Image-to-Image Translation

MUNIT

ECCV

2018

Multimodal Unsupervised Image-to-Image Translation

Xun Huang[1], Ming-Yu Liu[2], Serge Belongie[1], Jan Kautz[2]

Cornell University[1] NVIDIA[2]

Abstract. Unsupervised image-to-image translation is an important and challenging problem in computer vision. Given an image in the source domain, the goal is to learn the conditional distribution of corresponding images in the target domain, without seeing any examples of corresponding image pairs. While this conditional distribution is inherently multimodal, existing approaches make an overly simplified assumption, modeling it as a deterministic one-to-one mapping. As a result, they fail to generate diverse outputs from a given source domain image. To address this limitation, we propose a Multimodal Unsupervised Image-to-image Translation (MUNIT) framework. We assume that the image representation can be decomposed into a content code that is domain-invariant, and a style code that captures domain-specific properties. To translate an image to another domain, we recombine its content code with a random style code sampled from the style space of the target domain. We analyze the proposed framework and establish several theoretical results. Extensive experiments with comparisons to state-of-the-art approaches further demonstrate the advantage of the proposed framework. Moreover, our framework allows users to control the style of translation outputs by providing an example style image. Code and pretrained models are available at https://github.com/nvlabs/MUNIT.

Keywords: GANs, image-to-image translation, style transfer

1 Introduction

Many problems in computer vision aim at translating images from one domain to another, including super-resolution [1], colorization [2], inpainting [3], attribute transfer [4], and style transfer [5]. This cross-domain image-to-image translation setting has therefore received significant attention [6–25]. When the dataset

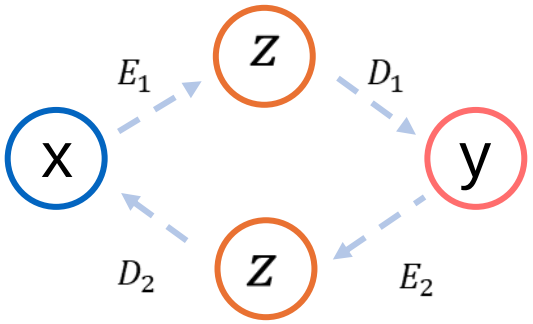
Huang, X., Liu, M. Y., Belongie, S., & Kautz, J. (2018). Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 172-189).

Image-to-Image Translation

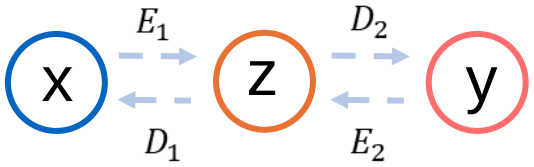
MUNIT

Comparison of Approaches

CycleGAN

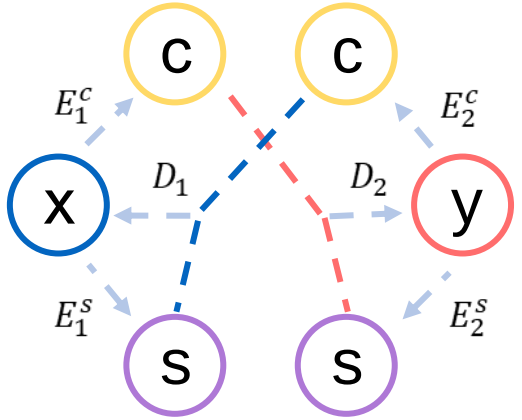


UNIT



Shared Latent Space

MUNIT



Content & Style

Image-to-Image Translation

MUNIT

Background & Goal

Previous research limitation

- Existing models relied on one-to-one mapping, producing only deterministic output.

Goal

- To generate diverse output images from a single input image.

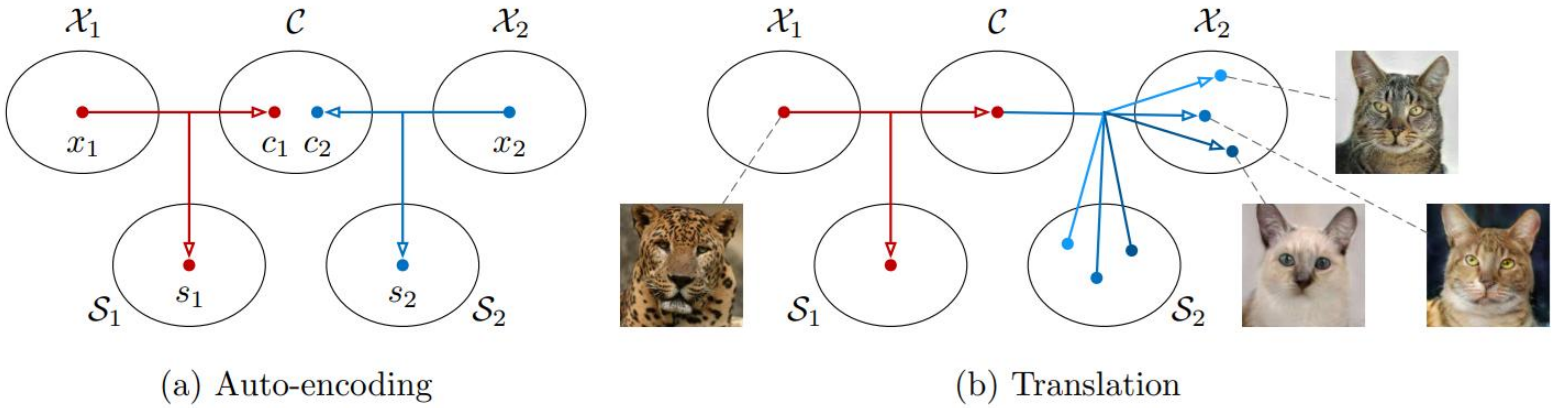


Fig 1. Images are encoded to a shared content space and a domain-specific style space

Huang, X., Liu, M. Y., Belongie, S., & Kautz, J. (2018). Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 172-189).

MUNIT

Background & Goal

Extension of UNIT

- Deterministic mapping : The same content input always produces the exact same output.

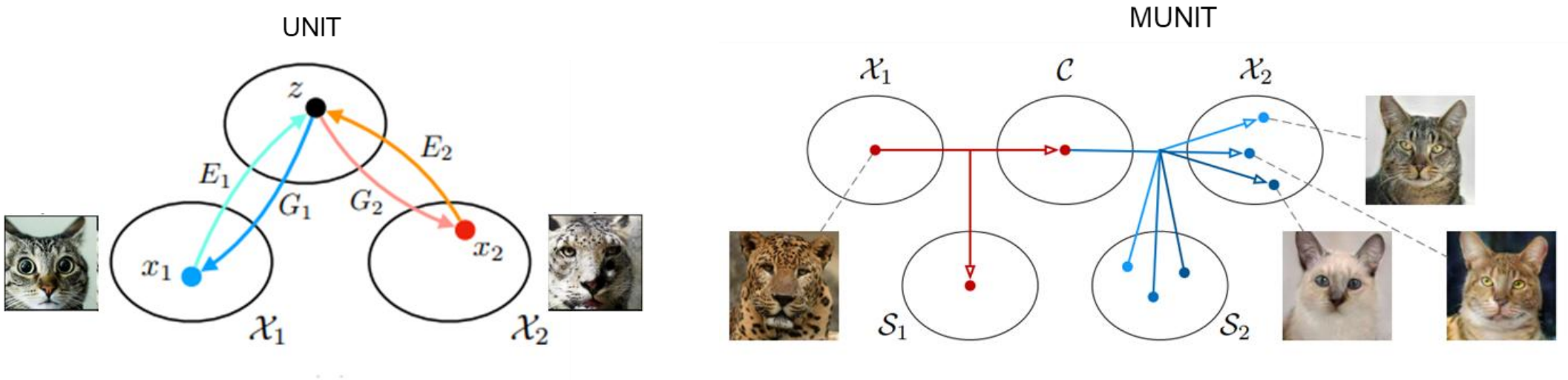


Fig 2. Differences in shared latent space assumption

Huang, X., Liu, M. Y., Belongie, S., & Kautz, J. (2018). Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 172-189).

Image-to-Image Translation

MUNIT

Background & Goal

Limitation of UNIT

- Fully shared latent space : The same content input always produces the exact same output.
  → Partially shared latent space

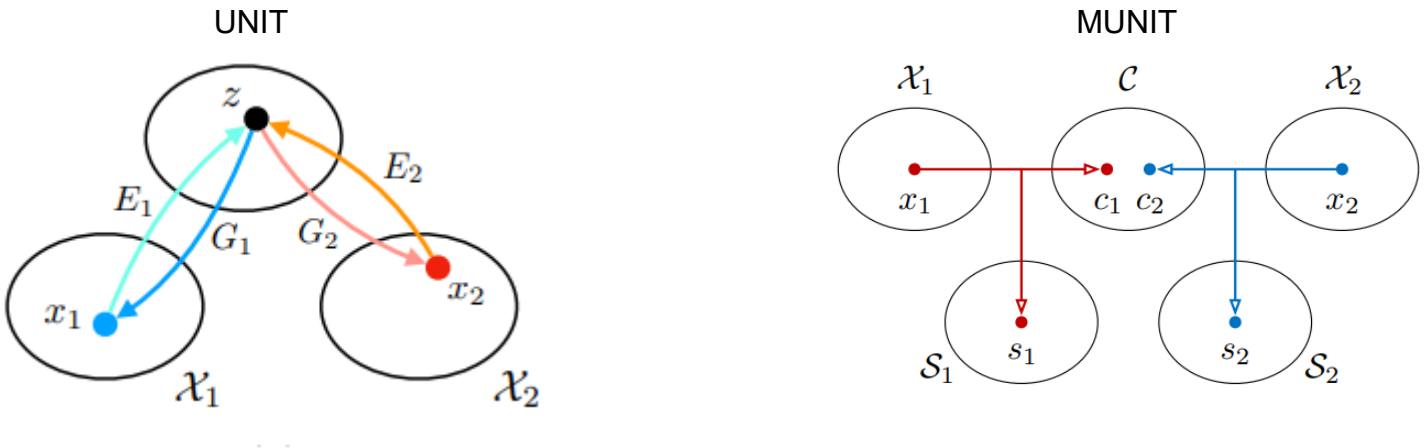


Fig 2. Differences in shared latent space assumption

Huang, X., Liu, M. Y., Belongie, S., & Kautz, J. (2018). Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 172-189).

Image-to-Image Translation

MUNIT

Network Architecture

Auto-encoder framework

- Separates an image into content and style and then recombines them.

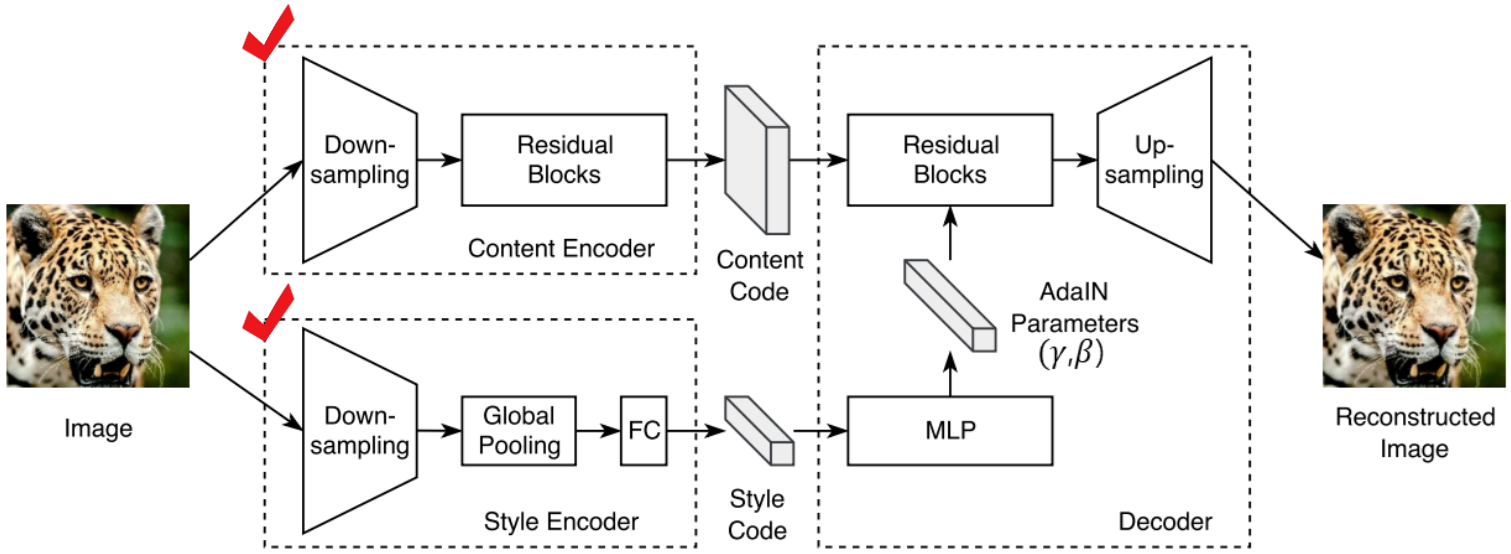


Fig 3. Overview of MUNIT

$$\mathrm{AdaIN}(z,\gamma,\beta)=\gamma\left(\frac{z-\mu(z)}{\sigma(z)}\right)+\beta$$

Eq 1. Adaptive instance normalization
(z: feature map, $\gamma\&\beta$: scaling and shifting scalars)

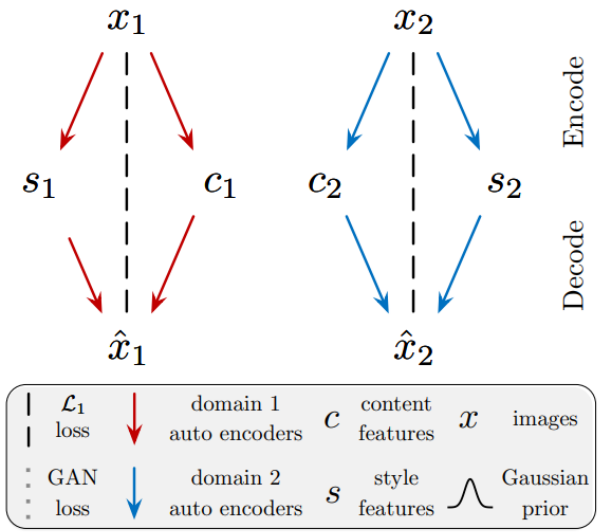
Huang, X., Liu, M. Y., Belongie, S., & Kautz, J. (2018). Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 172-189).

MUNIT

Loss Functions

Image reconstruction loss

- Image reconstruction : Image → Latent → Image



(a) Within-domain reconstruction

Fig 4. Overview of image reconstruction loss

Huang, X., Liu, M. Y., Belongie, S., & Kautz, J. (2018). Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 172-189).

Image-to-Image Translation

MUNIT

Loss Functions

Image reconstruction loss

- Image cycle consistency enforces similarity to the original, stabilizing content and style extraction and ensuring decoder synthesis.

$$\mathcal{L}_{\text{recon}}^{x_1} = \mathbb{E}_{x_1 \sim p(x_1)}[||G_1(E_1^c(x_1), E_1^s(x_1)) - x_1||_1]$$

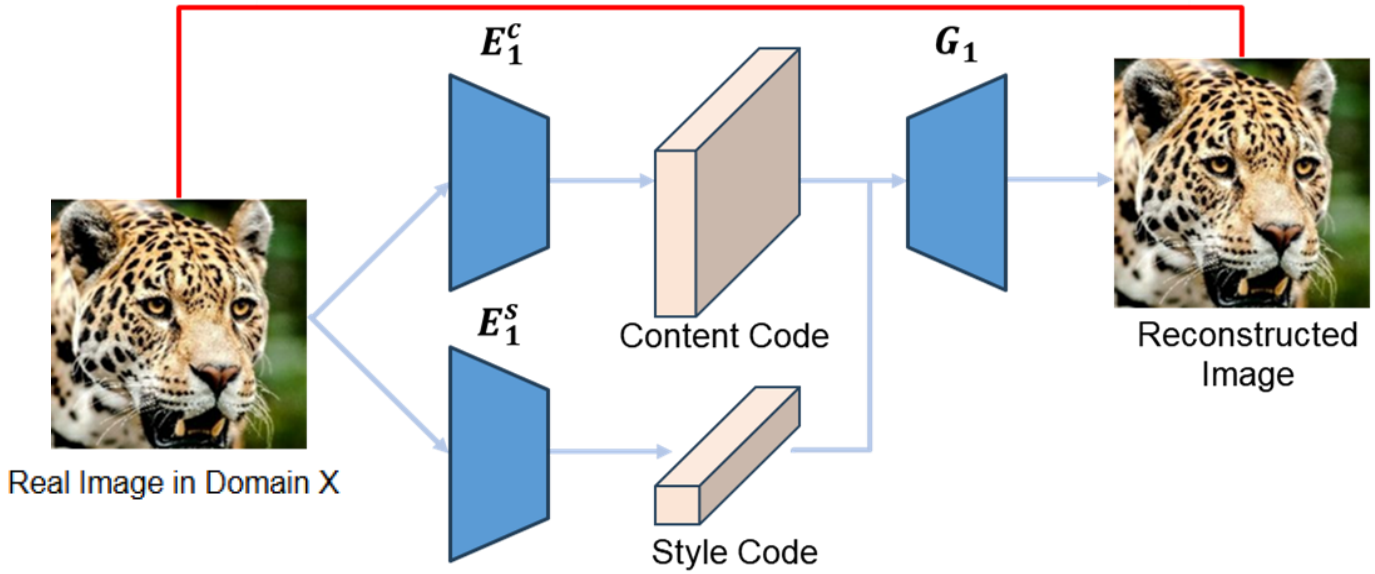


Fig 5. Image reconstruction loss

Huang, X., Liu, M. Y., Belongie, S., & Kautz, J. (2018). Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 172-189).

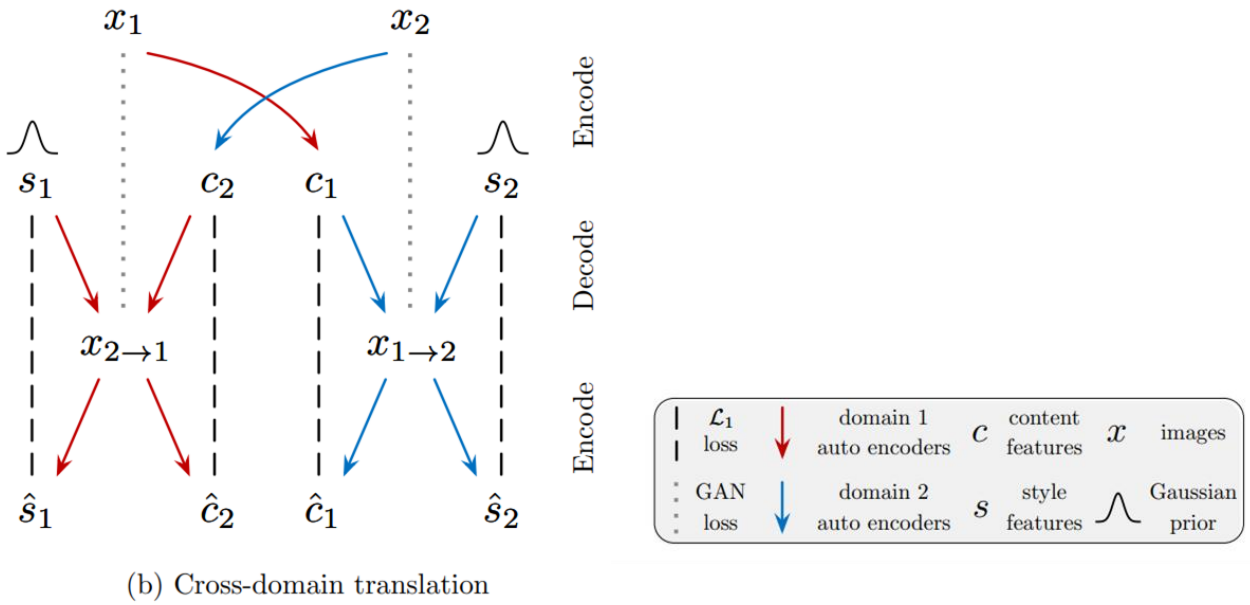
Image-to-Image Translation

MUNIT

Loss Functions

Latent reconstruction loss

- Latent reconstruction : Latent → Image → Latent



(b) Cross-domain translation

Fig 6. Overview of latent reconstruction loss

Huang, X., Liu, M. Y., Belongie, S., & Kautz, J. (2018). Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 172-189).

Image-to-Image Translation

MUNIT

Loss Functions

Latent reconstruction loss

$$\mathcal{L}_{\text{recon}}^{c_1} = \mathbb{E}_{c_1 \sim p(c_1), s_2 \sim q(s_2)}[||E_2^c(G_2(c_1, s_2)) - c_1||_1]$$
$$\mathcal{L}_{\text{recon}}^{s_2} = \mathbb{E}_{c_1 \sim p(c_1), s_2 \sim q(s_2)}[||E_2^s(G_2(c_1, s_2)) - s_2||_1]$$

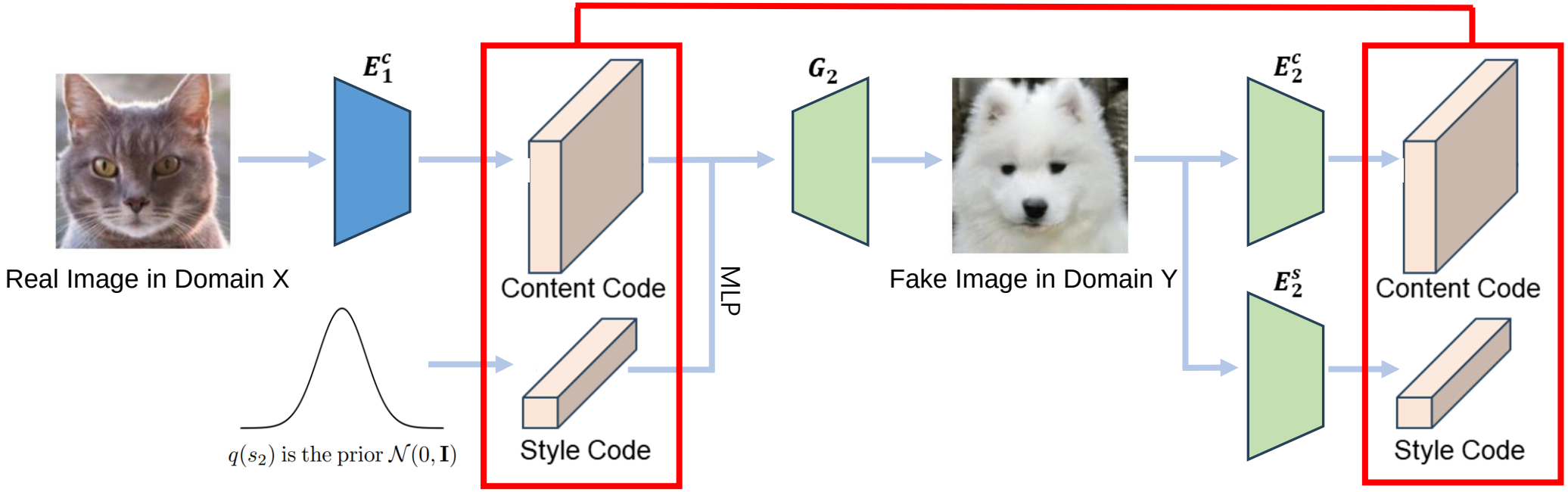


Fig 8. Latent reconstruction loss

Image-to-Image Translation

MUNIT

Loss Functions

Gaussian prior

- To enable the generation of diverse styles
- Stability to maintain a well-formed style code space

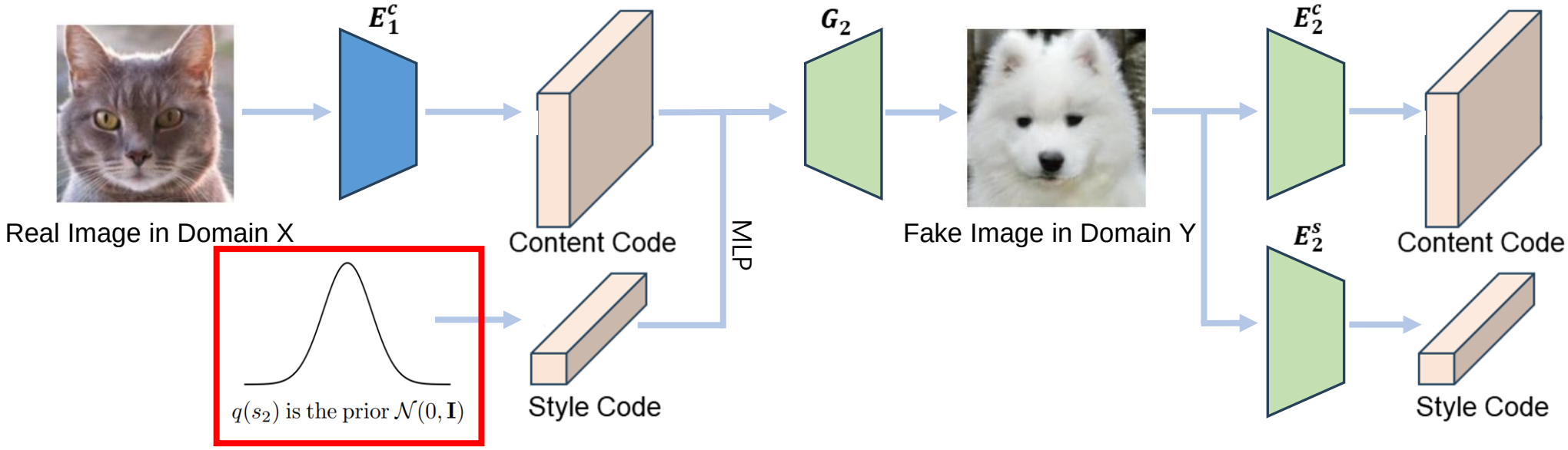


Fig 8. Latent reconstruction loss

Image-to-Image Translation

MUNIT

Loss Functions

Gaussian prior

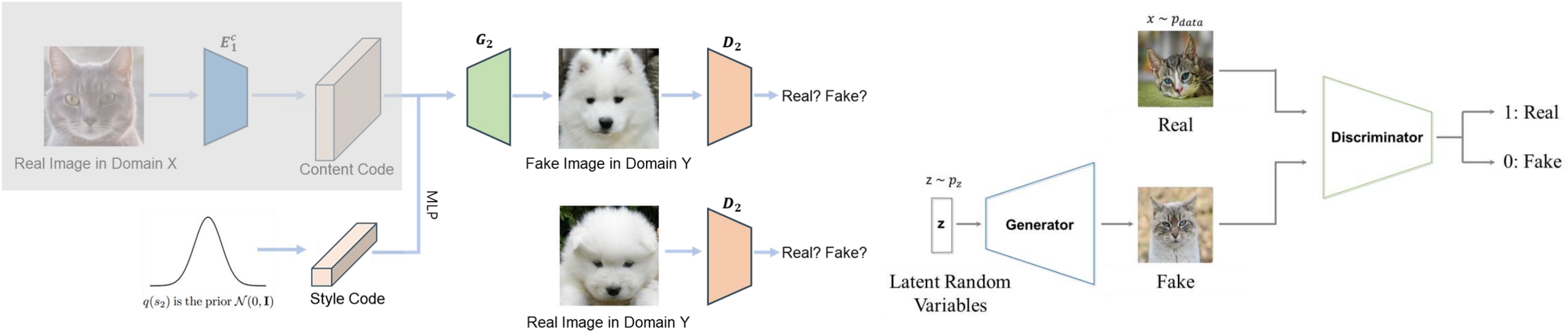


Fig 8. Latent reconstruction loss

MUNIT

Loss Functions

Relatively low-dimensional style code

- It captures simple attributes like brightness, and colors, rather than spatial information.
- It is transformed through a nonlinear MLP into a more complex representation.
  => However, 8 channels are insufficient for complex tasks.

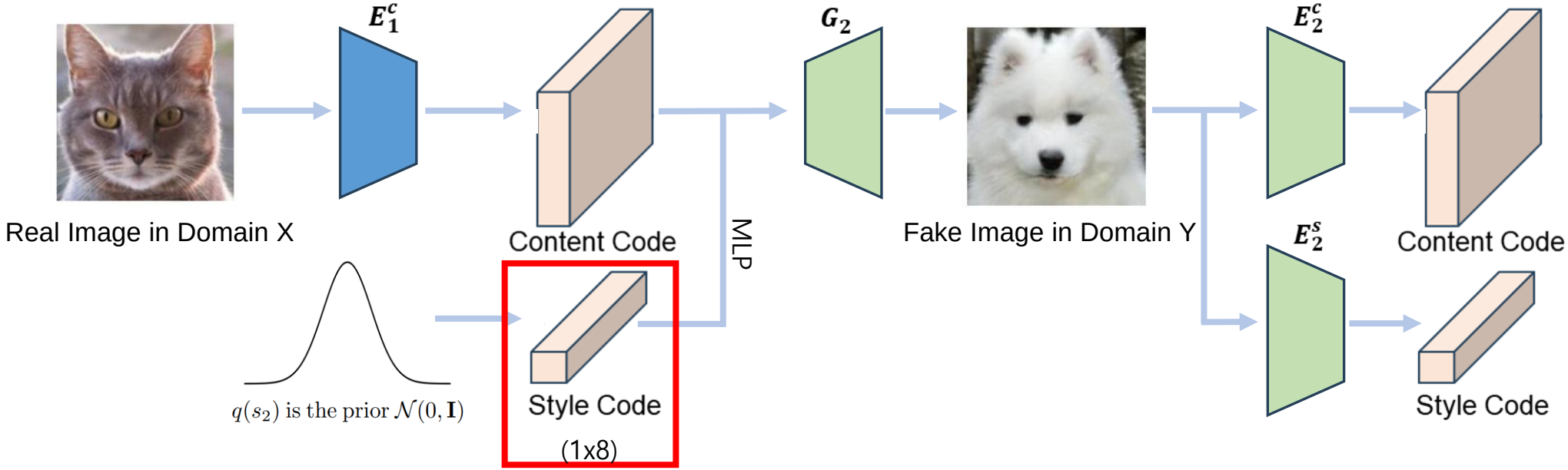


Fig 8. Latent reconstruction loss

Image-to-Image Translation

MUNIT

Loss Functions

Content space cycle consistency

- Latent space cycle consistency allows geometric transformations while preserving core content.

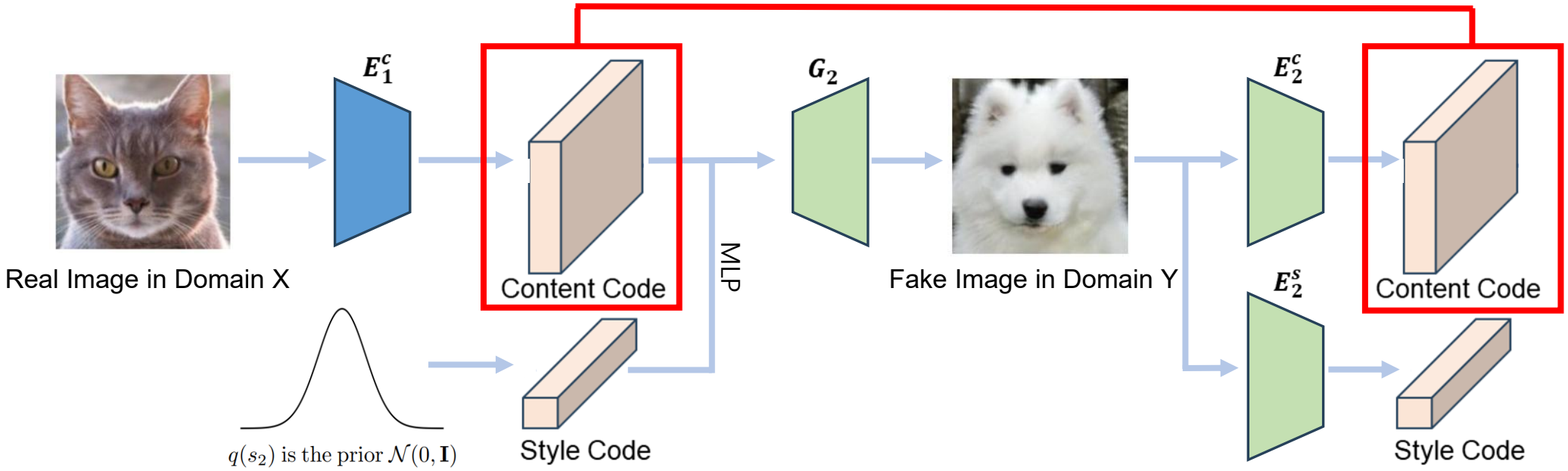


Fig 8. Latent reconstruction loss

Image-to-Image Translation

MUNIT

Loss Functions

Shared latent space without weight sharing

- The encoder disentangles content and style.
- Content-consistency loss preserves content even after style is removed.
  → Training to align content spaces across domains

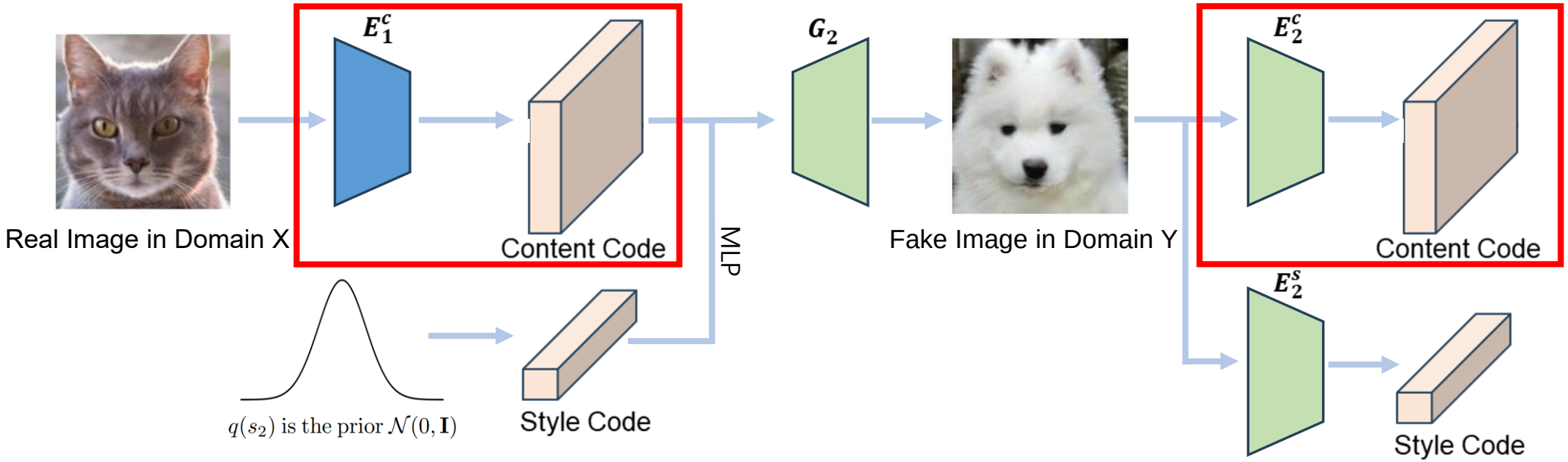


Fig 8. Latent reconstruction loss

Image-to-Image Translation

MUNIT

Loss Functions

GAN loss

$$\mathcal{L}_{\text{GAN}}^{x_2} = \mathbb{E}_{c_1 \sim p(c_1), s_2 \sim q(s_2)}[\log(1 - D_2(G_2(c_1, s_2)))] + \mathbb{E}_{x_2 \sim p(x_2)}[\log D_2(x_2)]$$

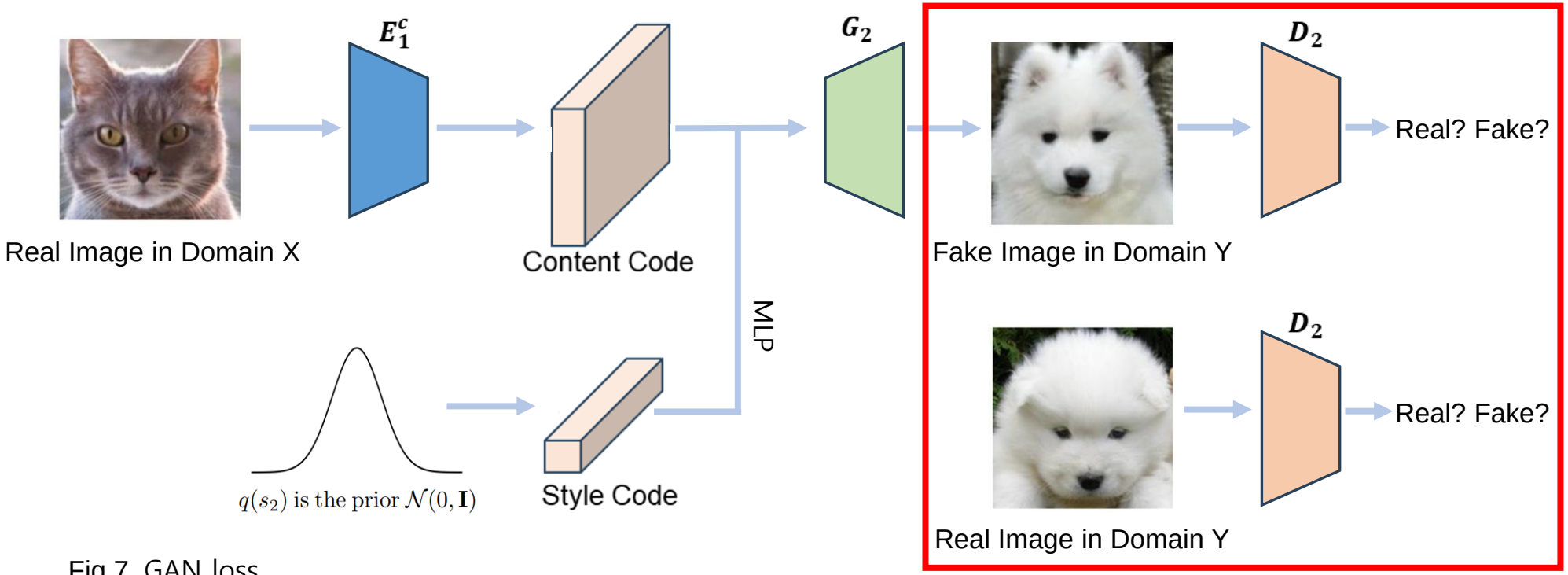


Fig 7. GAN loss

MUNIT

Inference Mode

- Applying a style sampled from a prior distribution.

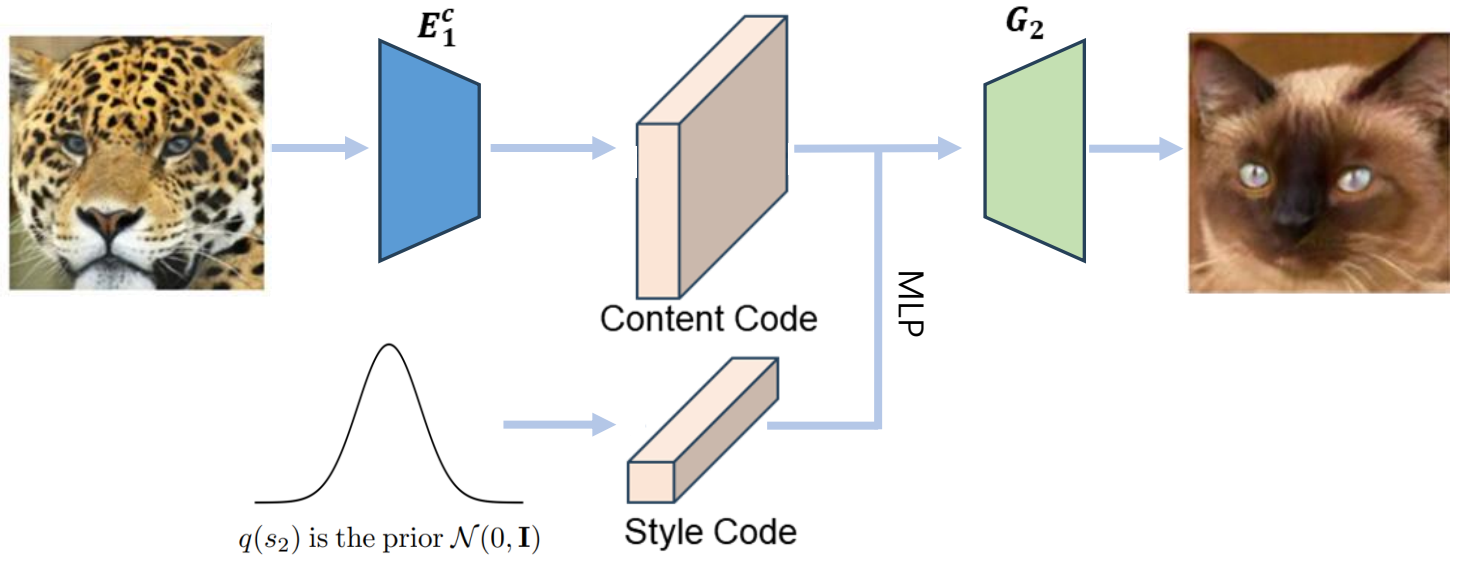


Fig 9. Applying a random style.

MUNIT

Inference Mode

- Applying a style extracted from a reference image.

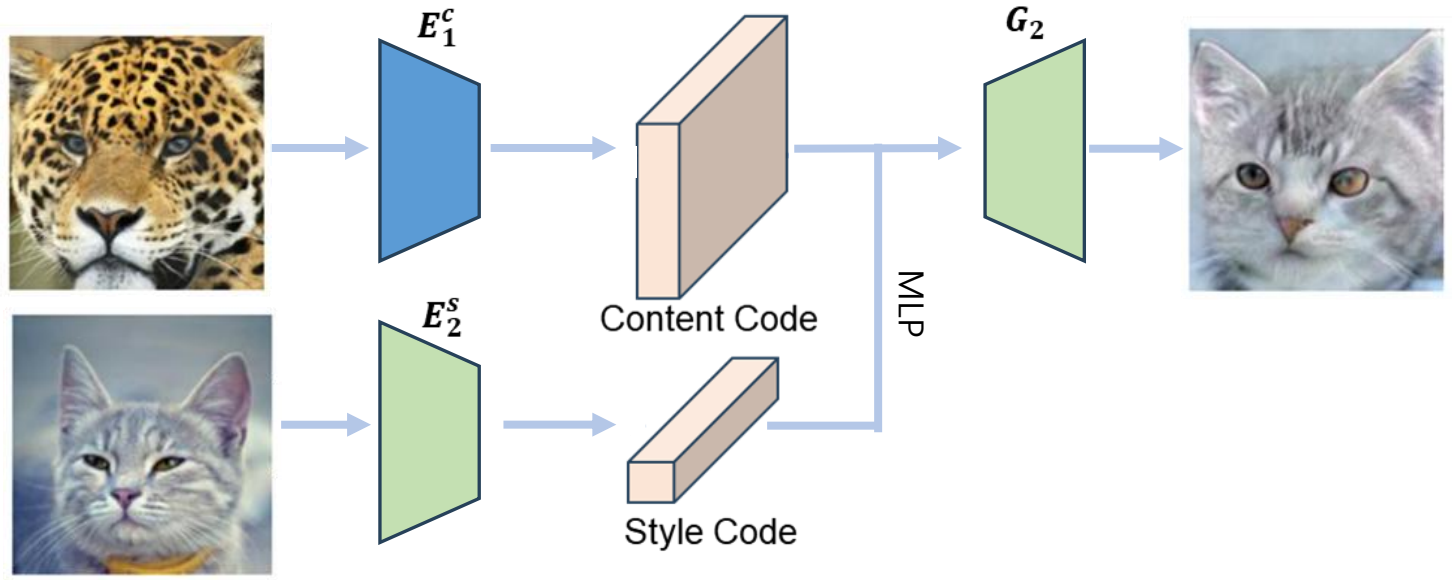


Fig 10. Applying a reference style.

Image-to-Image Translation

MUNIT

Experiment Results

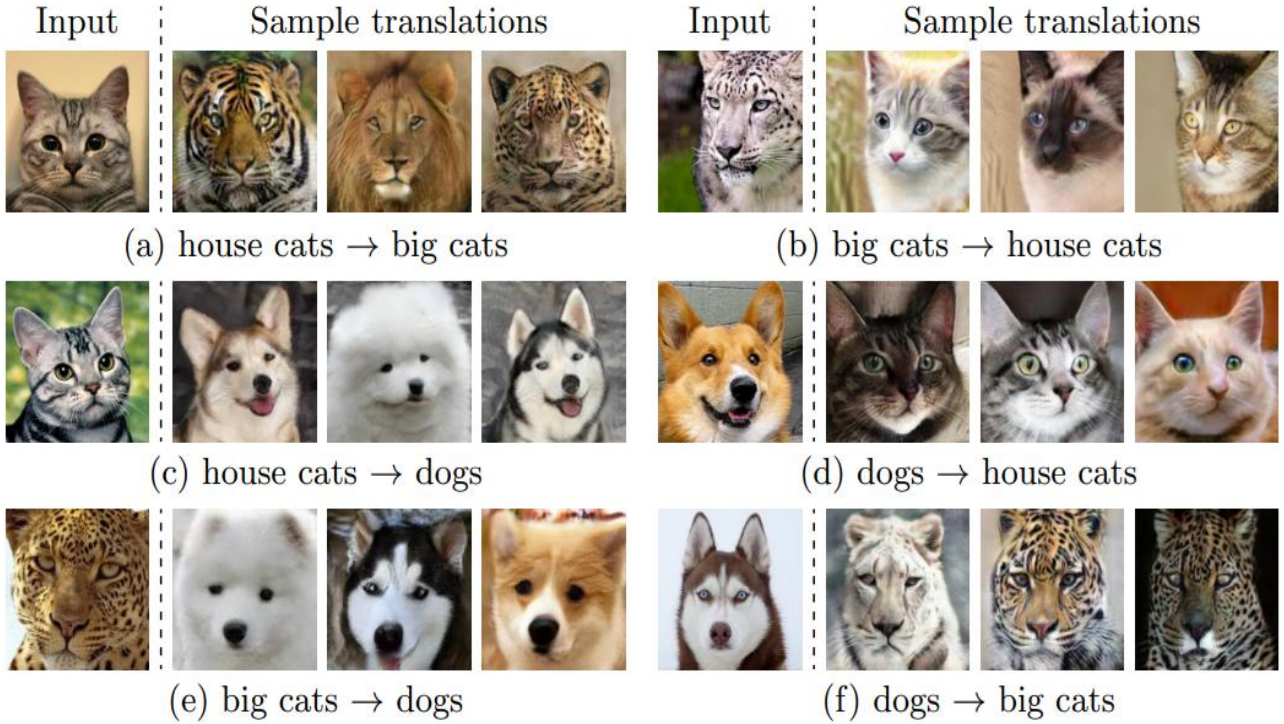
Fig 11. 3 random outputs from the methods.

Huang, X., Liu, M. Y., Belongie, S., & Kautz, J. (2018). Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 172-189).

Image-to-Image Translation

MUNIT

Experiment Results



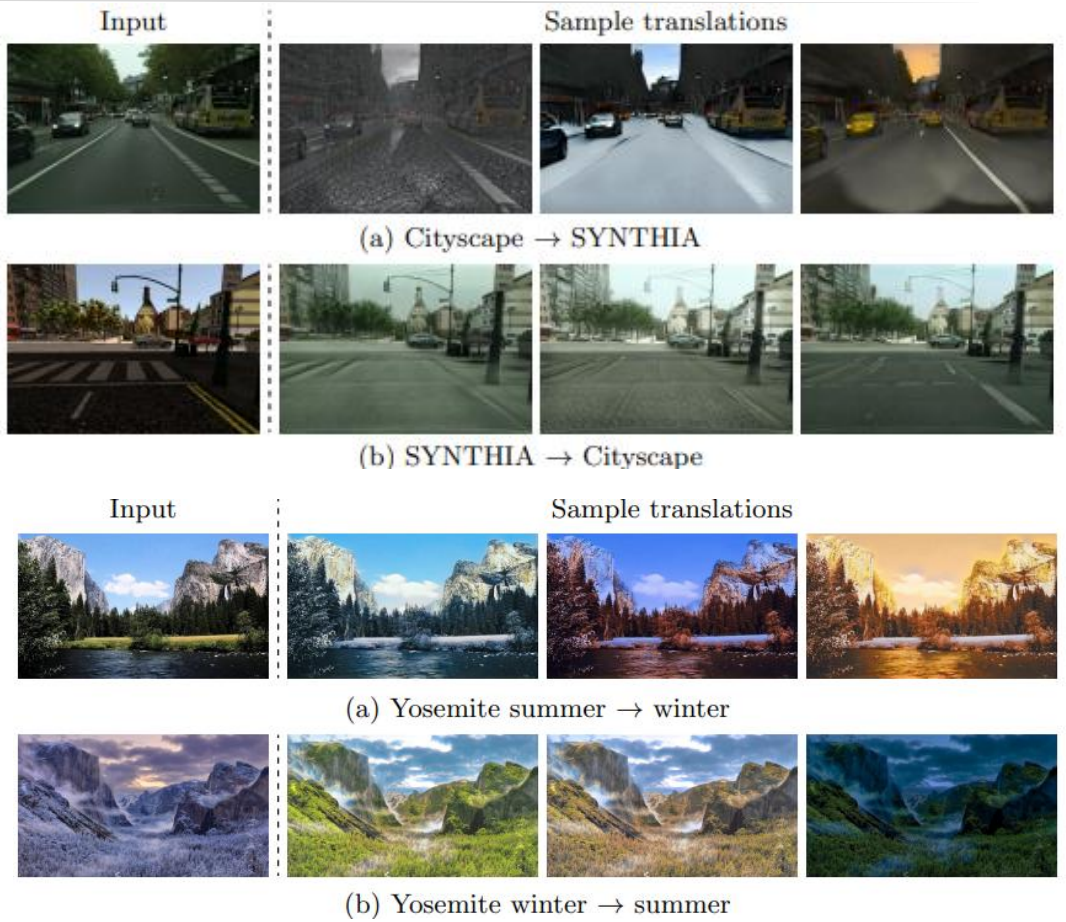


Fig 12. Example results

Huang, X., Liu, M. Y., Belongie, S., & Kautz, J. (2018). Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 172-189).

Image-to-Image Translation

MUNIT

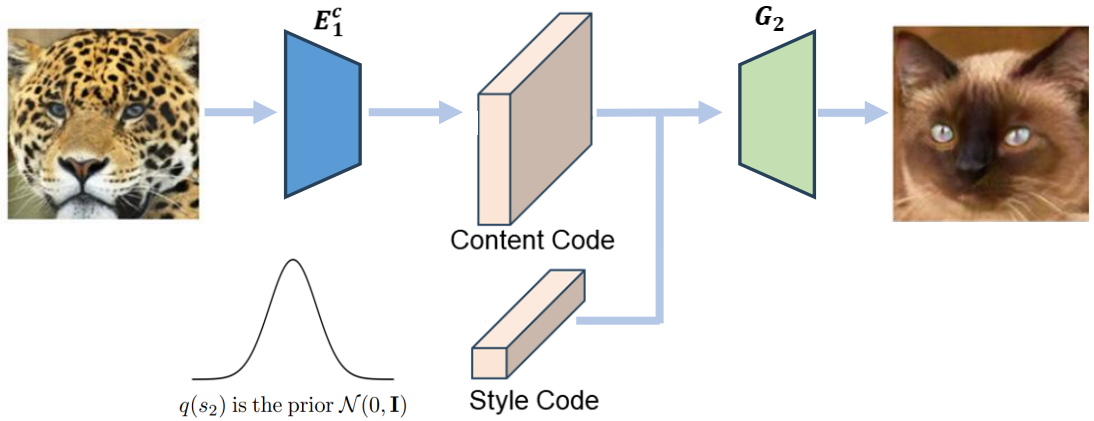
w, w/o Reference Style Image

Random sampling

- Diverse style generation
- Risk of unrealistic styles
- Unstable
- Difficult to control

Reference Style Extraction

- Ensures realism
- Easy to control
- Limited style diversity
- Requires reference image



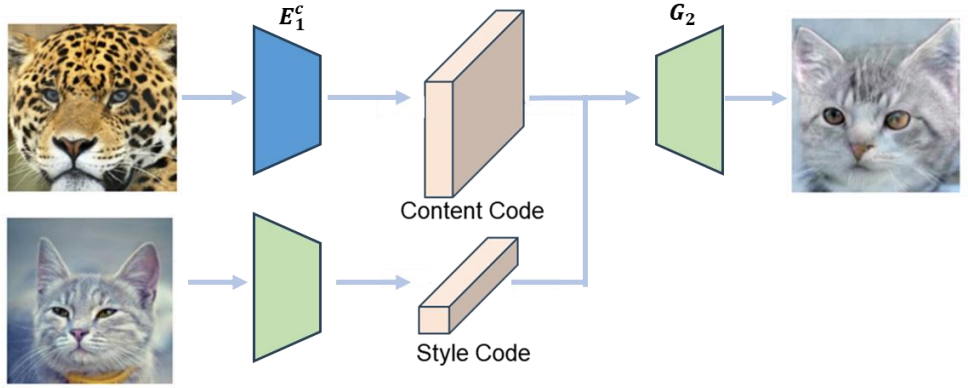


Fig 13. Comparison of methods for generating style codes.

Image-to-Image Translation

MUNIT

Implications & Limitations

Implications

- Disentangling images into shared content and domain-specific style enables flexible and diverse translations.

Limitations

- Difficulty in separating sophisticated content and style.

  => Content preservation is not perfect, causing distortions.

- Random sampling can lead to unrealistic styles, lowering realism.