VARIATIONAL DOMAIN ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper proposes *variational domain adaptation*, a unified, scalable, simple framework for learning multiple distributions through variational inference. Unlike the existing methods on domain transfer through deep generative models, such as CycleGAN (Zhu et al., 2017a) and StarGAN (Choi et al., 2017), the variational domain adaptation has three advantages. Firstly, the samples from the target are not required. Instead, the framework requries one known source as a *prior* $p(x)$ and binary *discriminators*, $p(\mathcal{D}_i|x)$, discriminating the target domain $\mathcal{D}_i$ from others. Consequently, the framework regards a target as a *posterior* that can be explicitly formulated through the Bayesian inference, $p(x|\mathcal{D}_i) \propto p(\mathcal{D}_i|x)p(x)$, as exhibited by a further proposed model of *multi-domain variational autoencoder* (MD-VAE). Secondly, the framework is scablable to large-scale domains. MD-VAE sophisticatedly puts together all the domains as well as the samples drawn from the prior into normal distributions in the same latent space as *embeddings*. The model enables us to expand the method to uncountable infinite domains such as continuous domains as well as interpolation. Thirdly, with MD-VAE, no need to search hyperparameter anymore. Although several domain transfer based on adversarial learning need sophisticated automatic/manual hyperparameter search, MD-VAE fast converges with less tuning because it has only one trainable matrix in addition to VAE. In the experiment part, we experimentally demonstrate the benefit with multi-domain image generation task on CelebA and facial image data that are obtained based on evaluation by 60 users, the model generates an ideal image that can be evaluated to be good by multiple users. Additionally, our experimental result exhibits that our model outperforms several state-of-the-art models.

1 INTRODUCTION

"...we hold that all the loveliness of this world comes by communion in Ideal-Form. *All shapelessness whose kind admits of pattern and form, as long as it remains outside of Reason and Idea, is ugly from that very isolation from the Divine-Thought."* — Plato (427 – 347 bc)

Agents that interact in various environments have to handle multiple distributions of observations. Domain adaptation (Bengio, 2012) is a methodology to exploit deep generative models, such as adversarial learning (Goodfellow et al., 2014) and variational inference (Kingma & Welling, 2013), for handling distributions that vary with environments and other agents. Further, multi-task learning and domain transfer are examples of the usage of domain adaptation methodology. We focus on domain transfer involving transfers across a distribution between domains. For instance, pix2pix (Isola et al., 2017) outputs the sample from the target domain that corresponds to the input sample from the source domain. This can be achieved by learning the pair relation of samples from the source and target domains. CycleGAN (Zhu et al., 2017a) transfers the samples between two domains using samples obtained from both the domains. Similarly, UNIT (Liu et al., 2017), DiscoGAN(Kim et al., 2017), and DTN(Taigman et al., 2016) have been proposed.

However, the aforementioned method requires samples that are obtained from the target domains; because of this requirement, it cannot be applied to domains for which direct sampling is expensive or sometimes impossible. For example, the desired, continuous, high-dimensional action in the environment, intrinsic reward (e.g., preference and curiosity) and the policy of the interacting agents other than itself cannot be sampled from inside, while they only can discriminate the proposed input. Even for ourselves, the concept of beauty or interest in our conscious is subjective, complex, and difficult to be sampled from the inside but is easy to discriminate on the outside.

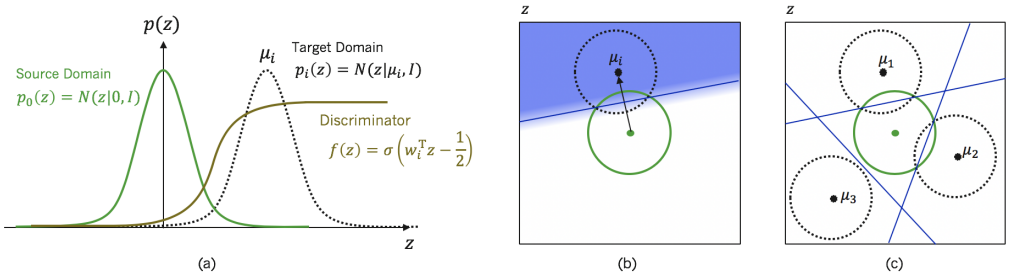


Figure 1: Major concept of variational domain adaptation. **a)** Given the proposal drawn from the prior, the discriminator discriminates the target domain from the others. Each domain is posterior for the prior $\mathcal{N}(z|0, 1)$; further, the distribution in the latent space is observed to be a normal distribution using the conjugate likelihood. **b)** Domain transfer is represented by the mean shift on the latent space. **c)** Domain embedding: After training, all the domains can be only represented by vectors μ_i.

In this study, we propose *variational domain adaptation*, which is a framework for targets that pose problems in direct sampling. One solution is *multi-domain semi-supervision*, which converts the problem to semi-supervised learning, thereby enabling us to perform variational inference. In this supervision, a source domain is regarded as a *prior* $p(x)$ and a target domain is considered to be a *posterior* $p(x|\mathcal{D}_i)$ by referring to the label given by a supervised *discriminator* $p(\mathcal{D}_i|x)$ that distinguishes the target domain from others. Our model imitates the behavior of the discriminator and models the target domain using a simple conclusion of the Bayesian theorem, $p_\theta(x|\mathcal{D}_i) \propto p_\theta(\mathcal{D}_i|x)p_\theta(x)$. The end-to-end learning framework enables us to learn good prior $p_\theta(x)$ with respect to all the domains. After the training was completed, the posterior $p_\theta(x|\mathcal{D}_i)$ succeeds in deceiving the discriminator $p(\mathcal{D}_i|x)$. This concept is similar to rejection sampling in the Monte Carlo methods. Further, the variational domain adaptation is the first contribution of this study.

The second contribution of this study is a model of *multi-domain VAE* (MD-VAE), which is a simple extension of the conditional VAE (Kingma et al., 2014), to exhibit our concept of multi-domain semi-supervision. MD-VAE learns multiple domains in one network by maximizing the variational lower bound of the total negative KL-divergence between the target domain and the model. MD-VAE uses VAE to model the prior $p(x)$ and abstract representation for the discriminator $p(\mathcal{D}_i|x)$. The major feature of MD-VAE is *domain embedding* that states that all the posteriors are modeled as a normal distribution $\mathcal{N}(z|\mu_i, \sigma^2)$ on the same latent space $\mathcal{Z}$ using the conjecture distribution of the prior. Here, μ_i is the domain embedding that represents the domain $\mathcal{D}_i$. This enables us to sample from $p_\theta(x|\mathcal{D}_i)$. Further, we exhibit that the discriminator of MD-VAE is a simple inner product between the two means of domain embedding and output of VAE—$\log p_\theta(\mathcal{D}_i|x) \propto \mu_i^{\mathrm{T}} \mu(x)$— that acts as a kernel function between the two objects. The probabilistic end-to-end model learns multiple domains in one network, making it possible to obtain the effect of transfer learning and to learn data that multi-domains cannot observe from sparse feedback. Domain embedding is a powerful tool and allows us to use VAE instead of GANs.

The third contribution of this study is that MD-VAE was validated for use in a recommendation task using celebA (Liu et al., 2015). In the experiment, using celebA and face image data obtained based on evaluation by 60 users, we generated an image based on the prediction of user evaluation and an ideal image that was evaluated to be good by multiple users. We demonstrated that the image could be modified to improve the evaluation by interpolating the image. We evaluated the image using the preferential inception score (PIS), which is the score of the model that has learned the preference of each user. We present the beauty inside each evaluators by simply sampling from $p_\theta(x|\mathcal{D}_i)$. The PIS of MD-VAE is 5% higher than that of a single domain. The dataset and code are available online.

2 Related Work

The existing literature related to domain transfer is based on the assumption that the samples are obtained from the target domain. For example, pix2pix(Isola et al., 2017) can output the samples

from the target domain that correspond to the input sample from the source domain by learning the pair relation between the samples of the source and target domains. CycleGAN (Zhu et al., 2017a), which differs from pix2pix, does not require sample pairs from both domains. Similarly, UNIT (Liu et al., 2017), DiscoGAN(Kim et al., 2017), and DTN(Taigman et al., 2016) also do not require sample pairs. Furthermore, because there are few cases in which samples from the source and target domains form a one-to-one pair in real world research after being extended to one-to-many conversion, including BicycleGAN(Zhu et al., 2017b) and MUNIT(Huang et al., 2018).

Several studies were conducted to model multiple distributions in a semi-supervised manner. StarGAN(Choi et al., 2017), UFDN(Liu et al., 2018), and RegCGAN(Mao & Li, 2018) are extensions of the aforementioned models and are frameworks that can convert the source domain samples into samples for various target domains with one network structure. However, the problem with these methods is hyperparameter tuning, which arises from the characteristics of adversarial learning. MD-VAE is a simple extension of conditional VAE in a multi-domain situation. Conditional VAE utilizes VAE for semi-supervised learning. Although the model is quite simple, it is powerful and scalable to learn multiple distributions with *domain embedding*. In fact, we demonstrated that MD-VAE quickly converged for more than 30 domains without sophisticated hyper-parameter tuning. In the experiment conducted in this study, we evaluated $\mathbb{E}_{\alpha}\left[J(\theta|\alpha)\right]$ instead of $J(\theta|\hat{\alpha})$ to denote that our method required less hyperparameter tuning.

3 METHOD

3.1 PROBLEM SETTING

With regard to n domains $\mathcal{D}_1, \ldots, \mathcal{D}_n$, and a sample x on an observation space $\mathcal{X}$, the objective of multi-domain learning is to minimize the KL-divergence between the target distribution and the model, $D_{\mathrm{KL}}\left(p^{(i)}(x) \| p^{(i)}(x,\theta)\right)$, over all the domains $\mathcal{D}_i$. From the perspective of optimizing parameter θ, minimizing the KL divergence is equivalent to maximizing the cross-entropy. Hence, the objective can be written as follows:

$$\text{Maximize}_{\theta}: \quad J(\theta) = \frac{1}{n}\sum_{i=1}^{n} \lambda_i \, \mathbb{E}_{x\sim p^{(i)}}\left[\log p_{\theta}^{(i)}(x)\right] + \bar{\lambda}\log p(\theta), \tag{1}$$

where $p_{\theta}^{(i)}(x) = p^{(i)}(x|\theta)$, $\lambda_i \in [0,1]$ is importance of each domain $\mathcal{D}_i$ and $\bar{lambda} = \sum_i^n \lambda_i/n$ If $\lambda_i = 1/n$ for all the i's, the objective function is simple mean, and if $\lambda_i = 0$ for certain i, the domain $\mathcal{D}_i$ is ignored. Note that $p^{(i)}(x,\theta) = p^{(i)}(x|\theta)p(\theta)$.

The difficulty arises from the fact that we cannot directly sample x from $p^{(i)}$ and that we can directly sample x from the likelihood $p(\mathcal{D}_i|x)$. This challenge motivated us to consider multi-domain semi-supervision.

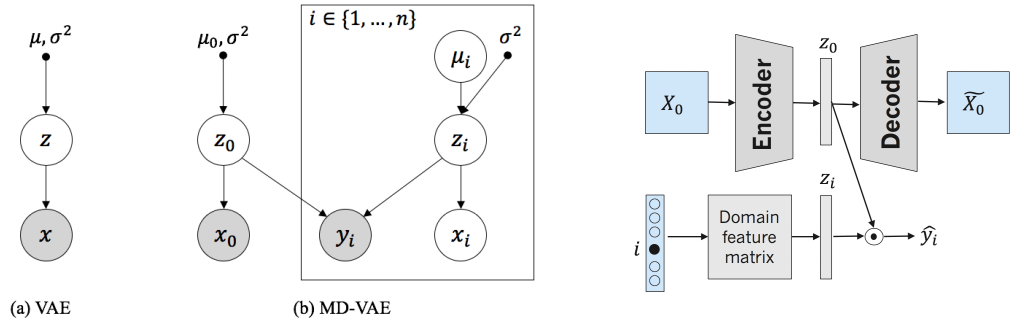
3.2 MULTI-DOMAIN SEMI-SUPERVISION

Multi-domain semi-supervision assumes a prior $p(x)$ and models each the domain as a posterior $p^{(i)} = p(x|\mathcal{D}_i)$. As the Bayesian inference, we reformulate the cross-entropy $\mathbb{E}_{x\sim p^{(i)}}\left[\log p_{\theta}(x|\mathcal{D}_i)\right]$ in Eq. (1) as follows:

$$\begin{aligned}
\mathbb{E}_{x\sim p^{(i)}}\left[\log p_{\theta}(x|\mathcal{D}_i)\right] &= \int p(x|\mathcal{D}_i)\log p_{\theta}(x|\mathcal{D}_i)dx = \int \frac{p(\mathcal{D}_i|x)p(x)}{p(\mathcal{D}_i)}\log\frac{p_{\theta}(\mathcal{D}_i|x)p_{\theta}(x)}{p_{\theta}(\mathcal{D}_i)}dx \\
&= \mathbb{E}_{x\sim p}\left[\frac{p(\mathcal{D}_i|x)}{p(\mathcal{D}_i)}\log\frac{p_{\theta}(\mathcal{D}_i|x)}{p_{\theta}(\mathcal{D}_i)}\right] + \mathbb{E}_{x\sim p}\left[\frac{p(\mathcal{D}_i|x)}{p(\mathcal{D}_i)}\log p_{\theta}(x)\right] \\
&= \mathbb{E}_{x\sim p}\left[f(\mathcal{D}_i|x)\log f_{\theta}(\mathcal{D}_i|x)\right] + \mathbb{E}_{x\sim p}\left[f(\mathcal{D}_i|x)\log p_{\theta}(x)\right],
\end{aligned} \tag{2}$$

where $f(\mathcal{D}_i|x) = p(\mathcal{D}_i|x)/p(\mathcal{D}_i)$ and $f_{\theta}(\mathcal{D}_i|x) = p_{\theta}(\mathcal{D}_i|x)/p_{\theta}(\mathcal{D}_i)$. By letting $\lambda_i = p(\mathcal{D}_i)$, the objective is identical to:

$$J(\theta) = \underbrace{\mathbb{E}_{x\sim p, i\sim[n]}\left[p(\mathcal{D}_i|x)\log f_{\theta}(\mathcal{D}_i|x)\right]}_{\text{discriminator}} + \underbrace{\mathbb{E}_{x\sim p}\left[\log p_{\theta}(x)\right]}_{\text{prior}} + \underbrace{\bar{\lambda}\log p(\theta)}_{\text{regularizer}}, \tag{3}$$



(a) VAE (b) MD-VAE

Figure 2: **Left**: Graphical models of the probabilistic models of VAE and MD-VAE. The gray and white circles indicate the observed variables and latent variables, respectively. Symbols without circles indicate the constants. Arrows between the symbols indicate probabilistic dependency (e.g., X generates Y). A rectangle with suffixes indicates a block, which comprises multiple elements. **Right:** The network structure of MD-VAE. The label is structured as the inner product of latent z_θ and domain embedding z_i.

where $[n]$ is a uniform distribution over $\{1, \ldots, n\}$ and $f(\bar{\mathcal{D}}|x) = \mathbb{E}_{i\sim[n]}\left[f(\mathcal{D}_i|x)\right]$. The first term is the likelihood from the discriminator; the second term is the prior learned by a generative model, including VAE; and the last term is the regularizer.

Because the equation is intractable, we use Monte Carlo sampling to estimate the function. During the estimation, we initially sample $x_1, \ldots, x_m$ from the prior $p(x)$ and subsequently obtain the binary labels $y_{ij} \in \{0, 1\}$ from each discriminator $y_{ij} \sim p(\mathcal{D}_i|x_j)$. Since the number of labels from supervises is nm, consider the situation that the sparse labels: $k << nm$. Further, some discriminators only provide parts of labels. In the situation, the missing value is 0-padded: $y_{ij} = 0$.

$$J(\theta) \approx \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}\log f_\theta(y_{ij}|x_j) + \frac{1}{m}\sum_{j=1}^{m}\log p_\theta(x_j) + \bar{y}\log p(\theta), \tag{4}$$

where $\approx$ indicates Monte Carlo estimation and $\bar{y} = \sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}/k$. In the limit of $n \to \infty$, the right hand side is identical to the left hand side.

3.3 Multi-domain Variational Autoencoder (MD-VAE)

We extended VAE for multi-domain transfer to demonstrate our concept of multi-domain semi-supervision. Our proposed model, multi-domain VAE (MD-VAE), models each domain $p_i(x)$ as a posterior distribution $p(x|\mathcal{D}_i)$ that is similar to that observed in conditional VAEs. Fig. 2 depicts the graphical models of VAE and MD-VAE.

The major feature of MD-VAE is *domain embedding* that denotes that all the domains and prior share the same latent space $\mathcal{Z}$. For the prior distribution, $p(z) = \mathcal{N}(z|0, 1)$ and $p(z|\mathcal{D}_i) = \mathcal{N}(z|\mu_i, \sigma^2)$, where $\mu_i \in \mathcal{Z}$ is an embedding. The domain $\mathcal{D}_i$ is characterized only by embedding μ_i. Here, μ_0 is the embedding of the prior that can be assumed as $\mu_0 = 0$.

Virtually, training MD-VAE is equivalent to simultaneously training $(n + 1)$ VAEs which sharing a parameter, including the prior. Using conjesture distribution to the prior $p(z)$, the posterior distribution is observed to be a normal distribution. Therefore, *all the posteriors are VAEs*. The joint distribution can be given as follows:

3.3.1 VAE: The Prior $p_\theta(x)$

We use a variational autoencoder (VAE) (Kingma & Welling, 2013) to model the prior $p(x)$, a deep generative model which uses autoencoder to model the hidden variable as random variable. The benefit of VAE is that we can model the each distribution as a normal distribution in $\mathcal{Z}$, achieved by maximizing the variational lower bound of $\log p(x)$ as follows.

$$\log p_\theta(x) \geq \mathcal{L}_\theta(x) = \mathbb{E}_{z\sim q_\phi(\cdot|x)}\left[\log p_w(x|z)\right] - D_{\mathrm{KL}}\left(q_\phi(z|x)\|p(z)\right), \tag{5}$$

where $\phi, w \in \theta$ is a parameter of the encoder and the decoder, respectively. The objective is to learn a pair of the encoder $p_w(x|z)$ and the decoder $q_\phi(z|x)$ to maximize $\mathcal{L}_(x)$. z is act as a prior $p(z) = \mathcal{N}(z|0, I)$.

The lower bound $\mathcal{L}_\theta(x)$ is constructed using the reconstruction error and penalty term as the KL divergence between the model and prior $p(z)$. Further, the gradient of reconstruction term can be calculated using the Monte Carlo method. Because the construction term is the KL divergence between two normal distributions, it can be analytically calculated.

3.3.2 DISCRIMINATOR $f_\theta(\mathcal{D}_i|x)$

Using the definition and the Bayesian theorem, $\log f_\theta(\mathcal{D}_i|x)$ can be written as follows:

$$\begin{aligned}\log f_\theta(\mathcal{D}_i|x) &= \log \int \frac{p_\theta(\mathcal{D}_i|z)p_\theta(z|x)}{p_\theta(\mathcal{D}_i)} dz = \log \int \frac{p_\theta(z|\mathcal{D}_i)p_\theta(z|x)}{p_\theta(z)} dz \\ &= \log \int \frac{\mathcal{N}(z|\mu_i, \sigma^2)\mathcal{N}(z|\mu_\phi(x), \sigma^2)}{\mathcal{N}(z|0, 1)} dz = \frac{\mu_i^{\mathrm{T}} \mu_\phi(x)}{\sigma^2}, \end{aligned} \tag{6}$$

The equation above indicates $\log f_\theta(\mathcal{D}_i|x)$ can be written as simple inner product between μ_i and $\mu_\phi(x)$. The objective can be written as follows:

$$\mathbb{E}_{i\sim[n]} \left[y_i \log f_\theta(\mathcal{D}_i|x) \right] = \frac{\mathbf{y}^{\mathrm{T}} U \mu_\phi(x)}{n\sigma^2} = \alpha \mu_U(\mathbf{y})^{\mathrm{T}} \mu_\phi(x), \tag{7}$$

where $U = (\mu_1, \ldots, \mu_n)^{\mathrm{T}} \in \mathcal{Z}^n$, $\mu_U = U^{\mathrm{T}}\mathbf{y}/n$ and $\alpha = \sigma^{-2}$. Interstingly, it requries only one additional parameter U except of a hyperparameter α. We call U as a *domain embedding matrix*, representing set of the prototypes. Domain embeding enables us to extend our method to infinite domains such as continuous domain. In fact, $\mu_U(\mathbf{y}) \in \mathcal{Z}$ represents prototype of mixed domains indicated by $\mathbf{y}$.

3.3.3 THE REGULARIZER $p(\theta)$

The overall parameters of MD-VAE is $\theta = (w, \phi, U)$, where w is encoder's parameter and ϕ is the decoders's parameter and U is domain embedding matrix. While typical VAE does not assume any distribution of w, ϕ, we set $p(U)$ as a exponential distribution to obtain sparse representation: $p(U) \propto \exp(-\beta \|U\|_1/\bar{\lambda})$, thus,

$$\log p(\theta) = \log p(U) = -\frac{\beta}{\bar{\lambda}} \|U\|_1 - m \dim \mathcal{Z} \log 2 + \log \beta - \log \bar{\lambda}. \tag{8}$$

As the terms except of the first are independent to θ, we ignore them later as a constant.

3.3.4 THE FINAL FORM

By putting together the prior and the discriminator, the variational lower bound $\mathcal{L}_\theta^{\mathrm{MD}}(x, \mathbf{y})$ of the point-wise objective of MD-VAE $J(\theta|x, \mathbf{y})$ can be written as a surprisingly simple form:

$$\mathcal{L}_\theta^{\mathrm{MD}}(x, \mathbf{y}) = \mathcal{L}_\theta(x) + \alpha \mu_U(\mathbf{y})^{\mathrm{T}} \mu_\phi(x) - \beta \|U\|_1, \tag{9}$$

where $\alpha, \beta \in [0, \infty)$ are hyperparameters and $\|\cdot\|_1$ is a 1-norm. Note that MD-VAE requires only *two* additional hyperparameters in addition to VAE. If $\alpha = \beta = 0$, it is equivalent to the VAE. Intuitively, $1/\alpha$ and $1/\beta$ control variance and bias of the domain embeddings, respectively.

The training algorithm of MD-VAE is shown in Algorithm 1.

4 EXPERIMENT

Based on an original numerical experiment in domain adaptation, we confirmed that MD-VAE learns multiple distributions both qualitatively and quantitatively. Similar to the case of the existing methods, domain adaptation was confirmed via an image-generation task in this study. We performed a *facial image recommendation task*, which is the task of a content-based recommendation for generating the preferences of users.

Algorithm 1 Variational domain adaptation through MD-VAE

Require: observations $(x_j)_{j=1}^m$, batch size M, VAE/encoder optimisers: g, g_e, hyperparameter λ, label matrix Y
Initialize encoder, decoder and domain embedding parameters: ϕ, θ, U
repeat
Randomly select batch $(x_j)_{j \in \mathcal{B}}$ of size M
Sample $z_j \sim q_\phi(z|x_j)\ \forall j \in \mathcal{B}$
$\phi, \theta \leftarrow g(\nabla_{\phi,\theta} \sum_{j \in \mathcal{B}} [\log p_\theta(x_j|z_j) - D_{\mathrm{KL}}\left(q_\phi(z|x_j) \| p(z)\right)])$
$\phi, U \leftarrow g_e(\nabla_{\phi,U} \sum_{j \in \mathcal{B}} [(Y_{:,j} - U^T z_j)^2 + \lambda \|U\|_1])$
until convergence of parameters (ϕ, θ, U)

The objective of the task was to generate an image that was preferred by a specific user. We set the input space $\mathcal{X}$ as a raw image, the prior $p(x)$ as faces, and the domain $\mathcal{D}_i$ as a user. We use the dataset of CelebA as the samples from the prior. The objective of the task was to generate samples from $p_\theta(x|\mathcal{D}_i)$, exhibiting the images that were preferred by a user. We used label $y_{ij} \sim p(\mathcal{D}_i|x)$ as the existing dataset of SCUT-PBT5500 with 5,500 faces and 60 users for content-based recommendation.

The results exhibited that MD-VAE successfully learned the model of target distribution $p_\theta(x|\mathcal{D}_i)$ both qualitatively and quantitively. Qualitatively, after a low-dimensional plot of the latent space $\mathcal{Z}$ exhibits the domain embedding track or the so-called distinct facial features, we demonstrated that the image could be transferred to improve the evaluation by interpolating the image. Further, we exhibited several beautiful facial images that the users were conscious of by decoding each domain embedding μ_i, which can be the projection of the ideal form inside the users. Quantitatively, we confirmed that the discriminator learns the distribution by evaluating the negative log-likelihood loss, $-\log p_\theta(\mathcal{D}_i|x)$. We evaluated the samples using the *preferential inception score* (PIS), which is the score of the model that has learned the preference for each person. It was denoted that the PIS of MD-VAE was 5% higher than that of the single domain.

In this study, we first introduce the datasets and the experiment settings (Sec. 4.1, 4.2). By considering the ideal image distribution for each user as one domain, we demonstrated that MD-VAE generated an ideal image that was evaluated by multiple users as good and that the source image can be modified so that the evaluation will be improved by interpolating the target ideal image (Sec. 4.3). We quantitatively evaluated the image using PIS (Sec. 4.4).

4.1 Dataset

The experiment employed several existing datasets, including celebA and SCUT-PBT5550.

CelebA CelebA(Liu et al., 2015) comprises approximately 200 thousand images of faces of celebrities with 40 attributes. However, because human evaluation is not included in CelebA, we used it for unsupervised learning. We resized each image to 256 $\times$ 256 and randomly selected 500 images each as test data for males and females.

SCUT-FBP5500 SCUT-FBP5500(Liang et al., 2018) comprises 5500 face images and exhibits a 5-point scale evaluation by 60 people in terms of beauty preference. The beauty scores of each image are observed to differ among people. The face images can be categorized as Asian male, Asian female, Caucasian male, and Caucasian female, with 2000, 2000, 750, 750 images, respectively. We resized each image to 256 $\times$ 256 and randomly selected 5000 images as the training set and 500 images as the test set.

4.2 Qualitative results

The qualitative result of the experiment can be demonstrated by visualizing domain embedding, interpolation, and latent space.

Domain embeddings By reducing the dimensions of the learned 60 user vectors from 63 to 2 using UMAP, we visualized the user vectors by means of a scatter plot. Furthermore, we visualized $\mathbf{x}_i$ by decoding samples from the domain distribution.

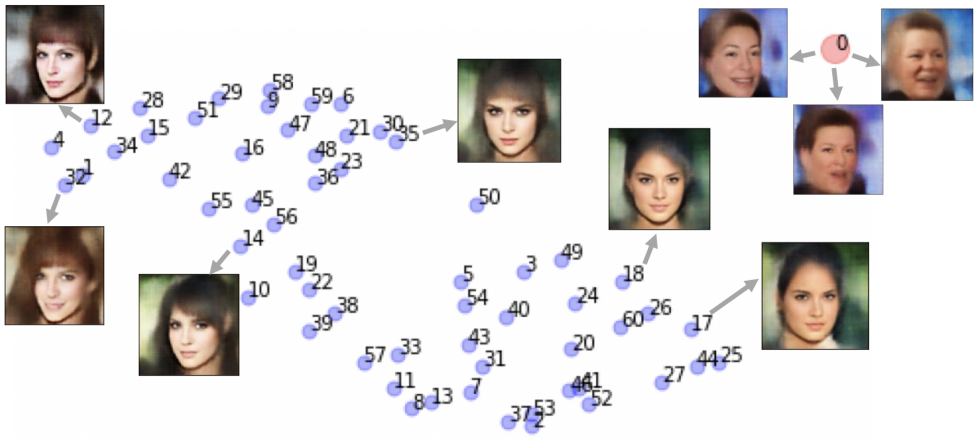


Figure 3: Scatter plot of the average vectors of the data distribution from each domain (user vector), and several decoded images of the samples from each domain. We visualized 60 user vectors using UMAP. We decoded 6 $\mathbf{z}_i$ from the target domain distribution and output $\mathbf{x}_i$. Furthermore, we decoded $\mathbf{z}_0$ from the source domain data distribution and output $\mathbf{x}_0$.

Interpolation We rectified the images by interpolating between the original and the ideal images. Because the evaluation for an image is $\boldsymbol{\mu}_i^T \mathbf{z}$, we observe that the bigger the norm of the item vector $\mathbf{z}$, the higher will be the score. This is applicable for all the evaluators. Therefore, we generated images by imposing some restrictions on the norm of $\mathbf{z}$ and by moving $\mathbf{z}$ to reach closer to each user vector $\boldsymbol{\mu}_i$ so that the score would increase. Further, we calculated the following vector $\mathbf{w}_i$:

$$\mathbf{w}_i = \frac{\mathbf{z} + \lambda \boldsymbol{\mu}_i}{||\mathbf{z} + \lambda \boldsymbol{\mu}_i||} ||\mathbf{z}||.$$

By changing λ, we denoted five images representing unideal to ideal reconstructions for each of the three sample users ($i = 14, 18, 32$). This corresponds to the user number in Figure 3, and interpolation is performed to approach close to the ideal image $\mathbf{x}_i$.

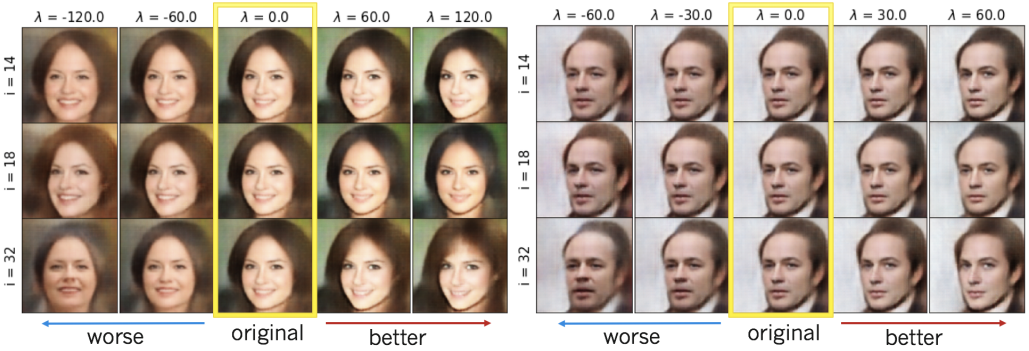


Figure 4: Images obtained from our model by decoding $w_i (i = 14, 18, 32)$ while changing the value of λ. The reconstructed images are present in the center.

4.3 Quantitative Results

MD-VAE vs single-domain VAE We evaluated the image using PIS. We proposed PIS as an evaluation metric for the generated images. PIS vectorizes the output of the Inceptionv3 (Szegedy

et al., 2016) to 60 dimensions, activates outputs using tanh, fine tunes the model parameters using the evaluation scores of the training set of SCUT-FBP5500, and finally outputs the predicted score for the generated image. To investigate the ability of PIS, we used the test data of SCUT-FBP5500 to calculate the RMSE of PIS as 0.296. Further, PIS can predict the score of a face image within an error of 0.6.

MD-VAE predicts all ideal images of target domains. As a competitive method, we use single-domain VAE (SD-VAE) that seperately learns the latent variables of five domains in a semi-supervised way and predicts five ideal images. For output images of these two models, the one has higher PIS value is considered to being able to output ideal images. Here, since the evaluation data of the training data is randomly lost by 90 for each user, the number of evaluations by each user is 500 out of 5000. We calculated the PIS of the ideal images generated by each model and normalized to $[0, 1]$. PIS of MD-VAE was 0.659 and PIS of SD-VAE was 0.609, 5 percent higher than PIS of SD-VAE.

MD-VAE vs StarGAN We compared MD-VAE with StarGAN as a model that can be converted into multiple domains on one network. In MD-VAE, the training was performed by randomly changing the dimension of the learning rate and the latent variable, hyper parameter of the prior distribution, and in the StarGAN, the training was performed by randomly changing the learning rate and the number of updates of the generator and classifier. MD-VAE gained a higher PIS as a result of classifying the generated images by the fine tuned Inception model.

Table 1: Comparison of StarGAN with MD-VAE

number of domains	10	40
StarGAN(Choi et al., 2017)	3.6×10^{-4}	0.064
MD-VAE	**0.87**	**0.36**

5 Conclusion

Variational domain adaptation, which is a unified framework for learning multiple distributions in one network, is proposed in this study. Our framework uses one known source as a prior $p(x)$ and binary discriminator $p(\mathcal{D}_i|x)$, thereby discriminating the target domain $\mathcal{D}_i$ from the others; this is in contrast with the existing frameworks in which samples undergo domain transfer through deep generative models. Consequently, our framework regards the target as a posterior that is characterized through the Bayesian inference, $p(x|\mathcal{D}_i) \propto p(\mathcal{D}_i|x)p(x)$. This is exhibited by the proposed model of MD-VAE. The major feature of MD-VAE is domain embedding, which is a powerful tool that encodes all the domains and the samples obtained from the prior into normal distributions in the same latent space as that learned by a unified network through variational inference. In the experiment, we applied our framework and model to a multi-domain image generation task. celebA and face image data that were obtained based on evaluation by 60 users were used, and the model generated an ideal image that was evaluated by multiple users to be good. We demonstrated that an image can be modified by interpolating it to improve the evaluation. We evaluated the images using PIS. The experimental result verified that our model outperformed several state-of-the-art frameworks.

Several directions can be considered for performing future studies. First, we intend to expand MD-VAE for learning in complex domains, such as high-resolution images with several models, for example, glow(Kingma & Dhariwal, 2018). Second, we will perform an experiment to consider wider domains for beauty. We expect that our method will contribute to the community and will help to deal with the paradigm of multiple contexts—multimodal, multi-task, and multi-agent contexts.

References

Yoshua Bengio. Deep learning of representations for unsupervised and transfer learning. In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, pp. 17–36, 2012.

Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. *arXiv preprint*, 1711, 2017.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. *arXiv preprint arXiv:1804.04732*, 2018.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint*, 2017.

Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. *arXiv preprint arXiv:1703.05192*, 2017.

Diederik P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *arXiv preprint arXiv:1807.03039*, 2018.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*, pp. 3581–3589, 2014.

Lingyu Liang, Luojun Lin, Lianwen Jin, Duorui Xie, and Mengru Li. Scut-fbp5500: A diverse benchmark dataset for multi-paradigm facial beauty prediction. 2018.

Alexander Liu, Yen-Chen Liu, Yu-Ying Yeh, and Yu-Chiang Frank Wang. A unified feature disentangler for multi-domain image translation and manipulation. *arXiv preprint arXiv:1809.01361*, 2018.

Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems*, pp. 700–708, 2017.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Xudong Mao and Qing Li. Unpaired multi-domain image generation via regularized conditional gans. *arXiv preprint arXiv:1805.02456*, 2018.

Leland McInnes and John Healy. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. *arXiv preprint arXiv:1611.02200*, 2016.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint*, 2017a.

Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 30*, pp. 465–476. Curran Associates, Inc., 2017b. URL http://papers.nips.cc/paper/6650-toward-multimodal-image-to-image-translation.pdf.

6 APPENDIX

6.1 LATENT SPACE

We visualized the latent descriptions of VAE and MD-VAE. VAE differs from MD-VAE because evaluation regression is not conducted during training. For each model, we can achieve 5500 latent vectors of 63 dimensions by encoding 5500 images from SCUT-FBP5500. We obtained a scatter plot after using UMAP (McInnes & Healy, 2018) to reduce the number of dimensions to two. The item average score is indicated by colors ranging from red to blue. As can be observed from the UMAP of MD-VAE, the gradient of the score is learned, and it represents the user vector in Figure 5.

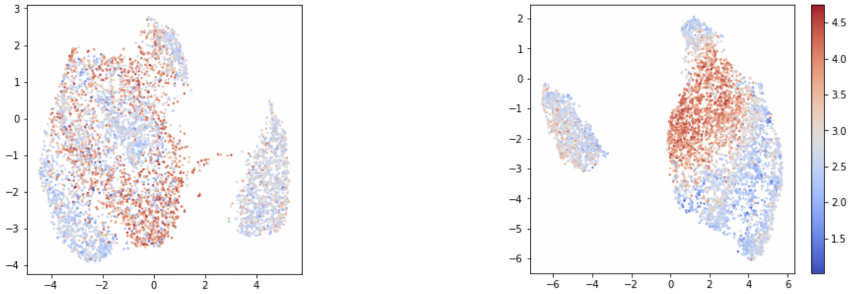


Figure 5: Latent visualization of VAE (left) and MD-VAE (right) demonstrates that MD-VAE learns a good prior to model the domains. The heat map indicates the mean score of all the users.

6.2 EXPERIMENT SETTINGS

We assumed that the beauty criterion required for evaluating the facial images depends on the gender of a person in the target image. Therefore, we added the gender information to the images. For this purpose, we applied CGAN (Mirza & Osindero, 2014) to VAE. We normalized the scoring in [?1, 1] to accelerate the learning.

Subsequently, we considered the specific model structure of MD-VAE. Both the input and output images were RGB images, $x \in R^{256\times256\times3}$. We used convolution networks for the encoder and stride 2 for convolution and no pooling. Convolution, batch normalization(Ioffe & Szegedy, 2015), and LeakyReLU were repeated four times and were subsequently connected to fully connected layers. Further, after batch normalization and LeakyReLU layers, a 63-dimensional latent variable was obtained. The decoder exhibited a completely symmetric shape with deconvolution layers instead of convolution layers. Furthermore, as the gender attribute, we set 0 as female and 1 as male. We added an image $x \in R^{256\times256\times1}$ comprising 0 or 1 data as the input of the encoder and a scalar of 0 or 1 for gender to the latent variable, which was the input to the decoder.

We optimized MD-VAE on SCUT-FBP5500. Because there were no face evaluation data in celebA, we only used it to optimize VAE. Learning was alternatively realized using these two datasets. Hereinafter, we refer to the latent variable of the source domain as an item vector and the average vector of the target domain data distribution as the user vector.

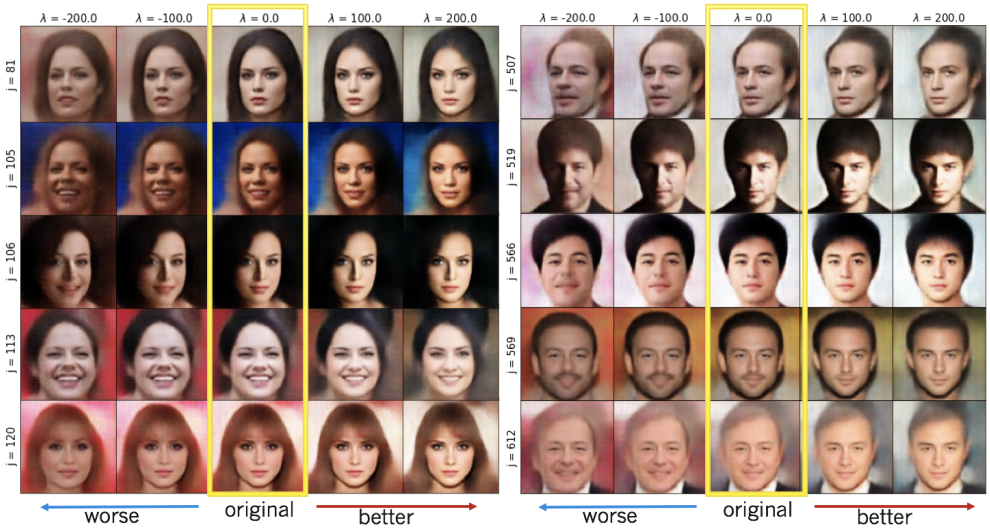


Figure 6: Images decoded latent variable so that it approaches or moves away from average user vector $\bar{\boldsymbol{\mu}}$. Reconstructed images are in the center.

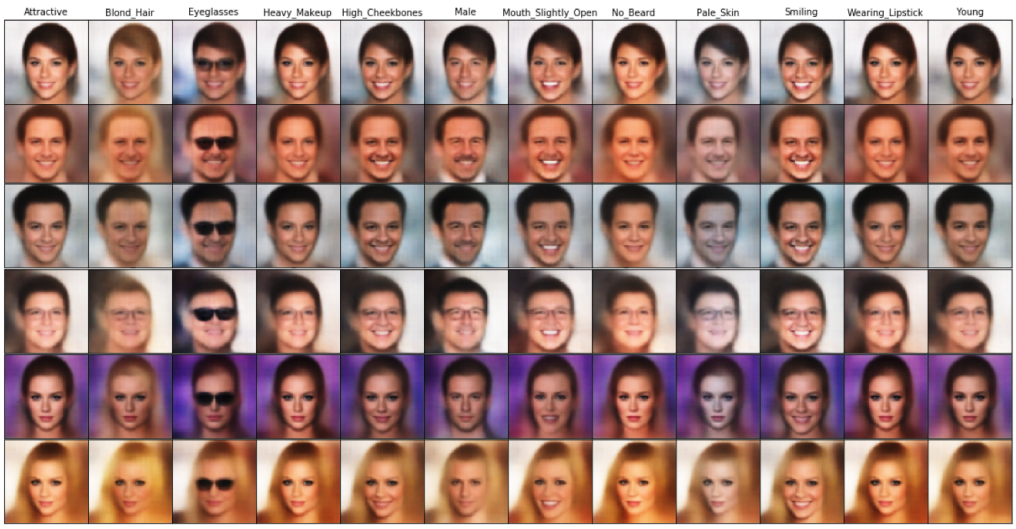


Figure 7: Result of domain transfer using CelebA by MD-VAE

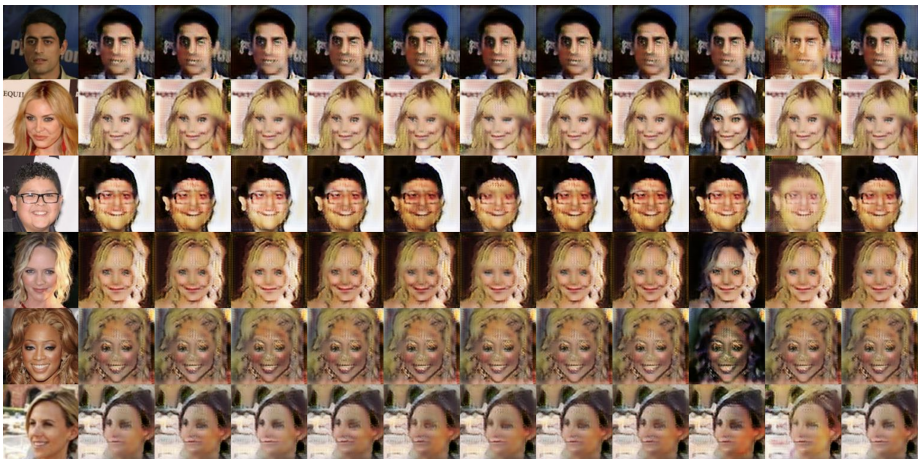


Figure 8: Result of domain transfer using CelebA by star-GAN

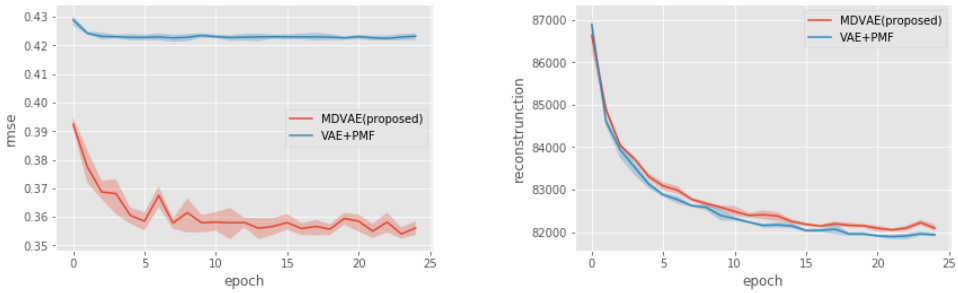


Figure 9: RMSE and Reconstruction loss. MD-VAE is far superior to VAE in classification accuracy, and there is almost no difference in reconstruction error between them.