Old Negative Films Restoration by Cross Channel Prediction

Qi LI
1155092087@link.cuhk.edu.hk
The Chinese University of Hong Kong

ABSTRACT

Old photo restoration has been an ill-posed problem for a long time. In this project, we proposed a new model based on the convolutional neural network(CNN) to remove the contamination on old photos and restore them by using the cross channel information. The network consists of three parts, which are the mask prediction sub-network to find the contaminated area, the cross channel prediction sub-network to predict an intermediate result, and the final prediction sub-network to produce the final result based on the information provided by previous sub-networks. The proposed model performs very well on the synthetic datasets and can also remove most of noticeable contamination on real old negative films.

1 INTRODUCTION

Nowadays, most people can use the digital camera and their cellphones to take high-quality photos very conveniently. With the development of technology, they can capture, save and share the pictures very easily. However, the old photos that shot by film camera decades ago were not that fortunate.At the age of film camera, each shoot cost one film, which was quite expensive at that time. And the films are not reusable. Not only were the films expensive, they were also very hard to preserve. After decades' degradation, there are more or less some contamination on these old films. It is a meaning task to restore those precious films, which contains valuable memories of that time.

Image restoration(**IR**) is a classical task in low-level computer vision, aiming at recovering an potential clean image x from its degraded observation $y = Dx + n$, where D is the degradation matrix and n is the additive noise.

In the past few years, convolutional neural network[5] has shown great ability on IR tasks and achieve state-of-the-art performance. The **CNN** in **IR** task is a regression model that is trained with quantities of paired data (x_i, y_i) to minimize the difference between predicted result $f_\theta(y_i)$ and ground truth clean image x_i

$$\arg\min_{\theta} \sum_{i} L(f_\theta(y_i), x_i), \tag{1}$$

where f_θ is the mapping function from the degraded image y_i to the predicted clean image x_i' with trainable parameters θ under the loss function L.

However, there is a difficulty if we want to use this method to solve our problem. Training a CNN to solve IR problem requires a huge amount of paired degraded data and clean data. However, due to the characteristics of old films, it is almost impossible to get a lot of paired contaminated and clean data. One possible way to solve this problem is to using synthesis data to train the network. However, the degradation on old films are very complex. Using synthesis data will introduce the domain gap, which is a new and tougher problem. Wan et al.[8] proposed a triplet domain translation network by leveraging real photos along with massive synthetic image pairs to solve the doamin gap problem, and here we are trying to propose a different insight to solve the problem.

By observing the red, green, and blue channels of the old negative films, we discover a property of the degradation on these old negative films, which may be helpful to our research. Fig.1 shows the red, green and blue channels of old negative films respectively. It is easy to notice that the degradation severity of different channels are different. Generally speaking, the blue channel suffers most while the deterioration on red and green channel are much more mild. It may because that the blue light sensitive layer is most exposed to the external environment compared to the other channels.

(a) Blue. (b) Green. (c) Red.

Figure 1: The three channels of a degraded negative film

Based on this discovery, we come up with a new method to restore those old negative films. Instead of recognizing and removing those contamination based on the whole image, we use cross channel information to remove the contamination on blue channel. Since the degradation on blue channel is the most severe, We can remove most of contamination of the image if we can restore the blue channel. As a result, when synthesizing data, we only add contamination on blue channel and use the relationship between RGB channels to remove the contamination on blue channel when designing and training the network. This method will have extra information based on the relationship between RGB channels, which can alleviate the dependency on the context information and the domain gap problem.

So, the task in this project can be described as follows: Given the clean green and red channels and contaminated blue channel of a picture, remove the contamination on the blue channel and restore the original clear image.

The main contribution of our work is as follows:

1. We proposed a new insight for image restoration, which used two clean channels to guide the restoration of one degraded channel.

2. We provide a new method for synthesizing the artificial training data, which has more arbitrary shape compared to the rectangle or circle shaped mask in previous methods.

2 RELATED WORK

In this project, we can take our task as two constrained sub-tasks. First, we need to generate a underlying blue channel based on the information from red and green channel, which can be regarded as a constrained colorization task. And then, we need to restore the original blue channel based on the underlying blue channel and the real contaminated blue channel, which can be taken as an inpainting task with extra auxiliary information.

Image colorization. Image colorizaiton is to hallucinate the plausible color version given a grayscale photograph as input. This task need to restore other two channels based on the information of one channel while our task only need to restore one channel based on the information of other two channels. Zhang et al.[11, 12] proposed a colorization network for the image colorization task and achieves satisfying result. Based on the model they proposed in the second paper,we design our cross channel prediction network on RGB space to serve as a sub-network in our overall network design.

Inpainting. Inpainting is a task to complete the missing content in an image. Since we provide the mask and other information as extra information. It can be regarded as a non-blind inpainting task. Wang et al. [9] proposed VCNET for blind inpainting task and performs well. We use their data preparation method to synthesize part of our training data.

3 METHOD

In this section, we show the architecture of our model in detail. In section 3.1, we describe the overall structure of our network. Then we show

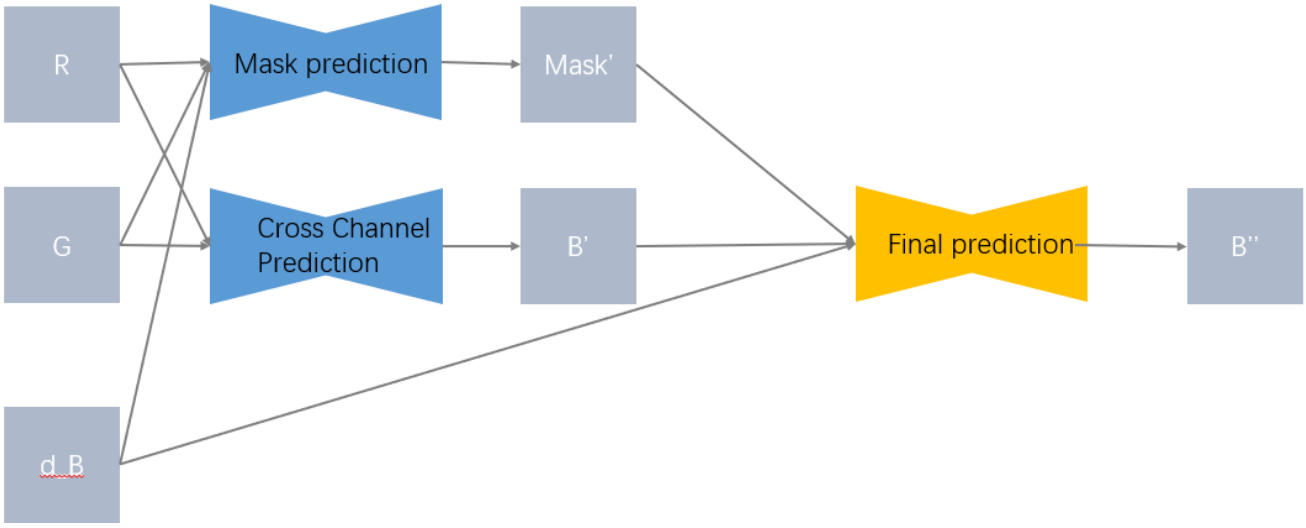


Figure 2: The Overview of Network.

the basic Unet[7] component in section 3.2. And three sub-networks are introduced in section 3.3.

3.1 Network Overview

The architecture diagram of proposed model is shown in Fig.2. The network takes the clean red and green channel as well as contaminated blue channel as input and feed them into different sub-networks as below. The network consists of three parts, the mask prediction sub-network, the cross channel prediction sub-network, and the final prediction sub-network. The detailed structure of these three sub-networks are described in the following sections.

3.2 Unet Component

The three sub-networks share the similar Unet based architecture. In this section, we will describe the structure of this Unet based network and point out some special design in our network.

The diagram in Fig.3 is the architecture of cross channel prediction network. The other two sub-networks have the similar structure with it. The only difference is the number of input channels. Hence, in this section, we will focus on this cross channel prediction network to introduce the architecture of this Unet based network.

The input channels are different in different sub-networks. The mask prediction network takes the clean red and green channel as well as contaminated blue channel as input. While the cross channel prediction network takes the clean red and green channel as input. And as for the final prediction network, its inputs are the mask predicted by the mask prediction network and the intermediate blue channel predicted by the cross channel prediction network as well as the contaminated blue channel. After the input channels are fed into the network, they are concatenated along the channel axis. After the concatenation, it becomes a tensor with shape of $H \times W \times C$, in which H, W, C represent the height, width and number of channels respectively. The concatenated tensor is then fed into ten convolutional blocks, which serves as an encoder-decoder network[1]. Each of convolutional block is made up of two or three Convolution-ReLU layers. The first six blocks are encoding blocks, which extract the features of the input tensor. The height and width of tensor will become half of before while the number of channels will become twice as before after processed by first four encoding block, while the last two

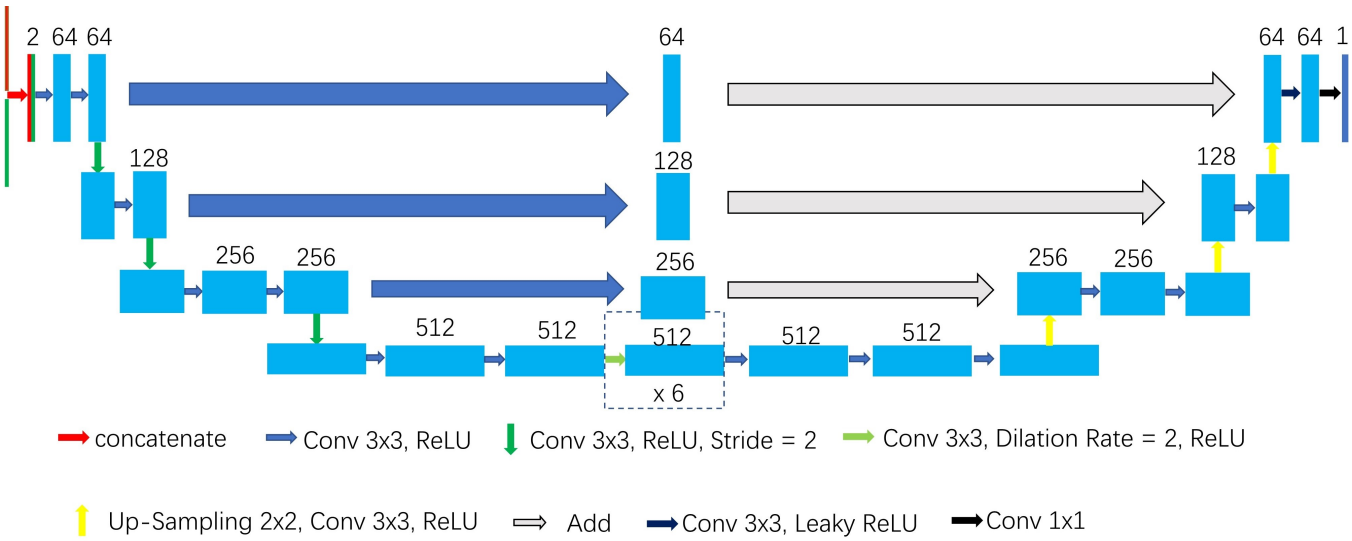


Figure 3: The Unet Component.

encoding block do not change the shape of tensor. The last two encoding blocks are dilated convolutional layers, which can increase the receptive field without introducing extra cost. The detailed functionality of dilated convolutional blocks will be explained later. And the last four convolutional blocks are decoding blocks, which turns the features back to its original domain. The height and width of these tensors will become twice as before and the number of channels will be halved after processed by each decoding blocks. The overall architecture of this Unet component is described before, and some important features will be introduced below.

Symmetric shortcut connections. The symmetric shortcut connection is the key feature of Unet, which means that the convolutional layers that have the same shape in encoders and decoders are connected respectively.*i.e.*, the first convolutional block in the encoder part are connected to the last convoluntional block in the decoder part, the second convolutional block in the encoder part are connected to the second last block in decoder, and so on. This connection between the encoder and decoder can help the network to pass the low level information more easily.

Batch Normalization at the end of every block. Since our network is a very deep network, Batch Normalization[3] is necessary in our network design. There is a Batch Normalization layer at the end of each convolutional block to reduce the covariate shift. The Batch Normalization layer can normalize the distribution of input tensor so that to alleviate the vanishing gradient and exploding gradient problem, which makes the network more easier to train and more robust.

Dilated Convolutional layer at the bottleneck of network. The dilated filter[10] is used in the fifth and sixth convolutional blocks to increase the receptive field of our network. The network with larger receptive field have stronger ability to capture the context information, so that its prediction can be more precise. The conventional method to enlarge the receptive field is to increase the network depth or the convolution kernel size. They both introduce more parameters to the network and increase the computational burden. Moreover, the increased depth of network makes the network more easier to diverge. The dilated convolution makes a balance in this tradeoff. It does not introduce more parameters or complicate the network architecture but increase the receptive field, which improve the network performance.

3.3 Roles of Sub-networks

In this section,we will describe the functionality of each sub-networks.

3.3.1 Mask Prediction. The mask prediction sub-network is an auxiliary network in this task. It takes clean red and green channel as well as the contaminated blue channel as input and output a tentative mask, which indicate the contaminated area. The predicted mask is then passed to the final prediction network as a reference to produce the final result.

3.3.2 Cross Channel Prediction. The cross channel prediction network produce an intermediate blue channel for further prediction. It takes clean red and green channel as input and output a possible blue channel as an intermediate result. The intermediate result predicted by this sub-network has no contamination, but the color information of it is not accurate. It IS then passed to the final prediction network for the final prediciton.

3.3.3 Final Prediction. The final prediction network takes the outputs of previous two sub-networks and the contaminated blue channel as input to predict the ultimate result. Based on information provided by the original contaminated blue channel and the intermediate blue channel predicted by the cross channel prediction network, and with the assistance of predicted mask that indicate the contamination area, the final prediction network is expected to produce the final result which is free from contamination and has accurate color tone.

4 EXPERIMENTS

In this section, we will describe which dataset we use, and how we pre-process the data and how we train our model with this dataset. The training result will also be presented in this section.

4.1 Dataset

The dataset we use in this project is Microsoft COCO2017[6] dataset. It is a dataset which collects the images of various everyday scenes containing common objects in their natural context. There are 328k images of 91 objects types in the dataset, which is sufficient for our model training. The dataset is splited into three parts, the training set containing 118k images, the validation set containing 5k images and the test set which contains 41k images.

4.2 Data Preprocessing

The dataset we use only contains the clean red, green and blue channels. So we need to synthesize the contaminated blue channel and the mask that indicate the contamination area. In this section,we will describe how we synthesize the data we need. There are two steps in the data synthesis procedure. We first determine the shape of contamination and then generate the corresponding mask of that contamination shape. Then we determine the contamination content and synthesize the final contaminated blue channel.

(a) Continuous shape mask. **(b) Discrete shape mask.**

Figure 4: Two types of masks

4.2.1 Contamination Shape and Mask. Arbitrary rectangle or text-shaped mask are often used in restoration tasks like inpainting task. However, this kind of mask may encourage the model to locate the contamination area based on the shape information rather then the context information or the cross channel relationship we want. So, we use two types of free-from shape as the mask in

our task. One of them is learnt from VCNET[9]. It draws arbitrary lines with certain width to produce a free-form mask like the left one in Fig.4. The other type of mask is designed by us. We first randomly draw several circles with reasonable radius, and then randomly choose several points in each circle we draw in the last step. These points will be the center of our new circle, whose radius are smaller then the previous step. After repeating this procedure for several times, we get the mask like the right one in Fig.4. It is more similar to the shape of contamination we encounter in the negative films.

(a) Discrete shape with another image as noise. (b) Continuous shape with another image as noise.

Figure 5: Examples of contaminated image

4.2.2 Contamination Content and Synthesis Result. After determining the shape of contamination, we need to decide what content should we place in the contaminated area. We design three types of contamination content, they are constant grayness, 50% impulse noise, and other natural images. The natural images used as noise are from the VOC2012 dataset[2]. After determining both contamination shape and contamination content, we can generate the contaminated image as

$$O = M \odot N + (1 - M) \odot I \tag{2}$$

where I and O is the original clean image and generated contaminated image respectively,M is the contamination mask and N is the contamination content. And $\odot$ is the Hadamard product operator. We apply Gaussian smoothing on M and employ alpha blending in the contact area between O and N to avoid noticeable edges, which may provide redundant information to the network and affect the performance of network. The result of contaminated images we generate are shown in Fig.5

4.3 Training Scheme and Result

4.3.1 Training Scheme. The overall network takes three inputs, clean red channel, clean green channel, and contaminated blue channel, they are combined and fed into the network according to the description in section 3.3. And the network also have three outputs, the predicted mask, the first predicted blue channel and the final predicted blue channel, we use M', I' and I'' to denote them respectively. The loss of mask prediction is evaluated by binary cross entropy

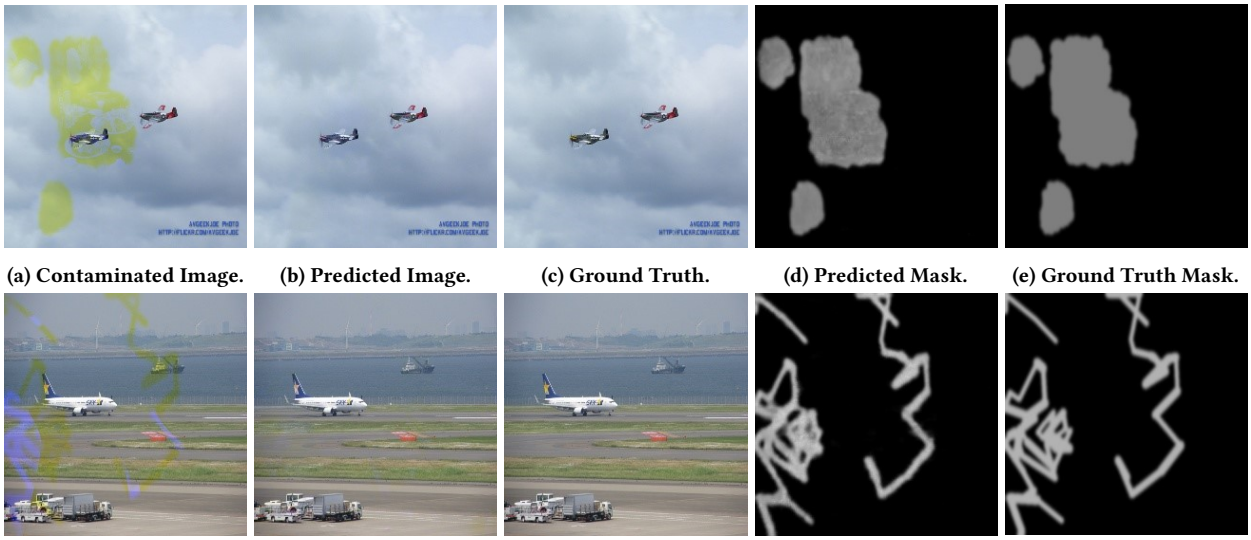
$$L = -(ylog(p) + (1 - y)log(1 - p)) \tag{3}$$

where y is the binary indicator if class label c is the correct classification for observation o and p is the predicted probability observation o is of class c. And the loss of predicted blue channel is evaluated by MSE

$$L = ||b_i, f_\theta(g_i, r_i)||^2, \tag{4}$$

, where b_i is the ground truth of the blue channel, g_i,r_i are the input red and green channels, and f_θ is the model we use. There are about 96M parameters in the whole network. For the training hyper-parameters, the learning rate is set to $1e-4$, and the optimizer we use is Adam[4]. The batch size is set to 32 so there are 3697 steps per epoch. After 20 epochs training, the result is presented in the following sections.

4.3.2 Result of Prediction. To show the effectiveness of our model design, we also train two other networks to do ablation study. One of them is a basic Unet which takes the clean red and green channel as well as the contaminated blue channel as input and predict the clean blue channel directly. The other one is the proposed model without the mask prediction sub-network. The quantitative result of these networks are shown in Tab.1. Accord-



(a) Contaminated Image. (b) Predicted Image. (c) Ground Truth. (d) Predicted Mask. (e) Ground Truth Mask.

(f) Contaminated Image. (g) Predicted Image. (h) Ground Truth. (i) Predicted Mask. (j) Ground Truth Mask.

Figure 6: Qualitative result

Network	PSNR	SSIM
Unet	33.21	0.9671
Proposed model without mask	33.75	0.9702
Proposed model with mask	35.73	0.9781

Table 1: Quantitative results

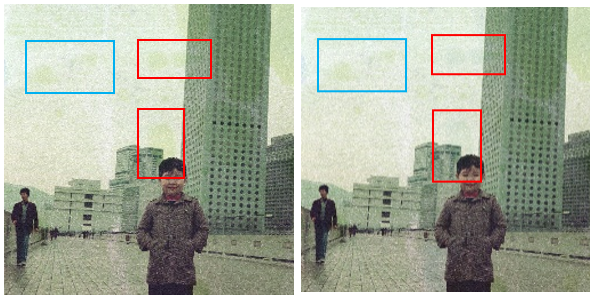
ing to the quantitative result above, we can find that the proposed model with mask prediction outperform the other two networks and have the best performance on the synthesis dataset. Here are some visualized qualitative results of prediction result in Fig.6.

5 DISCUSSION

5.1 Test Model on Real Data

Here is the visualized qualitative result of our network prediction on real negative films in Fig.7. Since we do not have the clean image, we cannot show the quantitative result. By observing the qualitative result, we can find that although our model can remove some obvious contamination on the film(marked by red rectangle), but it still cannot remove them thoroughly(marked by green rectangle), and some complex contamination cannot be recognized and removed, like the small contamination on the top of the image and the big contamination covers both the building and the sky at the right top corner.

Our model performs quite well on our synthesis contamination dataset. But it still have some limitations on the real negative films.



(a) The degraded old film (b) The output of our model

Figure 7: Result of Real Old Film

6 CONCLUSION

In this project, we proposed a new method to remove the contamination on a partially contaminated image(*i.e.*, only one channel of the image is

severely degraded) based on the cross channel relationship between the RGB channels. The network can find the contaminated area on the contaminated channel and then remove the contamination on the that channel. Our model works quite well on our synthesized dataset. However, it has some limitations on the real partially degraded images. Although we think the cross channel relationship can alleviate the dependency on the shape of contamination, the domain gap between the synthesis data still affect the performance of our model on real degraded films. It can only remove the noticeable contamination while cannot deal with the small stains and contamination in complex area. We are still trying to improve the performance of our model on real degraded data by modify the network architecture and redesign the loss term. Hopefully we can solve the domain gap problem and make our model have as great performance as on the synthesis data.

REFERENCES

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence* 39, 12 (2017), 2481–2495.

[2] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision* 88, 2 (2010), 303–338.

[3] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167* (2015).

[4] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2017. Imagenet classification with deep convolutional neural networks. *Commun. ACM* 60, 6 (2017), 84–90.

[6] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

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

[8] Ziyu Wan, Bo Zhang, Dongdong Chen, Pan Zhang, Dong Chen, Jing Liao, and Fang Wen. 2020. Bringing old photos back to life. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2747–2757.

[9] Yi Wang, Ying-Cong Chen, Xin Tao, and Jiaya Jia. 2020. VCNet: A Robust Approach to Blind Image Inpainting. *arXiv preprint arXiv:2003.06816* (2020).

[10] Fisher Yu and Vladlen Koltun. 2015. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122* (2015).

[11] Richard Zhang, Phillip Isola, and Alexei A Efros. 2016. Colorful image colorization. In *European conference on computer vision*. Springer, 649–666.

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