VARIATIONAL BAYESIAN FRAMEWORK FOR ADVANCED IMAGE GENERATION WITH DOMAIN-RELATED VARIABLES

Yuxiao Li†, Santiago Mazuelas‡, and Yuan Shen†

†Department of Electronic Engineering, Tsinghua University, Beijing, China
‡BCAM-Basque Center for Applied Mathematics, Bilbao, Spain
Emails: li-yx18@mails.tsinghua.edu.cn, smazuelas@bcamath.org, shenyuan_ee@tsinghua.edu.cn

ABSTRACT

Deep generative models (DGMs) and their conditional counterparts provide a powerful ability for general-purpose generative modeling of data distributions. However, it remains challenging for existing methods to address advanced conditional generative problems without annotations, which can enable multiple applications like image-to-image translation and image editing. We present a unified Bayesian framework for such problems, which introduces an inference stage on latent variables within the learning process. In particular, we propose a variational Bayesian image translation network (VBITN) that enables multiple image translation and editing tasks. Comprehensive experiments show the effectiveness of our method on unsupervised image-to-image translation, and demonstrate the novel advanced capabilities for semantic editing and mixed domain translation.

***Index Terms*—** DGMs, conditional generative problems, Bayesian framework, variational inference.

1. INTRODUCTION

Deep generative models (DGMs) [1, 2] are popular ways to learn complicated data distributions in an unsupervised manner. However, they have less promising capabilities towards the generation of conditional distributions, and are hard to scale to different problems in a consistent scheme. The related techniques include image-to-image translation and image editing, enabling a wide range of applications such as super resolution [3], image colorization [4], image inpainting [5, 6], and semantic attribute synthesis [7].

Different strategies have been proposed to improve the scalability of DGMs towards conditional distributions. Early conditional generative methods, like Conditional GAN [8] and Info GAN [9], use supervised annotations from the target distribution. While successful in basic conditional generative problems, these methods are insufficient towards advanced problems without direct annotations, such as unsupervised image-to-image generation. Existing techniques tackle this problem by adding constraints in either the image space or a low-dimensional latent space [10–12]. However, a unified framework for the underlying generative process of different semantic variables is seldom claimed, resulting in redundant fine-tunning work and limited scalability towards advanced tasks in a consistent scheme.

From a statistical viewpoint, these problems can be described well by a latent variable model (LVM). Specifically, semantic features can be viewed as latent variables while the generation can be conducted by inferring the conditional distribution of images given the variables corresponding to desired semantics. The idea of disentangling codes for different semantics is partially discussed by [13, 14], while seldom derived from first principles via statistic modeling.

In this paper, we present a novel probabilistic framework for a general class of conditional image generative problems. We then propose a deep generative network for image translation tasks, where latent variables of semantics are inferred via a variational lower bound in learning. Driven by a rigorous probabilistic model, the proposed method has a clear interpretation and improved generality to encompass multiple variants. Experimental results show that the proposed method achieves comparable performance with classic frameworks on unsupervised image-to-image translation, and enables novel variants like mixed domain translation.

2. BAYESIAN FRAMEWORK FOR IMAGE GENERATION WITH LATENT VARIABLES

We present a Bayesian framework for conditional image generation with respect to two latent variables, representing domain-related and domain-unrelated semantics respectively.

2.1. Bayesian Model for Image Generation

Suppose the generative process of an image sample $\mathbf{x}^{(k)} \in \mathbb{X}$ in certain domain involves two latent variables: a domain-related variable $\mathbf{y}$ that describes features specific to the domain, and an independent domain-unrelated variable $\mathbf{z}$ that describes general features. We refer to the domain-related variable as 'style' and the domain-unrelated variable as 'content', following the classical nomenclature in [15].

2.2. Unsupervised Image-to-Image Translation

Consider a dataset $\mathbb{X}_S = \{\mathbf{x}_S^{(k)}\}_{k=1}^N$ consisting of N i.i.d. samples of a random variable $\mathbf{x}_S$ corresponding with domain S, and a dataset $\mathbb{X}_T = \{\mathbf{x}_T^{(l)}\}_{l=1}^M$ consisting of M i.i.d. samples of $\mathbf{x}_T$ corresponding with domain T. The content variables $\mathbf{z}_S$ and $\mathbf{z}_T$ corresponding with domains S and T share the same prior distribution $p(\mathbf{z})$, while the style variables $\mathbf{y}_S$ and $\mathbf{y}_T$ corresponding with these domains have different distributions, denoted as $p(\mathbf{y}_S)$ and $p(\mathbf{y}_T)$, respectively.

The translation process from an image $\mathbf{x}_S^{(k)}$ in domain S to its counterpart $\mathbf{x}_{S\to T}^{(k)}$ in domain T, consists of three sequential steps: *1)* A value $\mathrm{y}_T^{(k)}$ for style variable is generated from distribution $p(\mathbf{y}_T)$ corresponding with domain T; *2)* A value $\mathrm{z}_S^{(k)}$ for content variable is generated from the conditional distribution $p(\mathbf{z}|\mathbf{x}_S^{(k)})$; and *3)* A translated image $\mathrm{x}_{S\to T}^{(k)}$ is generated from the conditional distribution $p(\mathbf{x}_T|\mathrm{y}_T^{(k)}, \mathrm{z}_S^{(k)})$.

2.3. Multiple Variants

The proposed model also enables to develop variants, achieved by modifications in the three steps above. We introduce three such variants, which can be further combined and varied.

Multi-modal style editing. The information for style semantics in the first step are obtained by sampling the distribution of the style variable, resulting in a spectrum of values. The translated image with these values can result in the generation of images with multi-modal styles, i.e.,

$$\mathrm{y}_T^{(k_1)}, \ldots, \mathrm{y}_T^{(k_l)} \sim p(\mathbf{y}_T), \tag{1}$$

with the other steps stay unchanged, images of l different styles in domain T $\mathrm{x}_{S\to T}^{(k_1)}, \ldots, \mathrm{x}_{S\to T}^{(k_l)}$ can be generated.

Multi-modal content editing. The information for content semantics in the second step are obtained by sampling the distribution of the content variable, resulting in a spectrum of values. The generation with these values can result in images of multi-modal contents, i.e.,

$$\mathrm{z}_S^{(k_1)}, \ldots, \mathrm{z}_S^{(k_m)} \sim p(\mathbf{z}|\mathbf{x}_S^{(k)}), \tag{2}$$

with the other steps stay unchanged, images of m content variants $\mathrm{x}_{S\to T}^{(k_1)}, \ldots, \mathrm{x}_{S\to T}^{(k_m)}$ can be achieved.

Mixed domain translation. The semantics determined by the style variable can represent a mixed style from more than one target domains, resulting in translated image in a mixed domain. The distribution for the mixed style can be constructed as the weighted sum of style distributions, i.e.,

$$\mathrm{y}_{Mix}^{(k)} \sim p(\mathbf{y}_{Mix}) = \sum_{i=1}^{n} w_i p(\mathbf{y}_{T_n}),\ \sum_{i=1}^{n} w_i = 1, \tag{3}$$

where $w_i, i = 1, \ldots, n$ are the weight values for these styles, e.g. $w_i = 1/n, i = 1, \ldots, n$.

3. VARIATIONAL BAYESIAN IMAGE TRANSLATION NETWORK

In this section, the variational Bayesian (VB) method is introduced and implemented to conduct image translation tasks.

3.1. Variational Bayesian Method

According to the VB technique, we construct a distribution $q(\mathbf{y}, \mathbf{z}|\mathbf{x})$ to approximate the true posterior $p(\mathbf{y}, \mathbf{z}|\mathbf{x})$. Following the mean field approximation, we assume that

$$q(\mathbf{y}, \mathbf{z}|\mathbf{x}) = q(\mathbf{y}|\mathbf{x})q(\mathbf{z}|\mathbf{x}). \tag{4}$$

The following proposition shows the lower bound of the log-likelihood $\log p(\mathbf{x})$ of each sample in the LVM.

Proposition 1. Given the Bayesian model for image translation and the variational distribution q on latent variables, the evidence lower bound (ELBO) $\mathcal{L}$ of log-likelihood $\log p(\mathbf{x})$ is expressed as follows,

$$\begin{aligned}\mathcal{L}(p, q; \mathbf{x}) =& \mathbb{E}_{q(\mathbf{y},\mathbf{z}|\mathbf{x})}\big[\log p(\mathbf{x}|\mathbf{y}, \mathbf{z})\big] \\ &- \mathrm{KL}\left(q(\mathbf{y}|\mathbf{x})\big|\big|p(\mathbf{y})\right) - \mathrm{KL}\left(q(\mathbf{z}|\mathbf{x})\big|\big|p(\mathbf{z})\right)\end{aligned} \tag{5}$$

where KL is the Kullback-Leibler (KL) divergence.

Such bound can be used to find a suitable approximated distribution q^* that matches the true distribution p in general.

3.2. Neural Modules for Distributions

We construct neural modules to represent the unknown p and the variational distribution q. In particular, the likelihood distribution is assumed to come from a parametric family $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{y}, \mathbf{z})$ learned by a decoder network $g_{\boldsymbol{\theta}}$, while the variational posterior distribution is from $q_{\boldsymbol{\phi}}(\mathbf{y}, \mathbf{z}|\mathbf{x})$ learned by an encoder network $f_{\boldsymbol{\phi}}$. Combing the neural modules, we can construct a modified VAE to learn variational parameter $\boldsymbol{\phi}$ jointly with the likelihood parameter $\boldsymbol{\theta}$ via the lower bound.

The likelihood distribution is from the parametric family, learned by the decoder network $g_{\boldsymbol{\theta}}$ as follows,

$$\mathrm{x}^{(k)} = g_{\boldsymbol{\theta}}(\mathrm{y}^{(k)}, \mathrm{z}^{(k)}) \sim p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{y}^{(k)}, \mathbf{z}^{(k)}) \tag{6}$$

According to the properties of the latent variables, we assume the prior distributions are as follows:

$$p(\mathbf{y}) = \mathcal{N}(\boldsymbol{\alpha}, \mathbf{I}),\ p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{7}$$

where $\mathbf{y}, \boldsymbol{\alpha} \in \mathbb{R}^{D_s}, \mathbf{z} \in \mathbb{R}^{D_c}$ and $\boldsymbol{\alpha}$ is domain-related. The choices of domain parameter $\boldsymbol{\alpha}_S, \boldsymbol{\alpha}_T$ can be arbitrary as long as $\boldsymbol{\alpha}_S \neq \boldsymbol{\alpha}_T$.

The approximated posterior distributions in this case are also Gaussian with learned parameters by network $f_{\boldsymbol{\phi}}$,

$$q_{\boldsymbol{\phi}_s}(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}; \hat{\boldsymbol{\mu}}_s, \hat{\boldsymbol{\sigma}}_s^2\boldsymbol{I}), q_{\boldsymbol{\phi}_c}(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}; \hat{\boldsymbol{\mu}}_c, \hat{\boldsymbol{\sigma}}_c^2\boldsymbol{I}) \tag{8}$$

where $\hat{\boldsymbol{\mu}}_s, \hat{\boldsymbol{\sigma}}_s^2 \in \mathbb{R}^{D_s}$ and $\hat{\boldsymbol{\mu}}_c, \hat{\boldsymbol{\sigma}}_c^2 \in \mathbb{R}^{D_c}$.

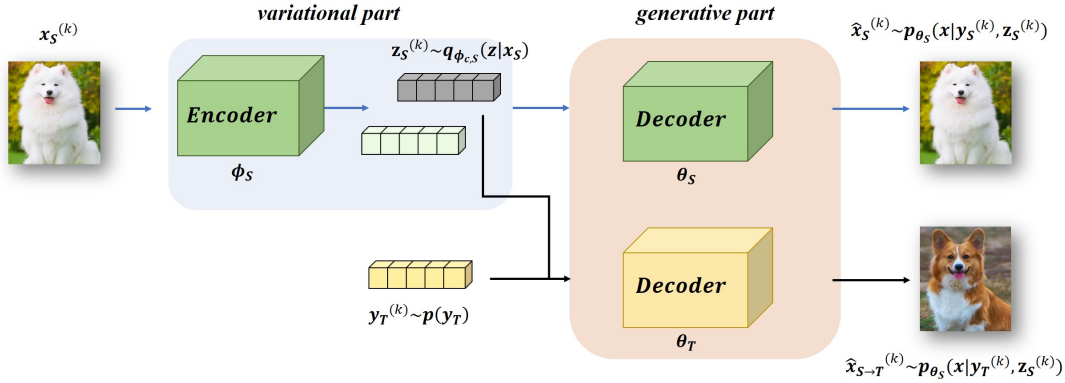

Fig. 1. The network architecture of the proposed VBITN. The framework consists of VAE-based networks which individually extract latent variables from different domain images. Then the learned latent variables are combined to generate new images.

3.3. Parametric Form of ELBO

To utilize the gradient descent algorithm for network learning, we derive the analytical version of the variational lower bound with respect to parameters $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$, expressed as follows,

$$\begin{aligned}\mathcal{L}(\boldsymbol{\phi},\boldsymbol{\theta};\mathbf{x}) = &\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{y},\mathbf{z}|\mathbf{x})}\big[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{y},\mathbf{z})\big] \\ &- \mathrm{KL}\left(q_{\boldsymbol{\phi}}(\mathbf{y}|\mathbf{x})||p(\mathbf{y})\right) - \mathrm{KL}\left(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})||p(\mathbf{z})\right)\end{aligned} \tag{9}$$

The last two terms can be integrated analytically with the Gaussian assumptions. The first term is evaluated as follows,

$$\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{y},\mathbf{z}|\mathbf{x})}\big[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{y},\mathbf{z})\big] = \frac{1}{L}\sum_{l=1}^{L}\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{y}^{(l)},\mathbf{z}^{(l)}) \tag{10}$$

where $\boldsymbol{\epsilon}^{(l)} \sim \mathcal{N}(\mathbf{0},\boldsymbol{I})$, $\mathbf{y}^{(l)} = \hat{\boldsymbol{\mu}}_s + \hat{\boldsymbol{\sigma}}_s^2 \odot \boldsymbol{\epsilon}^{(l)}$, $\mathbf{z}^{(l)} = \hat{\boldsymbol{\mu}}_c + \hat{\boldsymbol{\sigma}}_c^2 \odot \boldsymbol{\epsilon}^{(l)}$ using the so-called reparameterization trick [2].

3.4. Network Learning

Suppose we are given a dataset $\mathbb{X}_S$ from the source domain, and N unpaired datasets $\{\mathbb{X}_{T_i}\}_{i=1}^N$ from the target domains. The target is to translate some sample $\mathbf{x}_S^{(k)}$ from domain S to its counterpart with the mixed style of the regarding target domains. We adopt a compound loss with three terms: an inter-domain loss $\mathbb{L}_{\text{ind}}$ for latent variables inference, an adversarial loss $\mathbb{L}_{\text{adv}}$ to enforce realism of the translated images, and a reconstruction loss $\mathbb{L}_{\text{rec}}$ in for latent variable regularization.

Denote $\boldsymbol{\phi}_S$ and $\boldsymbol{\theta}_S$ for parameters of domain S, while $\boldsymbol{\phi}_{T_i}$ and $\boldsymbol{\theta}_{T_i}$ for domain T_i. We first implement the inter-domain loss $\mathbb{L}_{\text{ind}}$ as the expectation of negative inter-domain bounds on corresponding datasets:

$$\mathbb{L}_{\text{ind}} = \mathbb{E}_{\mathbb{X}_S}\big[\mathcal{L}(\boldsymbol{\theta}_S,\boldsymbol{\phi}_S;\mathbf{x})\big] + \sum_{i=1}^{N}\mathbb{E}_{\mathbb{X}_{T_i}}\big[\mathcal{L}(\boldsymbol{\theta}_{T_i},\boldsymbol{\phi}_{T_i};\mathbf{x})\big]. \tag{11}$$

The next two terms, $\mathbb{L}_{\text{rec}}$ and $\mathbb{L}_{\text{adv}}$ are constructed to form regularization in both latent space and image space to constraint learning, expressed as follows,

$$\begin{aligned}\mathbb{L}_{\text{rec}} = \sum_{i=1}^{N} &\mathbb{E}_{\mathbb{X}_{S\to T_i}}\mathbb{E}_{q_{\boldsymbol{\phi}_S}(\mathbf{z}|\mathbf{x})}\big[\|\mathbf{z}-\mathbf{z}_S\|^2\big] \\ &+ \mathbb{E}_{\sim\mathbb{X}_{S\to T_i}}\mathbb{E}_{q_{\boldsymbol{\phi}_{T_i}}(\mathbf{y}|\mathbf{x})}\big[\|\mathbf{y}-\mathbf{y}_{T_i}\|^2\big]\end{aligned} \tag{12}$$

$$\begin{aligned}\mathbb{L}_{\text{adv}} = \sum_{i=1}^{N} &\mathbb{E}_{\mathbb{X}_{T_i}}\Big[\log\big(1-D_{\boldsymbol{\varphi}}(\mathbf{x})\big)\Big] \\ &+ \mathbb{E}_{\mathbb{X}_{S\to T_i}}\big[\log D_{\boldsymbol{\varphi}}(\mathbf{x})\big]\end{aligned} \tag{13}$$

where $D_{\boldsymbol{\varphi}}(\cdot)$ denotes the discriminator network with parameter $\boldsymbol{\varphi}$ to distinguish between true and generated images.

4. EXPERIMENTS

4.1. Experimental Setup

We compare our Bayesian framework on image translation task with several classic methods, including Cycle GAN [10], Bicycle GAN [18], Disco GAN [19], Dual GAN [20] utilizing cycle-consistency techniques, and UNIT [11] and MUNIT [13] utilizing latent representation techniques[1].

We evaluate our techniques on the 'Monet's painting $\leftrightarrow$ photo' dataset and CelebA dataset [24], all at resolution of $128px$. Quantitative comparisons with related methods are conducted by the Learned Perceptual Image Patch Similarity (LPIPS) distance [16] for diversity, and Amazon Mechanical Turk (AMT) perceptual [17] for realism, claimed sufficient in other literature.

[1]Note that only classic framework-level methods for the basic image-to-image translation are adopted as baselines. More advanced works like StyleGAN [21, 22] and StarGAN [23] are not compared and can be viewed as implementable techniques on any basic frameworks.

Table 1. LPIPS [16] and AMT [17] scores for different methods on unsupervised image-to-image translation on dataset 'Monet's painting↔Photo'. The best two results are highlighted in red and blue colors respectively.

METHOD	PHOTO→MONET'S PAINTING		MONET'S PAINTING→PHOTO	
	LPIPS (DIVERSITY)	AMT (REALISM)	LPIPS (DIVERSITY)	AMT (REALISM)
CYCLEGAN [10]	.6705± .0025	37.28±2.26%	.6604± .0031	17.58±2.24%
BICYCLEGAN [18]	.5982± .0026	19.31±1.89%	.5805± .0026	15.46±2.43%
DISCOGAN [19]	.6775±.0026	31.49±2.67%	.6667± .0027	24.43±3.01%
DUALGAN [20]	.6957±.0029	15.84±2.28%	.7012±.0030	19.29±2.13%
UNIT [11]	.6734± .0026	34.22±2.46%	.6661± .0024	21.43±1.89%
MUNIT [13]	.4544± .0028	17.86±2.89%	.6536± .0027	13.85±2.75%
VBITN (OURS)	.6997± .0024	38.62±2.24%	.6725± .0022	27.30±1.87%

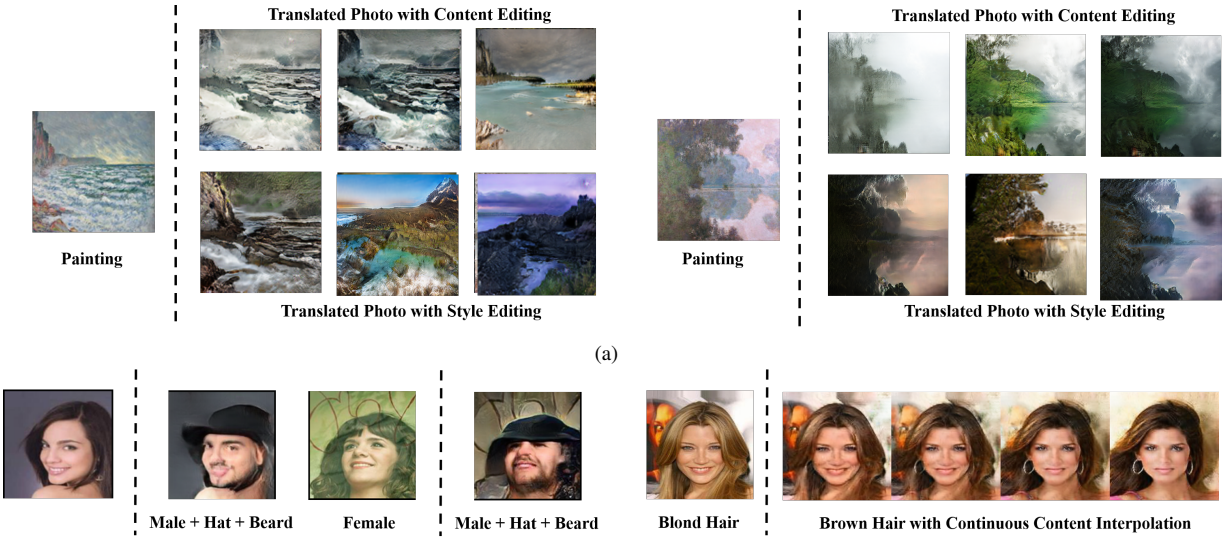

Fig. 2. VBITN enables efficient unsupervised image-to-image translation as well as semantic editing and mixed domain translation: (a) Paintings are translated to photos with different semantics; (b) Mixed domain translation on human face attributes.

4.2. Unsupervised Image-to-Image Translation

Table 1 reports the achieved performance different methods on LPIPS metric and AMT studies. We observe that our competitors tend to suffer from a trade-off between diversity and realism, though achieve remarkable results in one of the metrics. Our method gets the best of both sides, as it encourages diverse outputs with semantic variables and also has a well-defined objective function for regularization.

4.3. Multiple Variants

Qualitative results of our method on semantic editing are shown in Figure 2(a). We observe that both content and style semantics of the generated image can have meaningful variants with little cost to quality. Figure 2(b) shows our test on the novel mixed domain translation. The domain-related attributes (style) 'male', 'hat' and 'beard' have been successfully translated, while the domain-unrelated attributes (content) like 'looks' and 'expressions' are randomly sampled. Our method can produce translated image with multiple attributes with sharp edges and reliable details.

5. CONCLUSION

We introduced a Bayesian framework for conditional generative problems, and proposed VBITN for related tasks. The contributions include regularizing the ill-posed nature of image translation, and enabling novel capabilities like semantic editing. The developed techniques also suggest potential of combining DGMs and statistic tools to develop inference ability. Future work will tackle more scalable frameworks via delicate designs in latent space and graphic model.

6. REFERENCES

[1] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014.

[2] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *CoRR*, vol. abs/1312.6114, 2014.

[3] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 105–114, 2017.

[4] R. Zhang, J.-Y. Zhu, P. Isola, X. Geng, A. Lin, T. Yu, and A. A. Efros, "Real-time user-guided image colorization with learned deep priors," *ACM Trans. Graph.*, vol. 36, pp. 119:1–119:11, 2017.

[5] C. Yang, X. Lu, Z. L. Lin, E. Shechtman, O. Wang, and H. Li, "High-resolution image inpainting using multi-scale neural patch synthesis," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4076–4084, 2017.

[6] G. Liu, F. Reda, K. Shih, T. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," *ArXiv*, vol. abs/1804.07723, 2018.

[7] T. Park, M.-Y. Liu, T. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2332–2341, 2019.

[8] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *ArXiv*, vol. abs/1411.1784, 2014.

[9] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," in *NIPS*, 2016.

[10] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2242–2251, 2017.

[11] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *NIPS*, 2017.

[12] Y. Liu, M. De Nadai, J. Yao, N. Sebe, B. Lepri, and X. Alameda-Pineda, "Gmm-unit: Unsupervised multi-domain and multi-modal image-to-image translation via attribute gaussian mixture modeling," *arXiv preprint arXiv:2003.06788*, 2020.

[13] X. Huang, M.-Y. Liu, S. J. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *ECCV*, 2018.

[14] T. Park, J.-Y. Zhu, O. Wang, J. Lu, E. Shechtman, A. A. Efros, and R. Zhang, "Swapping autoencoder for deep image manipulation," *ArXiv*, vol. abs/2007.00653, 2020.

[15] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2414–2423, 2016.

[16] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

[17] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *ECCV*, 2016.

[18] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman, "Toward multimodal image-to-image translation," in *NIPS*, 2017.

[19] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," *ArXiv*, vol. abs/1703.05192, 2017.

[20] Z. Yi, H. Zhang, P. Tan, and M. Gong, "Dualgan: Unsupervised dual learning for image-to-image translation," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2868–2876, 2017.

[21] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8107–8116, 2020.

[22] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4396–4405, 2019.

[23] Y. Choi, M.-J. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8789–8797, 2018.

[24] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 3730–3738, 2015.