COCO-FUNIT: Few-Shot Unsupervised Image Translation with a Content Conditioned Style Encoder

Kuniaki Saito[1,2] Kate Saenko[1] Ming-Yu Liu[2]
Boston University[1] NVIDIA[2]

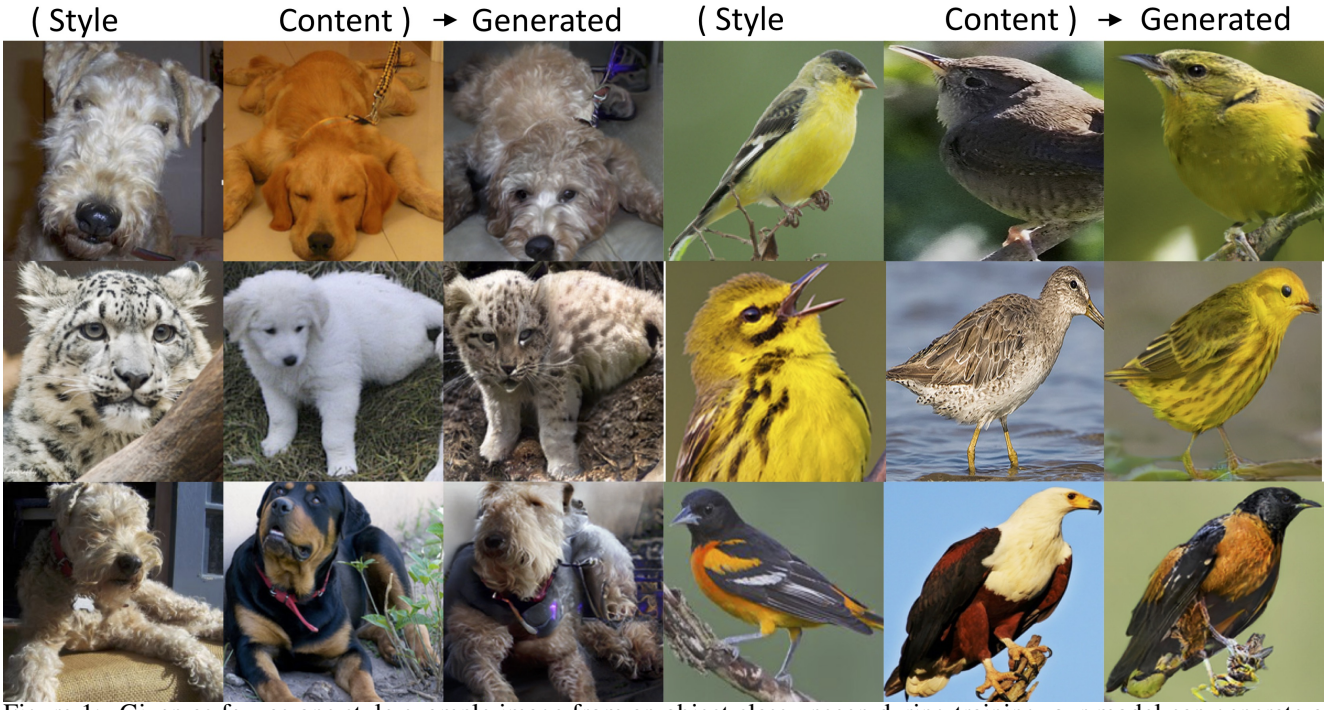

Figure 1: Given as few as one style example image from an object class unseen during training, our model can generate a photorealistic translation of the input content image to the unseen domain.

Abstract

Unsupervised image-to-image translation intends to learn a mapping of an image in a given domain to an analogous image in a different domain, without explicit supervision of the mapping. Few-shot unsupervised image-to-image translation further attempts to generalize the model to an unseen domain by leveraging example images of the unseen domain provided at inference time. While remarkably successful, existing few-shot image-to-image translation models find it difficult to preserve the structure of the input image while emulating the appearance of the unseen domain, which we refer to as the content loss problem. This is particularly severe when the poses of the objects in the input and example images are very different. To address the issue, we propose a new few-shot image translation model, which computes the style embedding of the example images conditioned on the input image and a new architecture design called the universal style bias. Through extensive experimental validations with comparison to the state-of-the-art, our model shows effectiveness in addressing the content loss problem.

1. Introduction

Image-to-Image translation [1, 2] concerns learning a mapping that can translate an input image in one domain into an analogous image in a different domain. Unsupervised image-to-image translation [3, 4, 5, 6, 7, 8, 9] attempts to learn such a mapping without paired data. The state-of-the-art has advanced significantly in the past few years. However, while existing unsupervised image-to-image translation models can generate realistic translations, they still have several drawbacks. First, they require a large amount of images from the source and target domains for training. Second, they cannot be used to generate images in unseen domains. These limitations are addressed in the

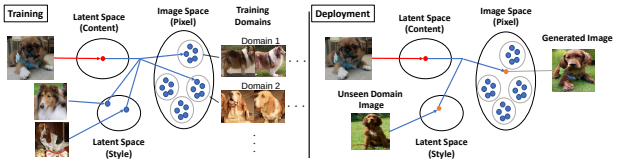

Figure 2: Few-shot image-to-image translation. **Training.** The training set consists of many domains. We train a model to translate images between these domains. **Deployment.** We apply the trained model to perform few-shot image translation. Given a few examples from a test domain, we aim to translate a content image into an image analogous to the test class.

few-shot *unsupervised* image-to-image translation framework [10]. By leveraging example-guided episodic training, the few-shot image translation framework [10] learns to extract the domain-specific style information from a few example images in the unseen domain during test time, mixes it with the domain-invariant content information extracted from the input image, and generates a few-shot translation output as illustrated in Fig. 2.

However, despite showing encouraging results on relatively simple tasks such as animal face and flower translation, the few-shot translation framework [10] frequently generates unsatisfactory translation outputs when the model is applied to objects with diverse appearance, such as animals with very different body poses. Often, the translation output is not well-aligned with the input image. The domain invariant content that is supposed to remain unchanged disappears after translation, as shown in Fig. 3. We will call this issue the *content loss* problem. We hypothesize that solving the content loss problem would produce more faithful and photorealistic few-shot image translation results.

In this paper, we propose a novel network architecture to counter the content loss problem. We design a style encoder called the *content-conditioned style encoder*, to hinder the transmission of task-irrelevant appearance information to the image translation process. In contrast to the existing style encoders, our style code is computed by conditioning on the input content image. We use a new architecture design to limit the variance of the style code. We conduct an extensive experimental validation with a comparison to the state-of-the-art method using several newly collected and challenging few-shot image translation datasets. Experimental results verify the effectiveness of the proposed method in dealing with the content loss problem. Our model can generate more faithful and more photorealistic few-shot translation outputs as shown in Fig. 1 and 3. The datasets, source code, and trained models will be made available upon publication. Thee contributions of the work are summarized below.

1. We identify the *content loss* problem in the existing few-shot unsupervised image-to-image translation framework.
2. To address the *content loss* problem, we propose a novel network architecture and show it can generate more faithful and more photorealistic few-shot translation outputs as shown in Fig. 1 and 3.
3. Extensive experimental results on diverse datasets with comparisons to the state-of-the-art method verify the effectiveness of the proposed framework.

Figure 3: Illustration of the *content loss* problem. The images generated by the baseline [10] fail to preserve domain invariant appearance information in the content image. The animals' bodies are sometimes merged with the background (column 3, & 4), scales of the generated body parts are sometimes inconsistent with the input (column 5), and some body parts absent in the content image show up (column 1 & 2). Our proposed method solves this "content loss" problem.

2. Method

In this section, we start with a brief explanation of the problem setup, introduce the basic architecture, and then describe our proposed architecture. Since the training objectives of our method is similar to the FUNIT method [10], we do not explain them in detail. Throughout the paper, the two words, "class" and "domain", are used interchangeably since we treat each object class as a domain.

Problem setting. Fig. 2 provides an overview of the few-shot image translation problem [10]. Let X be a training set consists of images from K different domains. For each image in X, the class label is known. Note that we operate in the unsupervised setting where corresponding images between domains are *unavailable*. The few-shot image-to-image translation model learns to map a "content" image in one domain to an analogous image in the domain of the input "style" examples. In the test phase, the model sees a few example images from an unseen domain and performs the translation.

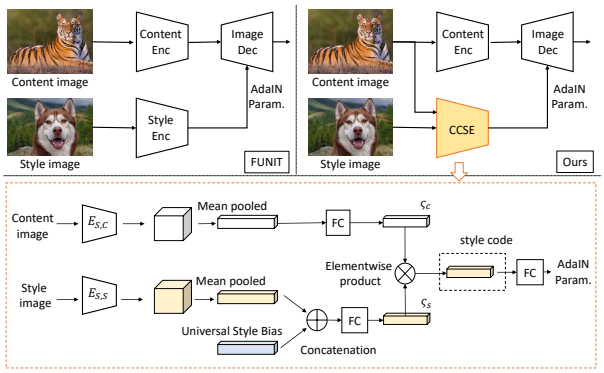

Figure 4: **Top**. The baseline FUNIT model [10] vs. the proposed model. To highlight, we use a novel style encoder called the content-conditioned style encoder where the content image is also used in computing the style code for few-shot unsupervised image-to-image translation. **Bottom**. Detail design of the content-conditioned style encoder. Please refer to the main text for more details.

During training, a pair of content and style images x_c, x_k is randomly sampled. Let x_k denote a style image in domain k. The content image x_c can be from any domains in K. The generator G translates x_c into an image of class k ($\bar{\boldsymbol{x}}_k$) while preserving the content information of x_c.

$$\bar{\boldsymbol{x}}_k = G(x_c, x_k) \tag{1}$$

In the test phase, the generator takes style images from a domain unseen during training, which we call the target domain. The target domain can be any related domain, not included in K.

The content loss problem. As discussed in the introduction as well as illustrated in Fig. 3, the FUNIT method suffers from the content loss problem—the translation result is not well-aligned with the input image.

In the following, we will present our content-conditioned style encoder, which, empirically, produces more stable style encoding against the variations in the style image.

Content-conditioned style encoder. We hypothesize that the content loss problem can be mitigated if the style embedding is more robust to small variations in the style image. To this end, we design a new style encoder architecture, called the content-conditioned style encoder (COCO). There are several distinctive features in COCO. The most obvious one is the conditioning in the content image as illustrated in the top-right of Fig. 4. Unlike the style encoder in FUNIT, COCO takes *both* content and style image as input. With this content-conditioning scheme, we create a *direct* feedback path during learning to let the content image influence how the style code is computed. It also helps reduce the direct influence of the style image to the extract style code.

In addition to the COCO, we also improve the design of the content encoder, image decoder, and discriminator in the FUNIT work [10]. For the content encoder and image decoder, we find that replacing the vanilla convolutional layers in the original design with residual blocks [11] improves the performance so does replacing the multi-task adversarial discriminator with the project-based discriminator [12].

3. Experiments

We evaluate our method on challenging datasets that contain large pose variations, part variations, and category variations. Unlike existing few-shot image-to-image translation works, which focus on translations between reasonably-aligned images or simple objects, our interest is in the translations between likely misaligned images of highly articulate objects. Throughout the experiments, we use 256×256 as our default image resolution for both inputs and outputs. Due to a limit of space, we focus on a qualitative evaluation of generated images.

Datasets. We benchmark our method using 4 datasets, *Carnivores, Mammals, Birds, and Motorbikes*. Each of the dataset contains objects with diverse poses, parts, and appearances.

Baseline. We compare our method with the FUNIT method because it outperforms many baselines with a large margin as described in Liu *et al*. [10].

Results. Fig. 5 and 3 compare the one-shot translation results computed by the FUNIT method and our approach. We find images generated by the FUNIT method contain many artifacts while our method can generate photorealistic and faithful translation outputs. In Fig. 6, we further visualize two-shot translation results.

4. Conclusion

We introduced the COCO architecture, a new style encoder for few-shot image-to-image translation that extracts the style code from the example images from the unseen domain conditioning on the input content image and uses a universal style bias design. We showed that the COCO can effectively address the content loss problem, proven challenging for few-shot image-to-image-translation.

References

[1] Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017) 1

[2] Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional GANs. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2018) 1

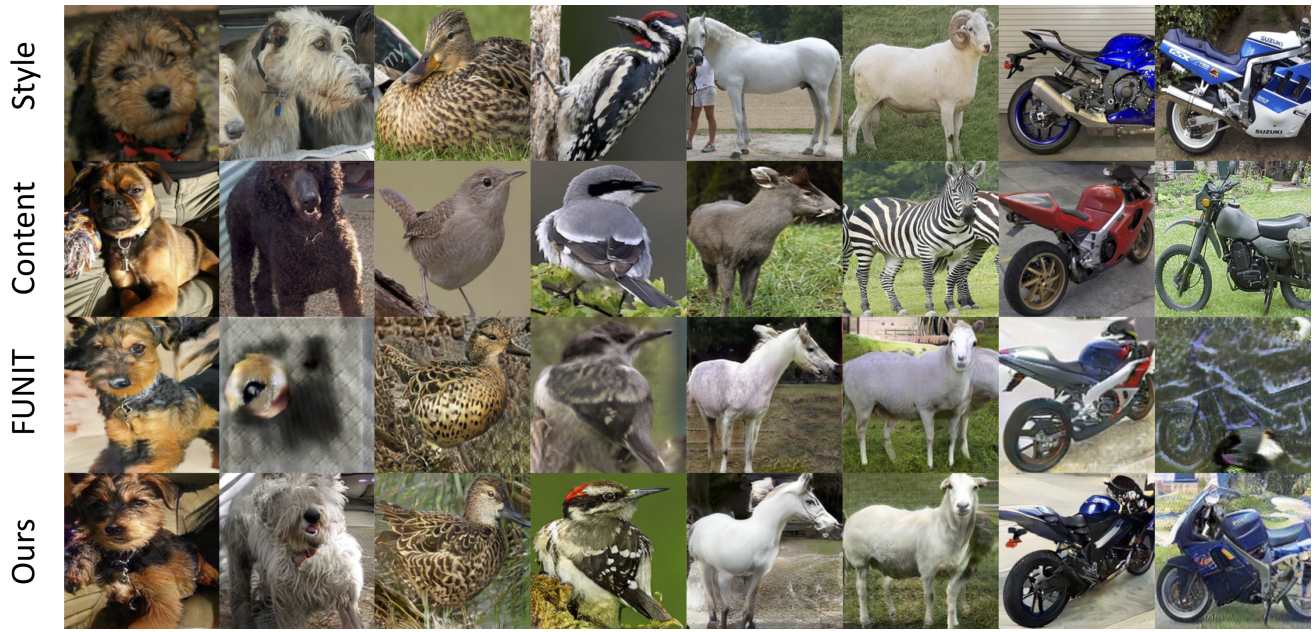

Figure 5: Results on one-shot image-to-image translation. Column 1 & 2 are from the Carnivore dataset. Column 3 & 4 are from the Bird dataset. Column 5 & 6 are from the Mammal dataset. Column 7 & 8 are from the Motorbike dataset.

Figure 6: Results on one-shot image-to-image translation. Column 1 & 2 are from the Carnivore dataset. Column 3 & 4 are from the Bird dataset. Column 5 & 6 are from the Mammal dataset.

[3] Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: IEEE International Conference on Computer Vision (ICCV). (2017) 1

[4] Liu, M.Y., Breuel, T., Kautz, J.: Unsupervised image-to-image translation networks. In: Advances in Neural Information Processing Systems (NeurIPS). (2017) 1

[5] Liang, X., Lee, L., Dai, W., Xing, E.P.: Dual motion GAN for future-flow embedded video prediction. In: Advances in Neural Information Processing Systems (NeurIPS). (2017) 1

[6] Kim, T., Cha, M., Kim, H., Lee, J.K., Kim, J.: Learning to discover cross-domain relations with generative adversarial networks. In: International Conference on Machine Learning (ICML). (2017) 1

[7] Liu, M.Y., Tuzel, O.: Coupled generative adversarial networks. In: Advances in Neural Information Processing Systems (NeurIPS). (2016) 1

[8] Taigman, Y., Polyak, A., Wolf, L.: Unsupervised cross-domain image generation. In: International Conference on Learning Representations (ICLR). (2017) 1

[9] Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2018) 1

[10] Liu, M.Y., Huang, X., Mallya, A., Karras, T., Aila, T., Lehtinen, J., Kautz, J.: Few-shot unsupervised image-to-image translation. In: IEEE International Conference on Computer Vision (ICCV). (2019) 2, 3

[11] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016) 3

[12] Miyato, T., Koyama, M.: cGANs with projection discriminator. In: International Conference on Learning Representations (ICLR). (2018) 3