Two Unpaired Images Colorization

Suhyeon Ha
School of Computer Science and Engineering, Chung-Ang University
Seoul, Korea
tngus3752@gmail.com

Abstract

Colorization task usually requires large-scale dataset to produce satisfactory results. However, collecting colorized data can be demanding in specific fields such as cartoons or animations. Also training over a million images as in existing methods is also time-consuming. In this paper, we suggest a generative model to colorize a given grayscale image with a reference without any other images. Multi-scale hierarchical GANs enable to colorize considering structures and colors of given two unpaired images. Experimental result shows that the proposed method can generate plausible colorized results with just tiny-scale data.

1. Introduction

Colors play a key role in understanding the world in that color images carry more information visually. Color images are considered completely different from grayscale images in aesthetics. Though deep learning has promoted the success of colorization performance, the fact that too much data and time is needed is undeniable. In detail, Iizuka *et al*. [4] used Places dataset[15] containing 2M training images. Larsson *et al*. [7] and Zhang *et al*. [14] made use of ImageNet dataset[9] including over 1.2M training images.

Recently, some current methods[13] tackle the issue. Yoo *et al*. achieved few-shot colorization using memory augmented networks. Yoo *et al*. utilized five different datasets which have a range of about 35 images in minimum and 10K images in maximum.

In this paper, we suggest a novel generative model which can colorize a grayscale image with just a single reference without other training images. Our proposed network using multi-scale of patchGANs learns the structure of given grayscale and colors of a reference. Proposed conditioned discriminators enable generators to generate a colorized image without any ground truth image. The experimental results show that our method is potential to colorize with only tiny-scale data.

2. Related work

2.1. Deep learning-based colorization

According to the success of deep learning in computer vision, convolutional networks with large-scale dataset[14, 7, 4] has been utilized as one of methods to solve colorization problems. These methods known as fully automatic colorization train a model over a million images and produce a colorized image. However, the methods which don't allow any user controllability are difficult to satisfy users with the single colorized image.

To reflect the users' desire on colorization results, reference-based methods[3, 11, 12] has been explored. Colorization networks transfer colors of given reference to a grayscale image so that users are able to gain diverse colorized images according to different references.

On the other hand, some works[13] attempt to exploit small-scale data to relieve burdensome effort to collect significant amount of data, especially in cartoons and animations fields.

In this paper, we aim to build colorization networks which only needs two images for training and testing: one for grayscale and one for reference.

2.2. Internal learning

Internal learning which is a learning from a single training image has achieved remarkable results in image manipulation tasks.

SinGAN[10] shows one single generative model can produce results in multiple tasks in a way to reconstruct the given image. Since multiple scale architecture of patchGANs[1] is able to capture internal statistics of each scale of a given single image. However, the network hardly handle any colors or structures which it haven't seen before because the networks only know about the single training image.

TuiGAN[8] is two symmetric pyramids of GANs to solve image-to-image translation task with two unpaired images. However, TuiGAN is only able to learn transformations between color images.

In this paper, we attempt to consider the relations between two unpaired images to colorize a grayscale image into a reference.

3. Proposed method

The goal of our model is to generate a colorized image which imitates the reference's colors maintaining the structure of given grayscale image. In this paper, we suggest to train two unpaired images with a generative model following the architecture of multiple patchGANs similar to SinGAN[10].

3.1. Overall framework

Given a grayscale target image T and reference image R, $\{T_0, \cdots, T_N\}$ and $\{R_0, \cdots, R_N\}$ are sets of downsampled versions of T and R respectively by a factor r^n, for some $r > 1$. According to resolutions of T, we train generators G_n and discriminators D_n for each scale. Generator tries to generate patches to fool the discriminator and discriminator attempts to distinguish real patches from generated fake patches. As the number n is decreasing, from N to 0, size of receptive field becomes typically $\sim 1/2$ of previous scale image's height to produce more details in finer scale.

3.2. Training

At scale $n = N$, new image is generated from White Gaussian noise of three channels denoted as z_N as follows

$$\tilde{T}_N = G_N(z_N). \tag{1}$$

From $\{0, \cdots, N\}$, each scale has same architecture which contains a generator and a discriminator. At scale of n, the colorized $\tilde{T}_n$ are upscaled with respect to factor r and fed into next scale generation. Then, noise z_n is injected to allow some diversity during colorization as follows

$$\tilde{T}_n = G_n(z_n, (\tilde{T}_{n+1}) \uparrow^r), n < N. \tag{2}$$

Generator G_n and discriminator D_n both are consisting of 7 conv-blocks. Generator learns to produce a colorized results to fool the discriminator. The generated image $\tilde{T}_n$ is separated into two elements(grayscale and color palette) and fed into the discriminator. The discriminator attempts to distinguish real samples from fake samples. Real samples consist of given grayscale image T_n and a color palette from given reference R_n. Fake samples are the sets of generated grayscale image and color palette. We used K-means clustering to extract a color palette consisted of five dominant colors.

In other words, the overall procedure is to generate a colorized image reflecting the structure of given grayscale image T_n and colors of given reference R_n. The generator G_n performs the operation as follows:

$$\tilde{T}_n = (\tilde{T}_{n+1}) \uparrow^r + \psi_n(z_n) + (\tilde{T}_{n+1}) \uparrow^r, \tag{3}$$

where ψ_n denotes a 7 conv-blocks containing Conv(3×3)-BatchNorm[5]-LeackyReLU. For the discriminator, Markovian discriminator (patchGANs[1]) which has $11x11$ patch-size is used.

3.3. Objective function

The loss function is a weighted sum of adversarial loss, reconstruction loss and content loss as follows:

$$\min_{G_n} \max_{D_n} \mathcal{L}_{adv}(G_n, D_n) + \alpha_r \mathcal{L}_{rec}(G_n) + \alpha_c \mathcal{L}_{content}. \tag{4}$$

We use the WGAN-GP[2] for adversarial loss and the loss $\mathcal{L}_{adv}$ penalizes for the distance between the distribution of patches in T_n and the distribution of patches in generated samples $\tilde{T}_n$. The discriminator plays an important role in combine two unpaired images which are different in structures and colors.

At the beginning of training, generating plausible image is more difficult than differentiating real images from fake ones. To elevate generator's ability, we define the reconstruction loss similar to SinGAN[10]. However, we set $\{z_N^{rec}, z_{N-1}^{rec}, \cdots, z_0^{rec}\} = \{z_N^*, z_{N-1}^*, \cdots, z_0^*\}$, where z^* denotes the fixed noise map. A fixed noise map is set for each scale and kept fixed during training. It allows the generated image to have globally similar but locally different structures from given reference R_n. The reconstruction loss is defined as follows:

$$\mathcal{L}_r = \left\| G_n(z_n^*, (\tilde{T}_{n+1}) \uparrow^r) - R_n \right\|^2, n < N, \tag{5}$$

and $\mathcal{L}_r = \|G_n(z^*) - R_n\|^2$ for $n = N$.

The structure of given grayscale T_n has to be remained in generated image. To force the generated image to preserve the structure of given grayscale T_n regardless of color changes, the content loss is defined as follows:

$$\mathcal{L}_c = \left\| T_n - \tilde{T}_n \right\|^2 \tag{6}$$

4. Experiment

4.1. Implement details

We trained our network using Adam[6] with leaning rate 0.0005. We set the scale factor as 0.75, the number of clusters for color palettes as 5. We experimented all test examples with the weight parameters $\alpha_r = 10, \alpha_c = 30$. We used two images(grayscale and refence) for training and testing with $2,000$ iteratons for each scale. We performed the experiments on a workstation with an NVIDIA GeForce RTX 2080Ti GPU. The proposed network was implemented using Pytorch.

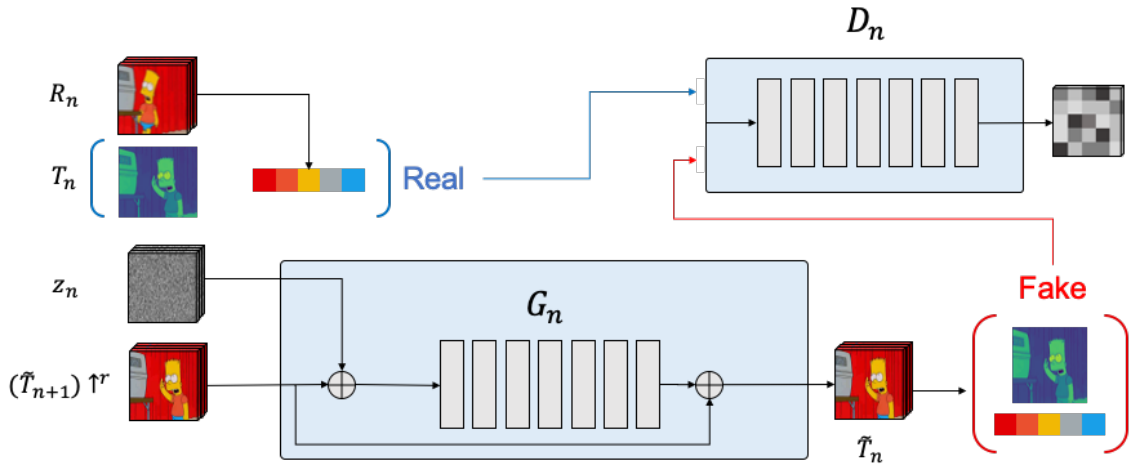

Figure 1. **Single scale generation.** At each scale, down-sampled version of two unpaired images(R_n for reference and T_n for target grayscale image) are given. A newly generated noise z_n, up-scaled colorized image carried from previous scale $\tilde{T}_{n+1}$ are fed into the generator G_n. Discriminator takes a set of grayscale and extracted color palette to tell whether the image is real or fake.

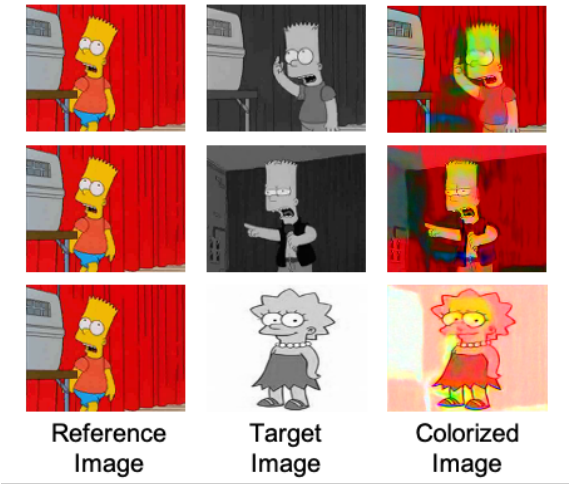

Figure 2. **Qualitative results.** Given a reference image and a target image, our proposed method generate a plausible colorized image.

4.2. Qualitative results

Figure 2 shows the qualitative result of our proposed method. Given a reference(colored) and a target(grayscale) image, our networks can generate the colorized image which has same structure with target and colors of reference. Note that existing colorization methods require over a million images for training. Our method only knows about two images but it finds some relation between input images and colorizes according to their relation. Though our method has to be improved in spatially consistency, the method shows a potential of tiny-scale image colorization.

5. Conclusion

Existing methods usually demand large-scale dataset to generate visually satisfying results. However, collecting and training over a million data can be burdensome. To solve the issues, we suggest to train a network to colorize with two unpaired images. The qualitative results represents colorization is operated in semantically matching manner though area to be colored is changed. For future work, an improvement for spatially consistency has to be processed for better satisfactory results.

References

[1] Ugur Demir and Gozde Unal. Patch-based image inpainting with generative adversarial networks. *arXiv preprint arXiv:1803.07422*, 2018.

[2] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in neural information processing systems*, pages 5767–5777, 2017.

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

[4] 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 Transactions on Graphics (ToG)*, 35(4):1–11, 2016.

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

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

[7] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In *European conference on computer vision*, pages 577–593. Springer, 2016.

[8] Jianxin Lin, Yingxue Pang, Yingce Xia, Zhibo Chen, and Jiebo Luo. Tuigan: Learning versatile image-to-image translation with two unpaired images. *arXiv preprint arXiv:2004.04634*, 2020.

[9] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[10] Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. Singan: Learning a generative model from a single natural image. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4570–4580, 2019.

[11] Chufeng Xiao, Chu Han, Zhuming Zhang, Jing Qin, Tien-Tsin Wong, Guoqiang Han, and Shengfeng He. Example-based colourization via dense encoding pyramids. In *Computer Graphics Forum*, volume 39, pages 20–33. Wiley Online Library, 2020.

[12] Zhongyou Xu, Tingting Wang, Faming Fang, Yun Sheng, and Guixu Zhang. Stylization-based architecture for fast deep exemplar colorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9363–9372, 2020.

[13] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11283–11292, 2019.

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

[15] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27, pages 487–495. Curran Associates, Inc., 2014.