

Coloring With Limited Data: Few-Shot Colorization via Memory Augmented Networks

Seungjoo Yoo¹

seungj ooyoo@korea. ac. kr

Hyojin Bahng¹

hjj 552@korea. ac. kr

Sunghyo Chung¹

s94021@korea. ac. kr

Junsoo Lee¹

j unsul ee@korea. ac. kr

Jaehyuk Chang²

j aehyuk. chang@webtooncorp. com

Jaegul Choo¹

j choo@korea. ac. kr

¹ Korea University

² NAVER WEBTOON Corp.

Abstract

Despite recent advancements in deep learning-based automatic colorization, they are still limited when it comes to few-shot learning. Existing models require a significant amount of training data. To tackle this issue, we present a novel memory-augmented colorization model MemoPainter that can produce high-quality colorization with limited data. In particular, our model is able to capture rare instances and successfully colorize them. We also propose a novel threshold triplet loss that enables unsupervised training of memory networks without the need of class labels. Experiments show that our model has superior quality in both few-shot and one-shot colorization tasks.

1. Introduction

When Dorothy stepped into Land of Oz in the 1939 movie *Wizard of Oz*, a transition from black and white to vibrant colors makes it one of the most breathtaking moments in the history of cinema. There is no doubt to colors being an effective tool of expression, but they usually come at a cost. Coloring images is one of the most laborious and expensive stages when making modern day animation movies and comics. Automating the colorization process can help to reduce both cost and time required in producing comics or animated movies.

Despite advances in deep learning-based colorization models [33, 7, 11, 34], they are still limited when it comes to real-world applications like coloring animations and cartoons. There exist two main problems that make it difficult to use deep colorization models in real-world settings.

First, data for animations and cartoons are often limited, but training deep learning-based colorization models requires a large amount of data. Cartoon images are difficult to create because they must be drawn and intricately

Figure 1. Not much data in your hands? Make the most out of your limited data with our fully-automatic colorization model *MemoPainter*. (Res-cGAN is *MemoPainter* without the memory networks.)

colored by hand. In contrast, obtaining real-world images is easier because they can be taken by a camera and simply converted to grayscale. This leads to cartoon data not being as abundant as real-world images. Numerous existing colorization models are trained on real-world images, and their

Figure 2. **Dominant color effect commonly encountered by deep colorization models.** Deep colorization models tend to ignore diverse colors present in a training set and opt to learn only a few dominant colors. Using the most dominant color can be effective in minimizing the overall loss but yields unsatisfactory results. One can see that the outputs of [34] are dominated by the most prevalent color (red).

application is mostly limited to coloring old legacy photographs. This task is no longer needed because modern-day photographs are produced in color. Thus, learning to color animations and cartoons with little data would allow a more practical application of deep colorization models.

Second, existing colorization models ignore rare instances present in data and opt to learn the most frequent colors to generalize over the data. Remembering rare instances is important when diverse characters appear in a story that we want to color. Rare side characters will be ignored by colorization networks and all side characters will be colored similarly to the main character. Existing colorization models suffer from the dominant color effect, illustrated in Fig. 2. This effect occurs when a colorization model only learns to color with a few dominant colors present in the training set. This leads to existing models being unable to preserve *color identity*, which we define as the distinctive colors that separate a particular object class from the other. An example of color identity can be found in flowers. Different flower classes are distinguished by both their color and shape (buttercups are yellow and roses are red). Coloring in the most dominant color may succeed in producing plausible and natural-looking outputs, but each image loses its color identity.

We aim to alleviate these problems with our novel memory-augmented colorization model *MemoPainter*. To the best of our knowledge, there has been no colorization networks augmented by external neural memory networks. The main contributions of this paper include:

(1) Our model can learn to color with little data, allowing one-shot or few-shot colorization. This is possible because our memory networks extract and store useful color information from a given training data. When an input is given to our model, we can query our external memory networks to extract color information relevant to coloring the input.

(2) Our model can capture images of rare classes and suffer less from the dominant color effect, which previous

methods have not been able to accomplish.

(3) We present a novel threshold triplet loss, which allows training of memory networks in an *unsupervised* setting. We do not need labeled data for our model to successfully colorize images.

2. Related Work

Deep Learning-Based Colorization. Existing colorization methods [33, 34, 11] use deep neural networks to improve colorization performance. Zhang *et al.* [33] train convolutional neural networks and re-weights the loss function at training time to emphasize rare colors, yielding more vibrant results. Zhang *et al.* [34] incorporate local and global color hint information to increase colorization performance, which enables interactive colorization during test time. Isola *et al.* [11] use conditional generative adversarial networks to improve colorization performance as well as other image-to-image translation tasks. Although existing deep colorization methods produce high-quality results, they inevitably require large-scale data to train the deep neural networks. However, preparing abundant training data for real-world applications such as animation colorization is highly expensive as they need to be produced by professional animators. Moreover, deep colorization networks are successful *on average* (i.e., successful in coloring prominent objects yet failing in coloring rare instances). No previous studies have tackled few-shot colorization on rare instances, which is the main focus of this work.

Memory Networks. Several approaches have augmented neural networks with an external memory module to store critical information over long periods of time. It has been applied to solve algorithmic problems [6], perform natural-language question answering [28, 15, 17], and allow life-long and one-shot learning, especially in remembering rare events [12]. Other approaches have applied memory networks to store image data, specifically for image captioning [22, 23], summarization [13], image generation [14], and video summarization [16]. We are the first to augment colorization networks with memory networks to allow few-shot learning in image colorization.

Conditional Generative Adversarial Networks. Generative adversarial networks (GANs) [5] have achieved remarkable success in image generation. The key to its success lies in its adversarial loss, where the discriminator tries to distinguish between real and fake images while the generator tries to fool the discriminator by producing realistic fake images. Several studies leverage *conditional* GANs in order to generate samples conditioned on the class [18, 20, 21], text description [25, 31, 29], domain information [3, 24], input image [11, 30], or color features [2].

Figure 3. **Our proposed MemoPainter model.** Our model consists of memory networks and colorization networks. During training, memory networks learn to retrieve a color feature that best matches the ground-truth color feature of the query image, while the colorization networks learn to effectively inject the color feature to the target grayscale image. During test time, we retrieve the top-1 color feature from our memory and give it as a condition to the trained generator.

In this paper, we adopt the adversarial loss conditioned on a grayscale image and its color feature extracted from our memory module to generate colored images indistinguishable from real images.

3. Proposed Method

As illustrated in Fig. 3, our model *MemoPainter* is composed of two networks: memory networks and colorization networks. *MemoPainter* is the first model to augment colorization networks with memory to remember rare instances and produce high-quality colorization with limited data. Our memory networks are distinguished from previous approaches by how its key and value memory are constructed. We also introduce a new threshold triplet loss (TTL), which allows unsupervised training of memory networks without additional class label information. Finally, our colorization networks utilize adaptive instance normalization (AdaIN) [9] to boost colorization performance.

3.1. Memory Networks

We construct memory networks to store three different types of information: key memory, value memory, and age. A key memory \mathbf{K} stores information about spatial features of input data. The key memory is used to compute the co-

sine similarity with input queries. A value memory \mathbf{V} stores color features which are later used as the condition for the colorization networks. Both memory components are extracted from the training data. An age vector \mathbf{A} keeps track of the age of items stored in memory without being used. Our entire memory structure \mathbf{M} can be denoted as

$$\mathbf{M} = (\mathbf{K}_1, \mathbf{V}_1, \mathbf{A}_1), (\mathbf{K}_2, \mathbf{V}_2, \mathbf{A}_2), \dots, (\mathbf{K}_m, \mathbf{V}_m, \mathbf{A}_m), \quad (1)$$

where m represents the memory size. Our memory networks are inspired by the previously proposed architecture [12].

A query \mathbf{q} is constructed by first passing the input image \mathbf{X} through ResNet18-pool5 layers [8] pre-trained on ImageNet [4]. It is denoted as $\mathbf{X}_{rp5} \in \mathbb{R}^{512}$. We use feature vectors from pooling layers to summarize spatial information. For instance, a rose should be perceived as the same rose regardless of where it is spatially positioned in an image. We pass the feature representation through a linear layer with learnable parameters $\mathbf{W} \in \mathbb{R}^{512 \times 512}$ and $\mathbf{b} \in \mathbb{R}^{512}$. Finally, we normalize the vector to construct our query \mathbf{q} as

$$\mathbf{q} = \mathbf{W} \mathbf{X}_{rp5} + \mathbf{b}, \quad \mathbf{q} = \frac{\mathbf{q}}{\|\mathbf{q}\|}, \quad (2)$$

where $\|\mathbf{q}\|_2 = 1$. Given \mathbf{q} , the memory networks compute

the k nearest neighbors with respect to cosine similarity between the query and the keys $d_i = q \cdot K[i]$, i.e.,

$$\begin{aligned} \text{NN}(q, M) &= \arg\max_i q \cdot K[i], \\ (n_1, \dots, n_k) &= \text{NN}_k(q, M), \end{aligned} \quad (3)$$

and returns the nearest value $V[n_1]$, which is later used as the condition for the colorization networks.

Color Features. We leverage two variants to represent color information stored in value memory: color distributions and RGB color values. The former has the form of color distributions over 313 quantized color values, denoted as $C_{\text{dist}} \in \mathbb{R}^{313}$. It is computed by converting an input RGB image to the CIE *Lab* color space and quantizing the *ab* values into 313 color bins. We use the previously proposed parametrization [33] to quantize *ab* values. Color distributions are suitable for images with diverse colors and intricate drawings.

The second variant we use is a set of ten dominant RGB color values of an image denoted as $C_{\text{RGB}} \in \mathbb{R}^{10 \times 3}$, which is extracted from input images by utilizing a tool called Color Thief.¹ Using C_{RGB} as color features works better in one-shot colorization settings, as neural networks seem to learn easily and fast from direct RGB values than from complex color distribution information. In short, our value memory is represented as

$$V = C_{\text{dist}} \text{ or } C_{\text{RGB}}. \quad (4)$$

The color information extracted in the above-described manner is later used as a condition given to our colorization networks. Even though either or both of the variants can be used, we will use the notation C_{dist} for value memory in subsequent equations, so as not to confuse the reader.

Threshold Triplet Loss for Unsupervised Training

Previously proposed triplet losses [26, 12] aim to make images of the same classes (positive neighbors) closer to each other while making images of different classes (negative neighbors) further away. Likewise, we adopt the triplet loss to maximize similarity between the query and positive key and minimize similarity to the negative key. An existing supervised triplet loss [12] introduces the smallest index p where $V[n_p]$ has the same class label as an input query q . This would make n_p a positive neighbor of q . A negative neighbor of q will be defined as the smallest index b where $V[n_b]$ has a different class label from our query q .

However, this supervised triplet loss requires class label information, leading to its limited applicability in our setting as such information is not available in most data for colorization tasks. For instance, it would be almost impossible to label every single frame of an animation with its

¹<http://lokesdhakar.com/projects/color-thief/>

Figure 4. **How our model works during test time.** The top-1 color feature from our memory is retrieved and given as the condition to our trained generator.

class label (i.e., whether a particular character, object, or background appears in a given frame).

To solve this issue, we extend the existing method and propose a threshold-based triplet loss applicable to fully unsupervised settings. Given two images with similar spatial features, we assume that if the distance between their color features are within a certain threshold, then they are more likely to be in the same class than those images with different color distributions. We introduce this threshold as a hyperparameter denoted as τ . As the distance measure C_{dist} between two color features, we compute the symmetric KL divergence of their color distributions over quantized *ab* values. For C_{RGB} , we compute color distance using CIEDE2000 [27] by converting RGB values to CIE *Lab* values. In our unsupervised triplet loss setting, we newly define a positive neighbor n_p as the memory slot with the smallest index where the distance between $V[n_p]$ and correct desired value v (i.e., the color feature of the query image) is within a color threshold τ , i.e.,

$$\text{KL}(V[n_p] \parallel v) < \tau. \quad (5)$$

Similarly, we define a negative neighbor n_b as the memory slot with the smallest index where the distance between $V[n_p]$ and v exceeds τ , i.e.,

$$\text{KL}(V[n_b] \parallel v) > \tau. \quad (6)$$

Finally, the threshold triplet loss is defined as

$$\mathcal{L}_t(q, M, \tau) = \max(q \cdot K[n_b] - q \cdot K[n_p] + \tau, 0). \quad (7)$$

This triplet loss minimizes distance between the positive key and the query while maximizing distance between the negative key and the query.

Memory Update. Our memory M is updated after a new query q is introduced to the networks. The memory gets updated as follows, depending on whether the color distance between the top-1 value $V[n_1]$ and the correct value v (i.e., the color feature of the new query image) is within the color threshold.

(i) If the distance between $V[n_1]$ and v is within the color threshold, we update the key by averaging $K[n_1]$ and q and normalizing it. The age of n_1 is also reset to zero. In detail, the update when $KL(V[n_1] - v) < \tau$ is written as

$$K[n_1] = \frac{q + K[n_1]}{2}, \quad A[n_1] = 0. \quad (8)$$

(ii) If the distance between $V[n_1]$ and v exceeds the color threshold, this indicates that there exists no memory slot that matches v in our current memory. Thus, (q, v) will be newly written in the memory. We randomly choose one of the memory slots with the oldest age (i.e., the least recently used one), denoted as n_r , and replace that slot with (q, v) . We also reset its age to 0. In detail, when $KL(V[n_1] - v) > \tau$, the update is performed as

$$K[n_r] = q, \quad V[n_r] = v, \quad A[n_r] = 0. \quad (9)$$

3.2. Colorization Networks

Objective Function. Our colorization networks are conditional generative adversarial networks that consist of a generator G and a discriminator D . The discriminator tries to distinguish real images from colored outputs using a grayscale image and a color feature as a condition, while the generator tries to fool the discriminator by producing a realistic colored image given a grayscale input X and a color feature C . A smooth L_1 loss between the generated output $G(x, C)$ and the ground-truth image y is added to the generator's objective function, i.e.,

$$L_{sL1}(y, \hat{y}) = \begin{cases} \frac{1}{2}(y - \hat{y})^2 & \text{for } |y - \hat{y}| \leq \tau \\ |y - \hat{y}| - \frac{1}{2}\tau & \text{otherwise.} \end{cases} \quad (10)$$

This encourages the generator to produce outputs that do not deviate too far from the ground-truth image. Our full objective function for D and G can be written as

$$L_D = E_{x \sim P_{data}} [\log D(x, C, y)] + E_{x \sim P_{data}} [\log(1 - D(x, C, G(x, C)))], \quad (11)$$

$$L_G = E_{x \sim P_{data}} [\log(1 - D(x, C, G(x, C)))] + L_{sL1}(y, G(x, C)). \quad (12)$$

During training, we extract the color feature from the ground-truth image to train G and D . During the test time, we utilize the color value retrieved from the memory networks and feed it as the condition to the trained G , as shown in Fig. 4. We adapt the architecture of our generator networks from [11] and that of the discriminator from [2].

Figure 5. **Colorization results using the top-3 memory slots.** The memory networks can retrieve appropriate color features for a given input. Different memory slots may be used to produce diverse results. All other samples in the paper are colored using the top-1 memory slot.

Colors as style. Style transfer is a task of transferring a style of a reference image to a target image. Colorization can be viewed as style transfer, where instead of a particular style, color features are transferred to a target grayscale image. We will regard color as a style and from this perspective, we use AdaIN, which has shown success in style transfer, to effectively inject color information. We compute the affine parameters used in the AdaIN module by directly feeding the color feature to our own parameter-regression networks, i.e.,

$$\text{AdaIN}(z, C) = \frac{z - \mu(z)}{\sigma(z)} \cdot \sigma(C) + \mu(C), \quad (13)$$

where z is the activation of the previous convolutional layer, which is first normalized. Then it is scaled by $\sigma(C)$ and shifted by $\mu(C)$, which are parameters generated by a multilayer perceptron adapted from [10]. Compared to existing colorization models [2, 34] that incorporate color conditions via a simple element-wise addition, AdaIN allows the model to produce vivid colorizations as shown in Fig. 6.

4. Experiments

Our experiments consist of an ablation study on memory networks, analysis on the threshold triplet loss, and both quantitative and qualitative comparisons on three baseline models.

4.1. Qualitative Evaluation

4.1.1 Datasets

We perform experiments on five different datasets and compare our model performance on diverse settings (abundant data, few-shot, and one-shot).

Oxford102 Flower Dataset. The Flower dataset [19] consists of 102 flower classes. Each class has 40 to 258 images. The class labels are not used in our experiments.

Monster Dataset. 1,315 images are collected from the trailer of the movie *Monsters, Inc.* [1] to perform coloriza-

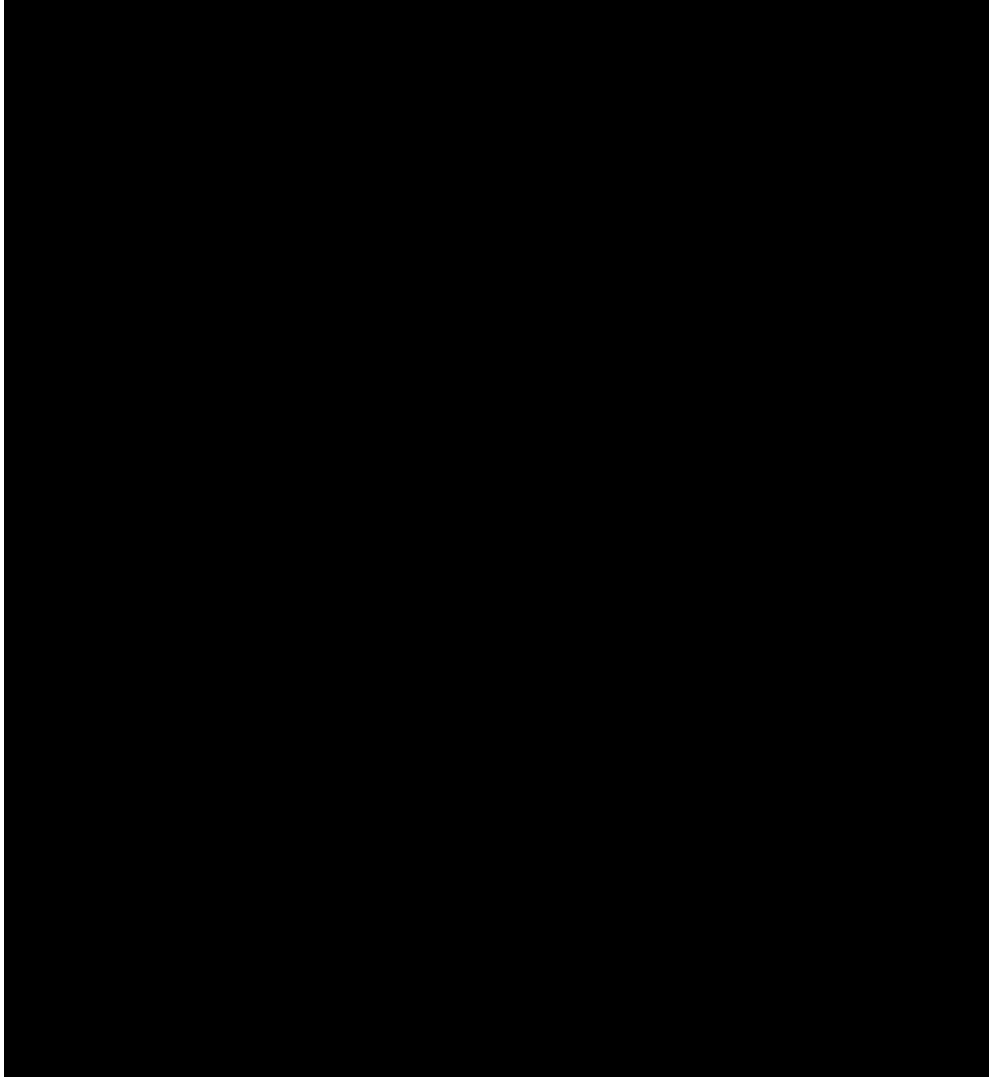


Figure 6. **Comparison to baselines on multiple datasets.** We compare our model with other colorization and image translation models: from left, Deep Priors [34], CIC [33], and Pix2pix [11]. Our model particularly excels at capturing and remembering objects that appear only a few times in a training set. The character in the first row is originally green, but he is drenched in pink paint in one scene. Other models color this character in green, but our model *MemoPainter* succeeds in remembering and coloring a scene where he is not green. Our model works even in settings with extremely limited data, where only one data item per class (last row) is available. In this one-shot learning setting, only our model manages to produce vibrant outputs.

tion on animations. An image frame is extracted every two seconds to reduce excessive redundancy in the dataset.

Yumi Dataset. We collect 9,955 images of the cartoon *Yumi's Cells*² to perform colorization on cartoons. It consists of images from 329 episodes, and each image is a single frame from of a sequence of each episode.

Superheroes Dataset. We collect images of superhero characters to perform few-shot colorization. It consists su-

perhero images from seven categories with less than five images per category.

Pokemon Dataset. We utilize the Pokemon dataset³ for one-shot colorization, which consists of 819 classes with a single image per class. Additional images are crawled from the internet to construct the test set.

²<https://comic.naver.com/webtoon/list.nhn?titleId=651673>

³<https://www.kaggle.com/kvpratama/pokemon-images-dataset>

Figure 7. **Analysis on memory networks.** We apply our model to a wide variety of datasets to show its applicability to different types of images ranging from diverse cartoons to real-world images. Our model shows superior performance especially in few-shot settings (first and third rows) when compared to our colorization networks (Res-cGAN) without memory networks. Colorization networks find it difficult to produce outputs with vivid color.

4.1.2 Analysis on Memory Networks

We run an ablation study to analyze the effect of augmenting colorization networks with memory. As shown in Fig. 7, we compare our proposed model *MemoPainter* against our colorization networks (Res-CGAN) without memory augmentation. Our memory-augmented networks are able to produce superior results on a wide variety of datasets from diverse cartoons to real-world images. In particular, it can accurately color an image even with only a single or few instances. Even though Res-cGAN produces high-quality colorization in most cases, it fails to preserve the ground-truth color of rare instances or completely fails in one-shot learning settings (e.g., results on the Pokemon dataset).

Moreover, an analysis on two hyperparameters (memory size and color threshold) is shown in Fig. 9. Performance is measured by comparing average Learned Perceptual Image Patch Similarity (LPIPS) [32]. Results show that LPIPS scores are stable across a wide range of hyperparameters and the model does not overfit to a particular color threshold or memory size.

4.1.3 Analysis on Threshold Triplet Loss (TTL)

The assumption behind the threshold triplet loss is that those images with (i) similar spatial features (i.e., the k nearest neighbors) and (ii) similar color features (i.e., color dis-

Figure 8. **Validation of our assumption on threshold triplet loss.** We demonstrate the corresponding images of the top-3 color features retrieved from our memory networks. By using the threshold triplet loss, our memory networks are trained to retrieve color features highly relevant to the content of the query image.

tance within a particular threshold) are likely to be in the same class. To confirm our assumption, we demonstrate the corresponding images of the top-3 color features retrieved from our memory networks during training. In this setting, we do not update the keys in the case $\text{KL}(V[n_1] \parallel v) < \epsilon$, so as to show its corresponding image. As shown in Fig. 8, we can see that the corresponding images of the top-3 color features have the same class as the query image. In particular, examples in the first row show that the top-3 images share the same character as well as similar clothes, objects, and backgrounds. This shows that TTL allows our memory

Figure 9. **Analysis of memory size and color threshold.** LPIPS scores are similar across various hyperparameters of the memory networks. Quality drops (high LPIPS) only with excessively small or large hyperparameters.

networks to retrieve color features relevant to the content of the query image, being able to color rare instances even if it was presented just once in the training data.

We also quantitatively validate the assumption of TTL by measuring the classification accuracy, evaluating whether the class of the images corresponding to the top-1 memory slot and the ground-truth label of the query are identical. During training, we additionally save the query’s class to compute classification accuracy. At test time, we compute the accuracy by computing the percentage of queries that has the same class as the top-1 memory slot. As an upper bound of our *unsupervised* method, we use a *supervised* version of our model. This version stores class values in the value memory and updates memory as presented by the life-long memory module made for few-shot classification [12]. Although our model is not specifically made for classification tasks, Table 2 shows that our unsupervised model (first row) retrieves the memory with the same class as accurately as the supervised method (second row) across a different number of classes and training sets.

4.1.4 Qualitative Comparisons

We qualitatively compare *MemoPainter* with three baselines: Deep Priors [34], CIC [33], and Pix2pix [11]. Fig. 6 shows qualitative results on multiple datasets. It shows that our model particularly excels at coloring rare instances in a training set. The Result from the monster dataset (first row) shows a rare scene where the main character (originally green) is drenched in pink paint. Although other models color this character in green, *MemoPainter* is able to remember this rare instance and color the character properly. Similarly, results from both the second and the fourth rows show that our model is able to remember rare classes along with minor details (i.e., even the clothes, objects, and backgrounds in a cartoon frame) while every other baseline fails to do so. Furthermore, our model is capable of producing high-quality results in both few-shot and one-shot learning settings compared to existing methods. Our model successfully produces accurate colorization given extremely limited data, e.g., given less than five training images per class

	One-shot		Few-shot	
	User-study	LPIPS	User-study	LPIPS
Ours	75%	8.48	71%	1.34
CIC	10%	9.89	7%	1.80
Pix2pix	5%	13.47	16%	2.34
Deep Prior	10%	19.26	4%	2.03

Table 1. **Quantitative comparisons with the state-of-the-art.** User study (higher is better) and LPIPS perceptual distance metric [32] (lower is better) shows superiority of our method.

	5-way		15-way	
	5-shot	10-shot	5-shot	10-shot
Ours (Unsup.)	87.50%	87.50%	69.44%	70.83%
Ours (Sup.)	91.66%	87.50%	72.22%	75.00%

Table 2. **Classification accuracy of the threshold triplet loss.**

(third row) or only a single data item per class (last row). In both cases, *MemoPainter* is the only model that can consistently produce accurate and vibrant colorization results.

4.1.5 Quantitative Evaluation

To quantitatively evaluate colorization quality, we conduct a user study with 30 participants, each answering 40 questions. We give a random source image and its corresponding colored outputs from our model and baselines. We then ask which generated output has the highest quality while maintaining the color identity of the source image (e.g., Hulk is green). We also compare the LPIPS distance [32] which is closer to human perception unlike MSE-based metrics. We compute the average LPIPS between the input image and its corresponding colored image. Table 1 shows that our model is superior to the state-of-the-art across both measures.

5. Conclusions

Results of this paper suggest that colorization networks with memory networks are a promising approach for practical applications of colorization models. We stress the importance of colorization models working with little data so that they can be used in coloring animations and cartoons. *MemoPainter* works on a wide variety of images, thus bearing great potentials in various applications that require few-shot colorization.

Acknowledgements. This work was partially supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIP) (No. NRF2016R1C1B2015924). We thank all researchers at NAVER WEBTOON Corp., especially Sungmin Kang. Jaegul Choo is the corresponding author.

References

- [1] Monsters, inc. **5**
- [2] H. Bahng, S. Yoo, W. Cho, D. K. Park, Z. Wu, X. Ma, and J. Choo. Coloring with words: Guiding image colorization through text-based palette generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 431–447, 2018. **2, 5**
- [3] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. **2**
- [4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. Ieee, 2009. **3**
- [5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014. **2**
- [6] A. Graves, G. Wayne, and I. Danihelka. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014. **2**
- [7] S. Guadarrama, R. Dahl, D. Bieber, M. Norouzi, J. Shlens, and K. Murphy. Pixcolor: Pixel recursive colorization. *BMVC*, 2017. **1**
- [8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. **3**
- [9] X. Huang and S. J. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, pages 1510–1519, 2017. **3**
- [10] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz. Multimodal unsupervised image-to-image translation. *ECCV*, 2018. **5**
- [11] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. **1, 2, 5, 6, 8**
- [12] L. Kaiser, O. Nachum, A. Roy, and S. Bengio. Learning to remember rare events. *International Conference on Learning Representations*, 2017. **2, 3, 4, 8**
- [13] B. Kim, H. Kim, and G. Kim. Abstractive Summarization of Reddit Posts with Multi-level Memory Networks. In *NAACL-HLT*, 2019. **2**
- [14] Y. Kim, M. Kim, and G. Kim. Memorization precedes generation: Learning unsupervised gans with memory networks. *International Conference on Learning Representations*, 2018. **2**
- [15] A. Kumar, O. Irsoy, P. Ondruska, M. Iyyer, J. Bradbury, I. Gulrajani, V. Zhong, R. Paulus, and R. Socher. Ask me anything: Dynamic memory networks for natural language processing. In *International Conference on Machine Learning*, pages 1378–1387, 2016. **2**
- [16] S. Lee, J. Sung, Y. Yu, and G. Kim. A memory network approach for story-based temporal summarization of 360 videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1410–1419, 2018. **2**
- [17] A. Miller, A. Fisch, J. Dodge, A.-H. Karimi, A. Bordes, and J. Weston. Key-value memory networks for directly reading documents. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016. **2**
- [18] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. **2**
- [19] M.-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing*, 2008. **5**
- [20] A. Odena. Semi-supervised learning with generative adversarial networks. *arXiv preprint arXiv:1606.01583*, 2016. **2**
- [21] A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier gans. *arXiv preprint arXiv:1610.09585*, 2016. **2**
- [22] C. C. Park, B. Kim, and G. Kim. Attend to you: Personalized image captioning with context sequence memory networks. 2017. **2**
- [23] C. C. Park, B. Kim, and G. Kim. Towards personalized image captioning via multimodal memory networks. *IEEE transactions on pattern analysis and machine intelligence*, 2018. **2**
- [24] A. Pumarola, A. Agudo, A. Martinez, A. Sanfeliu, and F. Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European Conference on Computer Vision (Proceedings of the European Conference on Computer Vision (ECCV))*, 2018. **2**
- [25] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text-to-image synthesis. In *Proceedings of The 33rd International Conference on Machine Learning (ICML)*, 2016. **2**
- [26] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. **4**
- [27] G. Sharma, W. Wu, and E. N. Dalal. The ciede2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. *Color Research & Application*, 30(1):21–30, 2005. **4**
- [28] S. Sukhbaatar, J. Weston, R. Fergus, et al. End-to-end memory networks. In *Advances in neural information processing systems*, pages 2440–2448, 2015. **2**
- [29] Q. H. H. Z. Z. G. X. H. X. H. Tao Xu, Pengchuan Zhang. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. 2018. **2**
- [30] D. K. P. I. S. J. C. Wonwoong Cho, Sungha Choi. Image-to-image translation via group-wise deep whitening and coloring transformation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. **2**
- [31] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. **2**
- [32] R. Zhang. The unreasonable effectiveness of deep features as a perceptual metric. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. **7, 8**

- [33] R. Zhang, P. Isola, and A. A. Efros. Colorful image colorization. In *European Conference on Computer Vision (ECCV)*, pages 649–666. Springer, 2016. **1, 2, 4, 6, 8**
- [34] R. Zhang, J.-Y. Zhu, P. Isola, X. Geng, A. S. Lin, T. Yu, and A. A. Efros. Real-time user-guided image colorization with learned deep priors. *SIGGRAPH*, 2017. **1, 2, 5, 6, 8**