the colorization sub-net; 2) instead of computing the multi-scale SPHist of the reference image, the multi-scale raw ab channels of the reference image are used as input; 3) only a single-scale color histogram is fused with the shallower layers of the encoder. The results reported in Table 4 has validated the importance and usefulness of the proposed multi-scale SPHist.

Multi-scale Similarity Maps. To validate the efficacy of multi-scale similarity maps relative to using a single similarity map, we conducted two experiments: 1) we use the single-scale similarity map proposed in [60]; 2) no similarity map is applied to the reference image. Table 5 reflects the benefits brought by the use of multi-scale similarity maps.

Table 4: Ablation study of multi-scale SPHist on Div2k.

Method	PSNR↑	SSIM↑	LPIPS↓
Input *ab* fusion	22.978	0.902	0.130
Multi-scale *ab* fusion	23.233	0.910	0.127
Single-scale histogram fusion	23.631	0.906	0.125
Multi-scale histogram fusion	**23.952**	**0.925**	**0.120**

Table 5: Ablation study of multi-scale similarity maps.

Method	PSNR↑	SSIM↑	LPIPS↓
No similarity map	22.817	0.910	0.131
Single-scale similarity map	23.803	0.922	0.126
Multi-scale similarity map	**23.952**	**0.925**	**0.120**

Table 6: Ablation study of multi-level RDN on Pascal.

Method	PSNR↑	SSIM↑	LPIPS↓
Single-level RDN	21.89	0.818	0.190
Multi-level RDN	**22.22**	**0.828**	**0.186**

Multi-scale RDN. To study the possible performance gains brought by building a multi-level Residual Dense Network, we also tried the origin RDN [68] as the backbone for old photo restoration and tested the modified system on the Pascal dataset [17] with degradation. The results in Table 6 show that the multi-level design significantly improve the quality of the restored outputs.

Sensitivity to Reference Image Selection. To examine the robustness of our model relative to the selection of reference images, we compared the results obtained on the DIV2K dataset using different reference pictures than the computed "most similar" one (rank 1) to the least similar one (rank 10) among the ten selected reference pictures. As shown in Fig. 5, the performance of Pik-Fix is robust against reference picture selection, with only slight performance drops, *viz.*, from 23.95 (rank 1) to 22.25 (rank 10) of PSNR, from 0.925 (rank 1) to 0.899 (rank 10) of SSIM, and from 0.12 (rank 1) to 0.15 (rank 10) of LPIPS.

5. Concluding Remarks

We propose a first end-to-end trainable system called Pik-Fix that is able to simultaneously restore and colorize old photos. The overall system contains several subnetworks, each designed to handle a single defect, but trained holistically. A hierarchical restoration subnet recovers the luminance channel from physical and capture distortions, followed by a colorization subnet that uses space-preserving color histograms to estimate the chroma components. Extensive experimental results show that Pik-Fix attains excellent performance both visually and numerically on synthetic and real old photo datasets, as compared with state-of-the-art models. Moreover, we created the first publicly available real-world old photo dataset repaired by Photoshop experts, which we hope will facilitate further research on deep learning-based old photo restoration problems.

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, July 2017.

[2] Jason Anctic. jantic/deoldify: A deep learning based project for colorizing and restoring old images (and video!).

[3] Johannes Ballé, Valero Laparra, and Eero P Simoncelli. End-to-end optimized image compression. *arXiv preprint arXiv:1611.01704*, 2016.

[4] Antoni Buades, Bartomeu Coll, and J-M Morel. A non-local algorithm for image denoising. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, volume 2, pages 60–65. IEEE, 2005.

[5] Guillaume Charpiat, Matthias Hofmann, and Bernhard Schölkopf. Automatic image colorization via multimodal predictions. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pages 126–139. Springer, 2008.

[6] Guillaume Charpiat, Matthias Hofmann, and Bernhard Schölkopf. Automatic image colorization via multimodal predictions. volume 2008, 09 2008.

[7] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3291–3300, 2018.

[8] Li-Heng Chen, Christos G Bampis, Zhi Li, Andrey Norkin, and Alan C Bovik. Proxiqa: A proxy approach to perceptual optimization of learned image compression. *IEEE Transactions on Image Processing*, 30:360–373, 2020.

[9] Weizhe Chen, Runsheng Xu, Hao Xiang, Lantao Liu, and Jiaqi Ma. Model-agnostic multi-agent perception framework. *arXiv preprint arXiv:2203.13168*, 2022.

[10] Zezhou Cheng, Qingxiong Yang, and Bin Sheng. Deep colorization. In *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pages 415–423, 2015.

[11] Alex Yong-Sang Chia, Shaojie Zhuo, Raj Kumar Gupta, Yu-Wing Tai, Siu-Yeung Cho, Ping Tan, and Stephen Lin. Semantic colorization with internet images. *ACM Trans. Graphics*, 30(6):1–8, 2011.

[12] Aditya Deshpande, Jason Rock, and David Forsyth. Learning large-scale automatic image colorization. In *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pages 567–575, 2015.

[13] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *arXiv preprint arXiv:2004.07728*, 2020.

[14] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Comparison of full-reference image quality models for optimization of image processing systems. *Int. J. Comput. Vis.*, 129(4):1258–1281, 2021.

[15] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pages 184–199. Springer, 2014.

[16] Michael Elad and Michal Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Trans. Image process.*, 15(12):3736–3745, 2006.

[17] Mark Everingham, S. Eslami, Luc Van Gool, Christopher Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *Int. J. Comput. Vis.*, 111, 01 2014.

[18] Zhaoxin Fan, Zhenbo Song, Jian Xu, Zhicheng Wang, Kejian Wu, Hongyan Liu, and Jun He. Object level depth reconstruction for category level 6d object pose estimation from monocular rgb image. *arXiv preprint arXiv:2204.01586*, 2022.

[19] Raj Kumar Gupta, Alex Yong-Sang Chia, Deepu Rajan, Ee Sin Ng, and Huang Zhiyong. Image colorization using similar images. In *Proc. ACM Multimedia Conf. (MM)*, pages 369–378, 2012.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 770–778, 2016.

[21] Mingming He, Dongdong Chen, Jing Liao, Pedro V Sander, and Lu Yuan. Deep exemplar-based colorization. *ACM Trans. Graphics*, 37(4):1–16, 2018.

[22] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Let there be color! joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. *ACM Trans. Graphics*, 35(4):1–11, 2016.

[23] Revital Ironi, Daniel Cohen-Or, and Dani Lischinski. Colorization by example. In *Rendering Techniques*, pages 201–210. Citeseer, 2005.

[24] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 1125–1134, 2017.

[25] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017.

[26] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. Enlightengan: Deep light enhancement without paired supervision. *IEEE Trans. Image Process.*, 30:2340–2349, 2021.

[27] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 1646–1654, 2016.

[28] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Int. Conf. Learn. Represent. (ICLR)*, 12 2014.

[29] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 8183–8192, 2018.

[30] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 4681–4690, 2017.

[31] Xiaopei Liu, Liang Wan, Yingge Qu, Tien-Tsin Wong, Stephen Lin, Chi-Sing Leung, and Pheng-Ann Heng. Intrinsic colorization. In *ACM SIGGRAPH Asia*, pages 1–9. 2008.

[32] Yiqun Mei, Yue Zhao, and Wei Liang. Dsr: An accurate single image super resolution approach for various degradations. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2020.

[33] Zibo Meng, Runsheng Xu, and Chiu Man Ho. Gia-net: Global information aware network for low-light imaging. In *European Conference on Computer Vision*, pages 327–342. Springer, 2020.

[34] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 3883–3891, 2017.

[35] David Paz, Hao Xiang, Andrew Liang, and Henrik I Christensen. Tridentnetv2: Lightweight graphical global plan representations for dynamic trajectory generation. *arXiv preprint arXiv:2203.14019*, 2022.

[36] David Paz, Hengyuan Zhang, Qinru Li, Hao Xiang, and Henrik I Christensen. Probabilistic semantic mapping for urban autonomous driving applications. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2059–2064. IEEE, 2020.

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Int. Conf. Medical Image Comput. Computer-assisted Intervention*, pages 234–241. Springer, 2015.

[38] Kristof Schütt, Farhad Arbabzadah, Stefan Chmiela, Klaus-Robert Müller, and Alexandre Tkatchenko. Quantum-chemical insights from deep tensor neural networks. *Nature Communications*, 8, 01 2017.

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

[40] Jheng-Wei Su, Hung-Kuo Chu, and Jia-Bin Huang. Instance-aware image colorization. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 7968–7977, 2020.

[41] Jian Sun, Wenfei Cao, Zongben Xu, and Jean Ponce. Learning a convolutional neural network for non-uniform motion blur removal. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 769–777, 2015.

[42] Xiejie Liu Tong Tong, Gen Li and Qinquan Gao. Image super-resolution using dense skip connections. In *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017.

[43] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxim: Multi-axis mlp for image processing. *arXiv preprint arXiv:2201.02973*, 2022.

[44] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. *arXiv preprint arXiv:2204.01697*, 2022.

[45] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 9446–9454, 2018.

[46] Ziyu Wan, Bo Zhang, Dongdong Chen, Pan Zhang, Dong Chen, Jing Liao, and Fang Wen. Bringing old photos back to life. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 2747–2757, 2020.

[47] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 8798–8807, 2018.

[48] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pages 0–0, 2018.

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

[50] Yair Weiss and William T Freeman. What makes a good model of natural images? In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 1–8. IEEE, 2007.

[51] Tomihisa Welsh, Michael Ashikhmin, and Klaus Mueller. Transferring color to greyscale images. In *Annual Conf. Comput. Graphics Interactive Techniques*, pages 277–280, 2002.

[52] Hao Xiang, Runsheng Xu, Xin Xia, Zhaoliang Zheng, Bolei Zhou, and Jiaqi Ma. V2XP-ASG: Generating Adversarial Scenes for Vehicle-to-Everything Perception. *arXiv e-prints*, page arXiv:2209.13679, Sept. 2022.

[53] Li Xu, Jimmy S Ren, Ce Liu, and Jiaya Jia. Deep convolutional neural network for image deconvolution. *Adv. Neural Information Process. Syst.*, 27:1790–1798, 2014.

[54] Runsheng Xu, Yi Guo, Xu Han, Xin Xia, Hao Xiang, and Jiaqi Ma. Opencda: An open cooperative driving automation framework integrated with co-simulation. In *2021 IEEE Intelligent Transportation Systems Conference (ITSC)*, 2021.

[55] Runsheng Xu, Zhengzhong Tu, Hao Xiang, Wei Shao, Bolei Zhou, and Jiaqi Ma. Cobevt: Cooperative bird's eye view semantic segmentation with sparse transformers. *arXiv preprint arXiv:2207.02202*, 2022.

[56] Runsheng Xu, Hao Xiang, Zhengzhong Tu, Xin Xia, Ming-Hsuan Yang, and Jiaqi Ma. V2x-vit: Vehicle-to-everything cooperative perception with vision transformer. *arXiv preprint arXiv:2203.10638*, 2022.

[57] Runsheng Xu, Hao Xiang, Xin Xia, Xu Han, Jinlong Li, and Jiaqi Ma. Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2583–2589. IEEE, 2022.

[58] Zhongyou Xu, Tingting Wang, Faming Fang, Yun Sheng, and Guixu Zhang. Stylization-based architecture for fast deep exemplar colorization. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2020.

[59] Seungjoo Yoo, Hyojin Bahng, Sunghyo Chung, Junsoo Lee, Jaehyuk Chang, and Jaegul Choo. Coloring with limited data: Few-shot colorization via memory-augmented networks. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2019.

[60] Bo Zhang, Mingming He, Jing Liao, Pedro V Sander, Lu Yuan, Amine Bermak, and Dong Chen. Deep exemplar-based

video colorization. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 8052–8061, 2019.

[61] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Trans. Image Process.*, 26(7):3142–3155, 2017.

[62] Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep cnn denoiser prior for image restoration. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 3929–3938, 2017.

[63] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Ffdnet: Toward a fast and flexible solution for cnn-based image denoising. *IEEE Trans. Image Process.*, 27(9):4608–4622, 2018.

[64] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pages 649–666. Springer, 2016.

[65] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 586–595, 2018.

[66] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018.

[67] Richard Zhang, Jun-Yan Zhu, Phillip Isola, Xinyang Geng, Angela S Lin, Tianhe Yu, and Alexei A Efros. Real-time user-guided image colorization with learned deep priors. *arXiv preprint arXiv:1705.02999*, 2017.

[68] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image restoration. *IEEE Trans. Pattern Anal. Mach. Intell.*, PP:1–1, 01 2020.

[69] Jiaojiao Zhao, Li Liu, Cees GM Snoek, Jungong Han, and Ling Shao. Pixel-level semantics guided image colorization. *arXiv preprint arXiv:1808.01597*, 2018.

[70] Zelin Zhao, Ze Wu, Yueqing Zhuang, Boxun Li, and Jiaya Jia. Tracking objects as pixel-wise distributions. *arXiv preprint arXiv:2207.05518*, 2022.