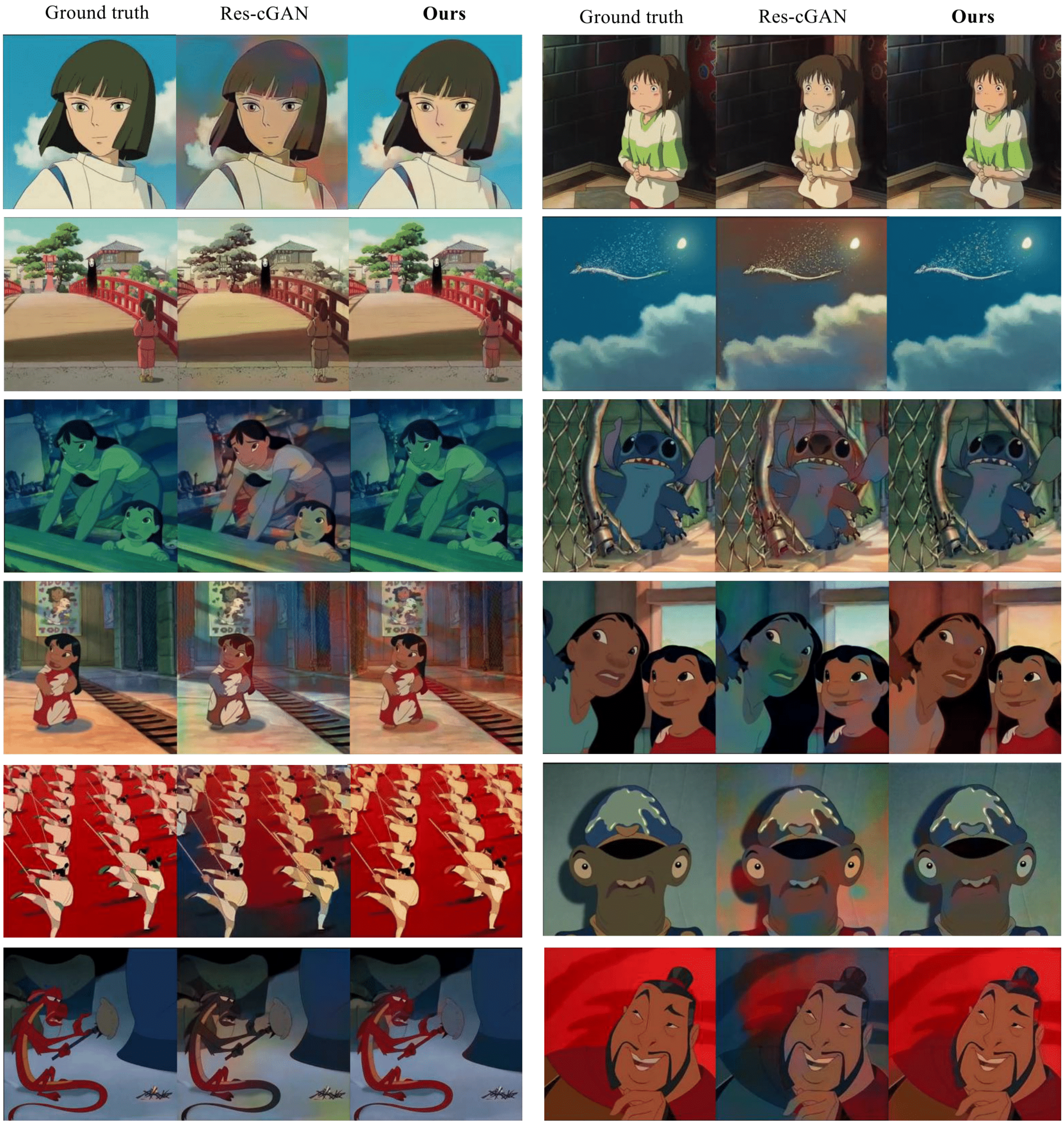

Figure 1. **Samples from *Lilo and Stitch*, *Mulan*, and *Spirited Away*.** Our memory-augmented model (*MemoPainter*) produces clearer and more vibrant outputs than colorization networks without memory networks (Res-cGAN) on complex cartoon images. Our model correctly colors various characters and diverse backgrounds even with limited data and no additional user inputs.

Appendix

1. Additional Few-Shot Learning Examples

We present additional few-shot colorization results on three other animation datasets: Mulan, Lilo and Stitch, and Spirited Away. Each dataset consists of 962,234, and 208 images, which are collected from Youtube. To test the few-shot learning capability of our model, we construct the datasets to have limited samples. Each training set consists of less than 10 images per scene, less than 50 images per main character, and less than 10 images per minor charac-

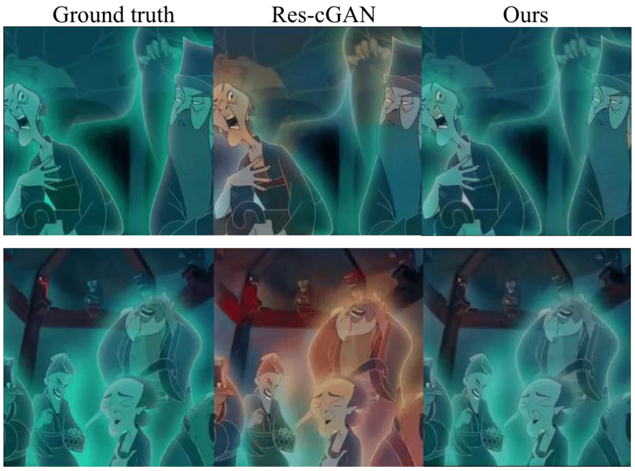

Figure 2. **Dominant color effect seen in human skin color.** Ghosts in the movie *Mulan* are blue. However, if spatial features in an input image resemble a human being, colorization networks without memory networks (Res-cGAN) will color the image in Asian skin colors. Our model successfully recognizes that the characters in the images are ghosts and colors them accordingly in blue.

	Ours	CIC	Pix2pix	Deep Priors
Parameters	**12m**	32m	14m	35m

Table 1. **Comparing the number of parameters with baseline models.** Even with external memory networks, our model has the smallest number of parameters. (m: million)

ter. The animation and cartoon datasets in the main paper (i.e., Monster dataset and Yumi dataset) were constructed in the same manner. Note that a scene is defined as a single event or conversation between characters, occurring during one period of time and in a single place [1]. In Fig. 1, we show that when given the same limited dataset, our *MemoPainter* excels in producing high-quality colorization compared to colorization networks without memory networks (Res-cGAN).

2. Comparisons in the Number of Parameters

Even with external memory networks, our model has the least number of parameters as shown in Table 1. Our memory networks requires only an additional 262k learnable parameters. Our model architecture is designed to be compact, and any model can be augmented with our memory networks with few additional parameters.

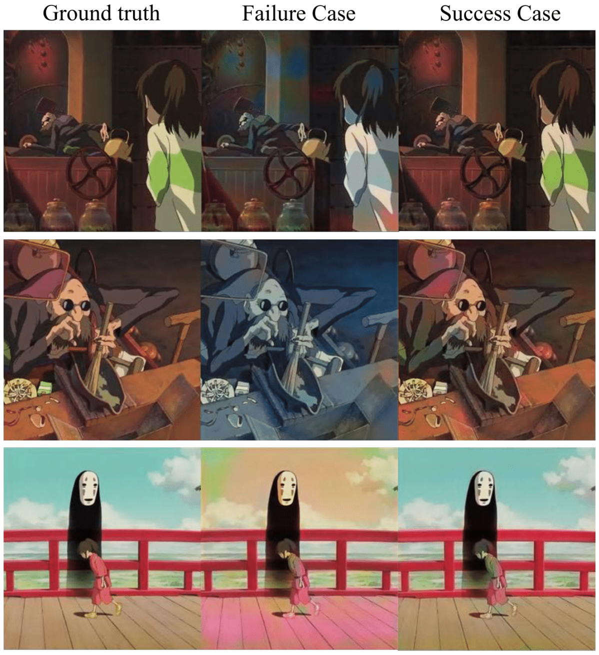

Figure 3. **Failure cases and success cases.** When our memory networks are insufficiently trained, they can retrieve an irrelevant memory slot for a query image as shown in the second column. Thus, careful optimization of memory networks is the key for producing good results.

	LPIPS	
Pre-trained	Oxford Flower	Mulan
ImageNet	**5.71**	**126.01**
Danbooru	15.79	126.91

Table 2. **Pre-training the feature extractor (real-image vs. cartoon).**

3. Training Details

We use the mini-batch size of 4, the learning rate of 0.0001, and select the Adam optimizer [2] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Regarding the memory networks, k is set to 256 when using C_{dist} and k is set to 32 when using C_{RGB} for computing the k nearest neighbors. For cartoon data, δ for C_{dist} was set between 0.5 to 0.7, and δ for C_{RGB} was set between 8 to 10. For real images, δ for C_{dist} was set between 0.7 to 0.9, and δ for C_{RGB} was set between 8 to 10. Setting the memory size m larger than 1.2 times the size of the training set yields best results.

4. Generalization Capability

Our model can generalize to not only images from the same dataset (i.e., movie and cartoon) but also to different datasets (drawing styles) of the same character. For

[1] https://www.awn.com/animationworld/animation-scene

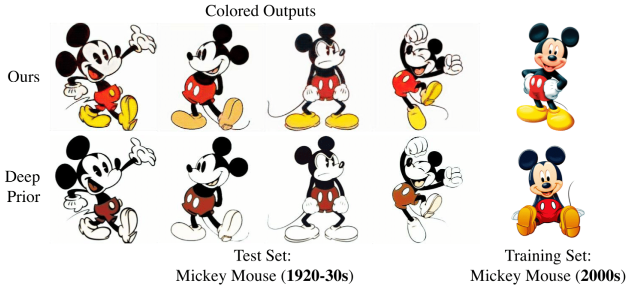

Figure 4. **Practical use case.** Our model can generalize to unseen cartoon styles of the same character and learn diverse vibrant colors.

instance, in Fig. 4, our model trained on modern Mickey Mouse can successfully color old Mickey Mouse images that originally exist only in black-and-white.

5. Pre-training on Comic Datasets

We compare the performance of using spatial features from the networks pre-trained on ImageNet and that of Danbooru [1], a large-scale cartoon database with 3.33 million images and 99 million tags. We use 1,000 most frequent tags and pre-train the ResNet18 for multi-class classification. Table 2 shows that using ImageNet achieves better performance for both cartoons and natural images, as ImageNet contains 1,000 objects while Danbooru contains only anime characters.

6. Failure Cases

While our *MemoPainter* can produce vibrant and high-quality outputs, it can also produce an output far from the ground truth. When the memory networks are insufficiently trained, they may retrieve an irrelevant memory slot for a query image as shown in Fig. 3. Thus, careful optimization of the memory networks is essential in producing good results. Also, We can mitigate this problem by using up to top-k memory slots rather than solely using the top-1 color feature. This solution works when the ground truth color value is within the top-k memory slots.

7. Low Cost of Memory Networks

Storing memory at a feature level requires little additional VRAM. We use 512 for our feature size, and storing 10,000 features in our memory networks only require an additional $10,000 * 512 * 32 = 19.53MB$ of memory. Storing training data into external memory networks may sound heavy at first, but it takes up little additional space.

References

[1] G. Branwen. Danbooru2018: A large-scale crowdsourced and tagged anime illustration dataset, 2019. 3

[2] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *ICLR*, 2015. 2