A Multi-domain and Multi-modal Representation Disentangler for Cross-Domain Image Manipulation and Classification

Fu-En Yang*, Jing-Cheng Chang*, *Student Member, IEEE,*
Chung-Chi Tsai, and Yu-Chiang Frank Wang, *Member, IEEE*

***Abstract*—Learning interpretable data representation has been an active research topic in deep learning and computer vision. While representation disentanglement is an effective technique for addressing this task, existing works cannot easily handle the problems in which manipulating and recognizing data across multiple domains are desirable. In this paper, we present a unified network architecture of Multi-domain and Multi-modal Representation Disentangler (M^2RD), with the goal of learning domain-invariant content representation with the associated domain-specific representation observed. By advancing adversarial learning and disentanglement techniques, the proposed model is able to perform continuous image manipulation across data domains with multiple modalities. More importantly, the resulting domain-invariant feature representation can be applied for unsupervised domain adaptation. Finally, our quantitative and qualitative results would confirm the effectiveness and robustness of the proposed model over state-of-the-art methods on the above tasks.**

***Index Terms*—Representation disentanglement, image translation, domain adaptation, deep learning**

I. Introduction

Recent advances in deep learning have shown promising progresses in the areas of computer vision and machine learning. In particular, visual analysis and synthesis across data domains attract the attention from researchers in these fields. For example, style transfer [1], [2], [3], [4], [5], image-to-image translation [6], [7], [8], [9], [10], and cross-domain visual classification (or domain adaptation) [11], [12], [13], [14], [15] can all be viewed as the associated applications.

To address the above tasks, previous works typically either learn a deterministic (i.e., unimodal) mapping from one data domain to another, or to embed desirable information into the resulting latent space to derive the data representation. The technique of representation disentanglement [16], [17], [18] particularly observes and manipulates specific feature attributes of interest, which has also been applied in the above tasks. Thus, one can view the attributes of interest as the meaningful factors inherent in image data, and further synthesize preferable outputs accordingly. For instance, one can manipulate the *style* attributes of the disentangled latent feature to achieve style transfer and image-to-image translation (e.g., photo ↔ sketch [19]).

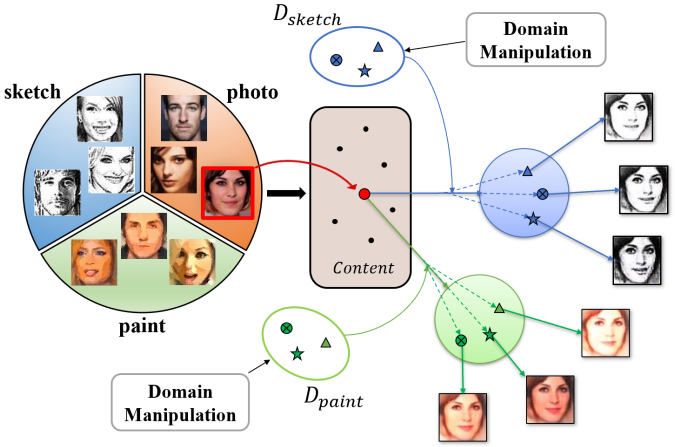


Fig. 1. Illustration of multi-domain and multi-modal representation disentanglement. Given an input (in red bounding box) and images in multiple domains (e.g., styles), we derive representations for describing domain-invariant and domain-specific information, while images can be manipulated and recovered in different domains with sufficient diversity. Note that D_{sketch} and D_{paint} denote domain-specific spaces for *sketch* and *paint* images, respectively.

In practice, adaptation or translation between data domains needs to exhibit multi-modal diversity. That is, a single input instance may correspond to diverse possible outputs, associated with the same attribute of interest (e.g., image style). Even with the promising models based on generative adversarial networks (GANs) [20], one might encounter mode collapse problems and fail to produce multi-modal outputs. Recently, MUNIT [21] and DRIT [22] utilize the disentangled representation for multi-modal translation, achieved by decomposing the latent feature into disjoint features to describe content and style information. While these models manipulate the style feature to synthesize diverse outputs, they cannot be easily extended to handle the image manipulation among multiple (i.e., more than two) domains due to their network architecture designs.

In this paper, we propose a unified framework of *Multi-domain and Multi-modal Representation Disentangler* (M^2RD) for cross-domain image synthesis and classification,

Fu-En Yang is with the Graduate Institute of Communication Engineering, National Taiwan University, Taiwan; e-mail: r07942077@ntu.edu.tw.
Jing-Cheng Chang is with the Graduate Institute of Communication Engineering, National Taiwan University, Taiwan; e-mail: b04901138@ntu.edu.tw.
Chung-Chi Tsai is with Qualcomm Technologies, Inc, San Diego, CA, USA.; email: chuntsai@qti.qualcomm.com
Yu-Chiang Frank Wang is with the Graduate Institute of Communication Engineering, National Taiwan University, Department of Electrical Engineering, National Taiwan University, MOST Joint Research Center for AI Technology and All Vista Healthcare, and ASUS Intelligent Cloud Services, Taiwan; e-mail: ycwang@ntu.edu.tw.
* Indicates equal contribution.

	Unpaired data	Bidirectional translation	Shared representation	Feature disentanglement	Unified structure	Multiple domains	Multi-modal translation	Unsupervised domain adaptation
Pix2Pix [6]	-	-	-	-	-	-	-	-
CycleGAN [8]	√	√	-	-	-	-	-	-
StarGAN [23]	√	√	-	-	-	-	-	-
DTN [7]	√	-	√	-	-	-	-	√
CyCADA [15]	√	√	-	-	-	-	-	√
UNIT [10]	√	√	√	-	-	-	-	√
E-CDRD [19]	√	√	√	√	-	√	-	√
BicycleGAN [24]	-	√	√	√	-	-	√	-
CDDN [25]	-	√	√	√	-	-	√	-
MUNIT [21]	√	√	√	√	-	-	√	-
DRIT [22]	√	√	√	√	-	-	√	√
UFDN [26]	√	√	√	√	√	√	-	√
M^{2}RD (Ours)	√	√	√	√	√	√	√	√

TABLE I
COMPARISONS WITH RECENT WORKS ON IMAGE TRANSLATION AND IMAGE MANIPULATION.

with the ability to manipulate image data with particular attribute of interest while exhibiting sufficient diversity, as illustrated in Fig. 1. Without collecting pairwise image data across domains, our model encodes image data into a domain-invariant and specific latent feature spaces. While the former observes content information from the input data, the latter exhibits multi-modal diversity during cross-domain image translation. In the experiments, we not only show that our model is able to perform image manipulation, but we further verify that derived domain-invariant content features can be applied to the task of unsupervised domain adaptation. With both qualitative and quantitative results provided, the effectiveness and robustness of our model can be successfully confirmed.

We now highlight the contributions as follows:

- Our proposed deep learning model is able to factorize latent image representations into disjoint features describing domain-invariant and specific information.
- Our network uniquely integrates adversarial learning, representation disentanglement, and generative modules in a unified architecture.
- Our derived domain-invariant feature representation allows unsupervised domain adaptation, while the domain-specific feature enables multi-modal image manipulation across multiple data domains.

II. RELATED WORKS

Representation Disentanglement. Aims at learning interpretable data representations ([16], [17], [18], [27], [28], [29], [30], [31]), Chen *et al.* [16] proposed InfoGAN to maximize the mutual information between the latent features and generated images, which realizes representation disentanglement in an unsupervised way. Similarly, Higgins *et al.* [17] introduced β-VAE which derives such representations by adding an adjustable hyperparameter to a variational auto-encoder (VAE) [32], balancing the latent channel capacity and the independence constraints. Tulyakov *et al.* [27] presented MoCoGAN to learn motion and content decomposition for video generation. Although the above methods realize representation disentanglement without label supervision, one cannot manipulate the latent factors directly since the semantic meanings behind the disentangled factors cannot be explicitly obtained. Thus, Odena *et al.* [18] augmented GANs with an auxiliary classifier, allowing image outputs to be conditioned on the desirable latent factors. Furthermore, Peng *et al.* [30] applied reconstruction-based disentanglement and self-supervision to guarantee completely decoupling of latent factors, which benefits pose-invariant face recognition. Tran *et al.* [28], and Liu *et al.* [29] proposed DR-GAN, and MTAN, which derived pose-invariant feature via disentanglement technique and adversarial learning to facilitate the face recognition. Tian *et al.* [31] employed GAN and cycle-consistency for disentangling latent features in multi-view image manipulation. Despite significant progresses, most existing works only focus on producing such representations from a single data domain.

Image-to-Image Translation. To convert images from one style to another, Isola *et al.* [6] chose to observe pairs of images for learning GAN-based models. Taigman *et al.* [7] presented Domain Transfer Network (DTN) to performed such tasks by employing feature consistency across domains. Without observing cross-domain image pairs, Zhu *et al.* [8] learned the bidirectional domain mappings in pixel space with a cycle consistency loss; similar ideas were also applied by [33] and [34]. Coupled GAN (CoGAN) [9] binds high-level information between two data domains for learning the joint distribution. UNIT [10] is extended from CoGAN, which integrates VAE and GAN to achieve image translation by mapping the data between two domains to the same latent space. While the above methods produce promising results, they cannot provide diverse outputs due to their model designs or issues like model collapse.

For multi-modal translation, Zhu *et al.* [24] observed pairs of images for deriving bijection mapping between the latent and output spaces. Gonzalez-Garcia *et al.* [25] decomposed the paired inputs into disjoint shared and exclusive parts to perform diverse image-to-image translation between two domains. Recently, Huang *et al.* [21] and Lee *et al.* [22] concurrently proposed MUNIT and DRIT respectively.

MUNIT and DRIT both factorize the latent representations into domain-invariant content feature and domain-specific style feature from unpaired data. However, their model designs limit the use of data across multiple domains.

Cross-Domain Image Manipulation. In addition to image-to-image translation, several recent works [19], [23], [35], [26] further address image synthesis tasks with the ability of manipulating the attributes of interest. For example, Liu *et al.* [19] considered cross-domain disentangled representation with supervision from single-domain data which aims to manipulate the desirable attributes across different domains. However, they can only deal with a pair of data domains using the proposed model. To handle such tasks across multiple domains, Choi *et al.* [23], He *et al.* [35], and Liu *et al.* [26] proposed StarGAN, AttGAN, and UFDN respectively, which all perform multi-domain image-to-image translation by manipulating the domain label directly. Although StarGAN allows training of multiple domains simultaneously by the unified model structure, it does not exhibit ability in disentangling desirable latent representation. Nevertheless, while the above models are able to manipulate face images, our model further allows one to perform image-to-image translation on a variety of images including images of faces and natrual sccenes. Most importantly, all of them cannot allow multi-modal outputs, which might not be desirable for practical uses.

Unsupervised Domain Adaptation (UDA). Domain adaptation [36], [37], [11], [12], [13], [15] addresses the same learning tasks across domains, with the goal of eliminating the domain shift (i.e., dataset bias). And, unsupervised domain adaptation (UDA) specifically deals with the scenario in which no label supervision is available during training in the target domain. For instance, GAKT [36] applied adaptive graph to transfer discriminative information from labeled source to unlabeled target domain. Also, Ding *et al.* [37] integrated low-rank coding with deep neural network for preserving global structures across source and target, to achieve more effective knowledge transfer. Recently, several GAN-based methods have been proposed for UDA. For example, Ganin *et al.* [11] introduced a Domain Adversarial Neural Network (DANN) framework which contains a domain classifier with its gradient reversal layer serving as a domain-invariant feature extractor. Tzeng *et al.* [12] adapted feature extractors and classifier of source and target domains by domain adversarial learning strategies to tackle UDA. Bousmalis *et al.* [13] utilized the decomposed representations to produce domain-invariant features to facilitate cross-domain classification. Hoffman *et al.* [15] further extended CycleGAN [8] and applied adversarial learning and cycle-consistency for both feature and pixel-level adaptation.

Nevertheless, the above models typically do not exhibit abilities in disentangling particular image attributes, nor to manipulate image outputs across domains with multi-modal diversity. In Table I, we compare our proposed model with recent deep learning methods in the aforementioned topics.

III. Multi-domain and Multi-modal Representation Disentangler (M^{2}RD)

A. Notation and Model Overview

Given an image set $\{\mathcal{X}_i\}_{i=1}^N$ across N distinct domains, our M^{2}RD jointly learns a domain-invariant content feature $\{z_i^c\}_{i=1}^N$ and domain-specific feature $\{z_i^d\}_{i=1}^N$ from the input image $x_i \in \mathcal{X}_i$, and then utilize discrete domain code $\{l_i\}_{i=1}^N$ to further exploit the domain information in the latent space. We note that the domain code l_i can be implemented by an one-hot vector, a real-value vector, or even concatenation of multiple one-hot vectors, which describes the domain of interest.

As illustrated in Fig. 2, our framework consists of two network modules. First, we have a representation disentangler with a content discriminator. This module contains a *content encoder* E_c and a *domain encoder* E_d, which are shared by input data across different domains. By advancing adversarial learning strategies, this disentangler module allows us to derive domain-invariant and specific features. The former provides the content information of the input data disregard of its domain of origin, while the latter describe the domain of interest, which allows multi-modal manipulation as described later.

On the other hand, we have a Multi-domain and Multi-modal Generative Adversarial Networks as the second network module in Fig. 2, which includes a generator G and a domain discriminator D_{dom}, while the same content encoder E_c is deployed to observe *content consistency*. With the observed domain-invariant content feature z^c, this module performs both multi-domain and multi-modal image translation by manipulating the derived domain-specific feature z^d and the domain code l. The details of our proposed network will be discussed in the following subsections.

B. Representation Disentangler

As illustrated in Fig. 2, our proposed network encodes cross-domain image inputs using shared content encoder E_c and domain encoder E_d. To enable the encoded content features to be domain-invariant, we apply a content discriminator D_c to eliminate the domain differences between the resulting features inspired by [11]. In other words, we have D_c aim to correctly produce domain code prediction $\hat{l}$ from the encoded content features z_i^c. Thus, the objective function of this content discriminator $\mathcal{L}_{adv}^{D_c}$ is derived as follows:

$$\mathcal{L}_{adv}^{D_c} = \mathbb{E}[\log(P(\hat{l} = l_i | E_c(x_i)))], \tag{1}$$

where P is the probability distribution over domains $\hat{l}$, which is calculated by the content discriminator D_c.

With the above design, our content encoder E_c would be able to extract the domain-invariant content features from input data, which are observed across multiple domains. As a result, the objective function of the encoder E_c is to maximize the cross-entropy of the content discriminator:

$$\mathcal{L}_{adv}^{E_c} = -\mathcal{L}_{adv}^{D_c} = -\mathbb{E}[\log(P(\hat{l} = l_i | E_c(x_i)))]. \tag{2}$$

Finally, in order to learn a joint and continuous representation for cross-domain data, and further perform stochastic sampling in testing phase, we enforce the *Kullback-Leibler* divergence for our generative network model. This encourages

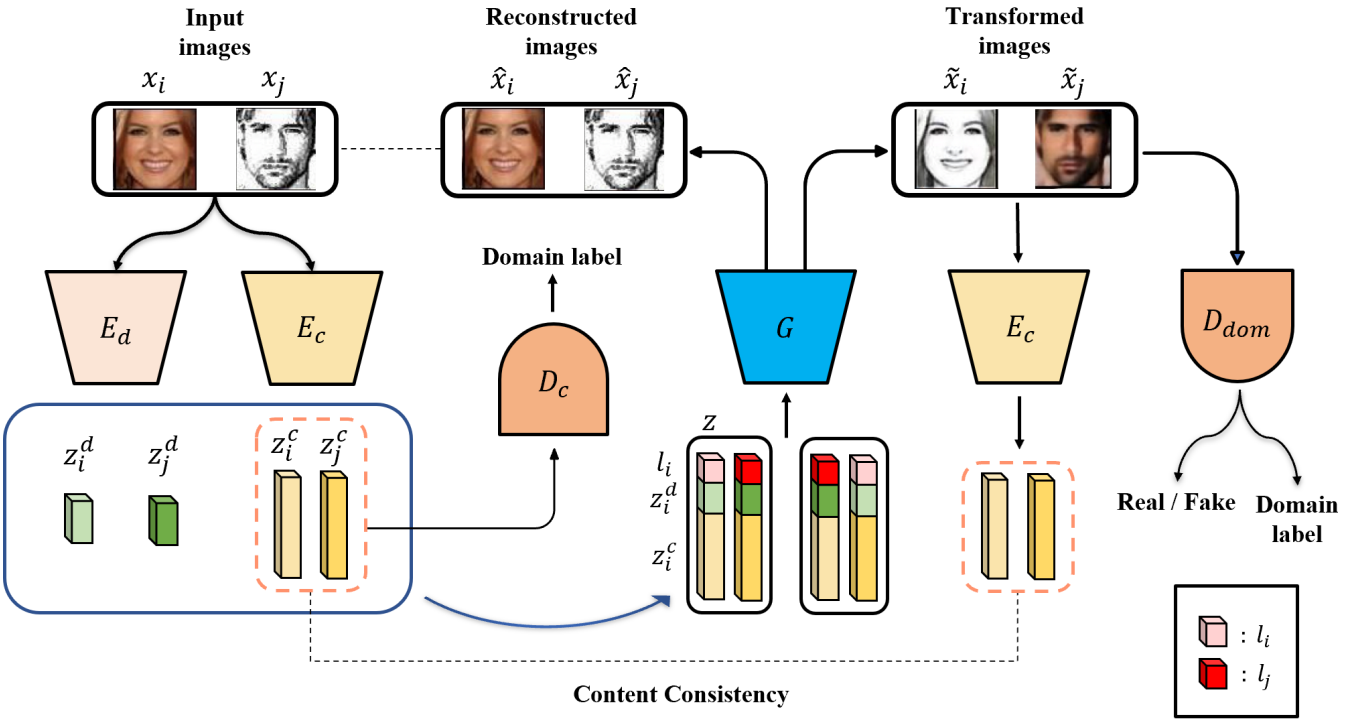


Fig. 2. The network architecture of our Multi-domain and Multi-modal Representation Disentangler (M^{2}RD), which consists of two modules: 1) Representation disentangler, composed of a content encoder E_c, a domain encoder E_d, and a content discriminator D_c, and 2) Multi-domain and Multi-modal GAN consisting of a generator G, and a domain discriminator D_{dom} with an auxiliary domain classifier. Note that z^c, z^d denote the domain-invariant and specific features extracted from different domains respectively. Together with a domain code l, the final feature representation $z = [z^c, z^d, l]$ can be utilized for cross-domain and multi-modal image manipulation.

the domain-specific feature z^d to fit a prior Gaussian distribution $N(0, I)$. Thus, the objective $\mathcal{L}_{KL}$ is calculated as:

$$\mathcal{L}_{KL} = \mathbb{E}[KL(E_d(x_i)||N(0, I))] \tag{3}$$

We note that, derivation of the above domain-invariant content representation is the reason why we can apply such features for unsupervised domain adaptation (UDA), which desires a common feature representation shared by different domains for adaptation purposes. With the above network design, we enforce the derived content features z^c does not contain any domain information, and thus the domain shift can be properly suppressed. As a result, we can simply deploy an extra classifier based on z^c if the UDA is of interest. To be more precise, the objective $\mathcal{L}_{cla}$ for this added UDA classifier can be expressed as follow:

$$\mathcal{L}_{cla} = -\sum_{k=1}^{N_{src}} y_k^{src} \cdot \log \tilde{y}_k^{src}. \tag{4}$$

where $\tilde{y}_k^{src}$ is the predicted output from the k-th labeled source input, and y_k^{src} is the ground truth label.

C. Multi-domain and Multi-modal GAN

Once the domain-invariant feature z^c and the domain-specific ones z^d are observed, the second module in our proposed architecture performs multi-domain and multi-modal image translation (i.e., cross-domain image manipulation with multi-modal diversity). We now discuss how these two tasks are jointly performed.

Similar to most existing image translation works, we perform image synthesis by combining the derived content feature z^c and with the domain feature z^d. Extended from AC-GAN [18], we additionally assign the domain code l into the above feature representation to form the final feature representation $z = [z^c, z^d, l]$, followed by the decoding process.

Recall that, the representation z^d is learned to describe domain-only information, while such representation is shared by cross-domain data inputs. Thus, with the sampling strategies noted in Section III-B, we will be able to reconstruct the image output and exhibit multi-modal diversity. In other words, *within-domain variants* of the recovered output associated with the same content feature z^c can be produced via sampling z^d. And, the above domain code l is added to ensure that the output image is recovered at or translated into the domain of interest. This is how our proposed model differs from existing image translation or disentanglement works.

With the above explanations, we now define the object functions applied in this network module. First, for image recovery guarantees, we calculate the reconstruction loss $\mathcal{L}_{rec}$ for the reconstructed image $\hat{x}_i$:

$$\mathcal{L}_{rec} = ||x_i - \hat{x}_i||_1, \tag{5}$$

Note that x_i is the (ground truth) input, and $\hat{x}_i = G([z_i^c, z_i^d, l_i])$.

Inspired by DTN [7], we further preserve the content consistency between translated images $\tilde{x}_i$ and input image x_i. Thus, an objective function $\mathcal{L}_{con}$ based on the same content encoder E_c is introduced in the feature level, which can be formulated as:

$$\mathcal{L}_{con} = ||E_c(x_i) - E_c(\tilde{x}_i)||_2, \tag{6}$$

where $\tilde{x}_i = G([z_i^c, z_j^d, l_j]), \ \ i \neq j$.

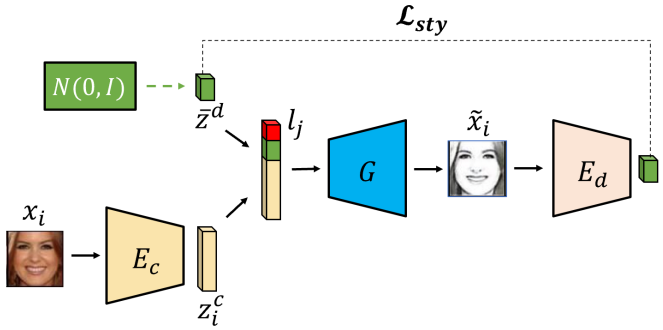


Fig. 3. In addition to the architecture described in Fig. 2, we further apply the objective function $\mathcal{L}_{sty}$ to enforce the reconstruction on the domain-specific feature. More details can be found in Section III-C

Also, similar to DRIT [22], we utilize style regression loss to enforce the reconstruction on the domain-specific feature, as illustrated in Fig. 3, with the objective $\mathcal{L}_{sty}$ expressed as:

$$\mathcal{L}_{sty} = ||E_d(G([z_i^c, \bar{z}^d, l_j]))) - \bar{z}^d||_2, \tag{7}$$

where $\bar{z}^d$ is sampled from a prior Gaussian distribution $N(0, I)$.

However, when manipulating images across domains using the above network module, there is no guarantee that the output image $\tilde{x}_i$ would properly satisfy the domain information based on the domain code l inserted. Thus, as a part of the AC-GAN extension, we deploy a domain discriminator D_{dom} in Fig. 2 which perform multi-task learning for combining adversarial learning with an auxiliary domain classification task.

To be more precise, this discriminator not only determines the authenticity of the output images, it also classify its domain label output to enforce the ability of the introduced domain code for domain disentanglement. Thus, the objective functions of this domain discriminator D_{dom} and generator G are calculated as follows:

$$\mathcal{L}_{adv}^{D_{dom}} = \mathbb{E}[\log(D_{dom}(\tilde{x}_i))] + \mathbb{E}[\log(1 - D_{dom}(x_i))], \tag{8}$$

$$\mathcal{L}_{aux}^{D_{dom}} = \mathbb{E}[\log(P(\bar{l} = l_j|\tilde{x}_i))] + \mathbb{E}[\log(P(\bar{l} = l_i|x_i))], \tag{9}$$

$$\mathcal{L}_{adv}^{G} = -\mathbb{E}[\log(D_{dom}(\tilde{x}_i))], \tag{10}$$

where $\bar{l}$ denotes the prediction output of D_{dom}. We note that the objective $\mathcal{L}_{aux}^{D_{dom}}$ aims at maximizing the mutual information between the domain code and the translated image [16].

D. Full Objectives

In summary, the full objective function $\mathcal{L}$ of our model can be summarized below:

$$\begin{aligned}\mathcal{L} = &\lambda_1\mathcal{L}_{adv}^{D_c} + \lambda_1\mathcal{L}_{adv}^{E_c} \\ &+ \lambda_2(\mathcal{L}_{adv}^{D_{dom}} + \mathcal{L}_{aux}^{D_{dom}}) + \lambda_2\mathcal{L}_{adv}^{G} \\ &+ \lambda_{rec}\mathcal{L}_{rec} + \lambda_{con}\mathcal{L}_{con} \\ &+ \lambda_{KL}\mathcal{L}_{KL} + \lambda_{sty}\mathcal{L}_{sty} + \lambda_{cla}\mathcal{L}_{cla},\end{aligned} \tag{11}$$

where the hyperparameters λ regularize each loss term. Nevertheless, we fix the values of λ for each dataset, and do not fine-tune them for each input instance.

To train our model, we alternatively update content encoder E_c, domain encoder E_d, generator G, content discriminator D_c, and domain discriminator D_{dom} via the following gradients:

$$\begin{aligned}\theta_{E_c} &\xleftarrow{+} -\Delta_{\theta_{E_c}}(\mathcal{L}_{rec} + \mathcal{L}_{adv}^{E_c} + \mathcal{L}_{con} + \mathcal{L}_{sty}) \\ \theta_{E_d} &\xleftarrow{+} -\Delta_{\theta_{E_d}}(\mathcal{L}_{rec} + \mathcal{L}_{KL}) \\ \theta_{G} &\xleftarrow{+} -\Delta_{\theta_{G}}(\mathcal{L}_{rec} + \mathcal{L}_{con} + \mathcal{L}_{KL} + \mathcal{L}_{adv}^{G} + \mathcal{L}_{aux}^{D_{dom}} + \mathcal{L}_{sty}) \\ \theta_{D_c} &\xleftarrow{+} -\Delta_{\theta_{D_c}}(\mathcal{L}_{adv}^{D_c}) \\ \theta_{D_{dom}} &\xleftarrow{+} -\Delta_{\theta_{D_{com}}}(\mathcal{L}_{adv}^{D_{dom}} + \mathcal{L}_{aux}^{D_{dom}}).\end{aligned} \tag{12}$$

We note that, if UDA is of interest, an additional classifier (as discussed in Section III-B) will be added with loss $\mathcal{L}_{cla}$. Thus, the gradient of θ_{E_c} is derived as follows:

$$\theta_{E_c} \xleftarrow{+} -\Delta_{\theta_{E_c}}(\mathcal{L}_{rec} + \mathcal{L}_{adv}^{E_c} + \mathcal{L}_{con} + \mathcal{L}_{sty} + \mathcal{L}_{cla}). \tag{13}$$

Once the training is complete, our model can be applied to image translation in the following ways:

1) For an input image, we utilize the content encoder E_c to extract its content feature. By conditioning on a randomly sampled domain-specific feature with a selected domain code l_i, generator G would manipulate and output the image in the domain of interest.

2) Give two images of interest, we extract the content feature z_i^c from one image, and the domain-specific feature z_j^d from another (together with its domain code l_j). This can be viewed as example-guided image translation.

It is worth noting that, our disentangled representations are achieved by jointly minimizing domain confusion loss ($\mathcal{L}_{adv}^{D_c}$, $\mathcal{L}_{adv}^{E_c}$), reconstruction loss $\mathcal{L}_{rec}$, content consistency loss $\mathcal{L}_{con}$, and style regression loss $\mathcal{L}_{sty}$. Specifically, we explicitly derive the domain-invariant content feature from input images via domain confusion loss ($\mathcal{L}_{adv}^{D_c}$, $\mathcal{L}_{adv}^{E_c}$) in an adversarial manner, allowing our content encoder E_c to extract domain-invariant features. Moreover, we have the content and style consistency losses ($\mathcal{L}_{con}$ and $\mathcal{L}_{sty}$) deployed in our architecture; the former ensures that the input and the translated images preserve the same content feature representation, while the latter enforces the transformed output to be of the style of interest. Finally, the reconstruction loss $\mathcal{L}_{rec}$ is applied to jointly observe the aforementioned disentangled representation with data recovery guarantees. In Table III, we have ablation studies to support the design of our proposed network in performing representation disentanglement.

Also, as shown in Fig. 4, 6, 9, we show that our proposed model is able to derive disentangled representations from input images of the *seen* domains and producing diverse outputs in the *seen* domain of interest during inference time. We note that, existing state-of-the-art image translation models via representation disentanglement (e.g., UNIT [10], E-CDRD [19], MUNIT [21], DRIT [22], UFDN [26]) cannot generalize to images in *unseen* domains. This also verifies that our model exhibit excellent abilities in decoupling content and style-dependent features for image translation.

E. Comparisons to Recent Models

It is correct that, while our model is related to a recent multi-domain image translation method of UFDN [26], and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2952707, IEEE Transactions on Image Processing

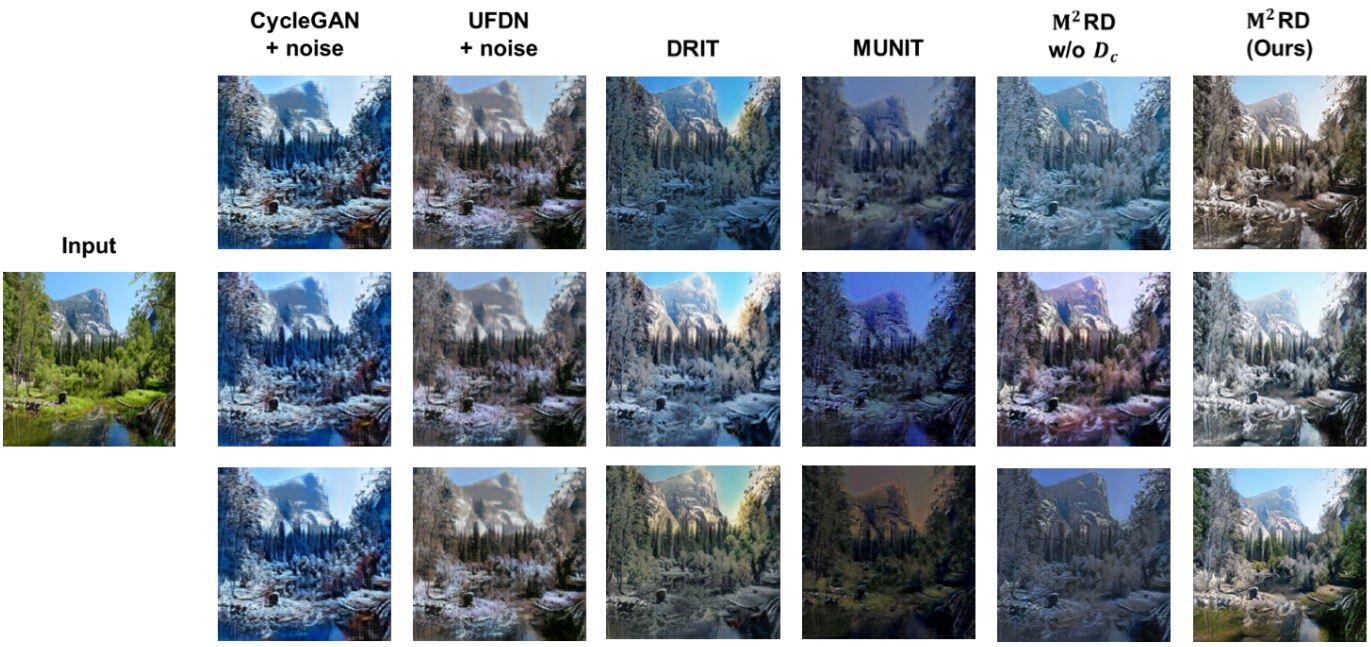


Fig. 4. Example results of our multi-modal image translations and the comparison with the existing image-to-image translation methods. We observe that our model is able to generate high-quality images with meaningful diversity.

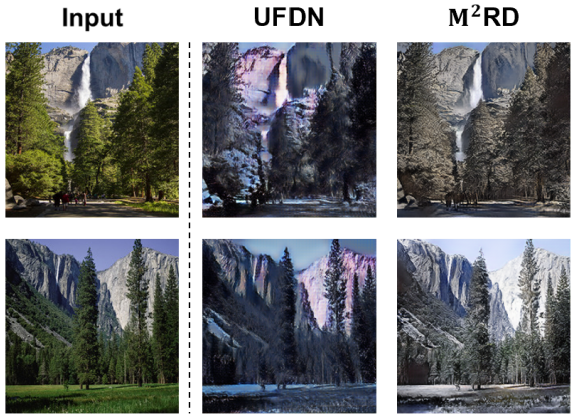


Fig. 5. Example results of the comparison with UFDN [26] in summer-to-winter translation. Note that since UFDN [26] does not observe and exhibit intra-domain image variety, its output might be irrational in terms of appearance or lighting (e.g., mix of daytime and nighttime appearance), while ours are more realistic and have a higher visual quality.

a number of network modules are shared by this work and ours, multi-modality is the major highlight of our work, plus the introduced feature-level consistency to improve the output image quality. As we noted in Table I, our M^{2}RD further exhibits multi-modal property during the translation/synthesis process, which cannot be achieved by UFDN [26]. However, such extension is not trivial. First, our M^{2}RD needs to derive disjoint domain-specific features (z^d) from the domain-invariant features z^c at the output of the domain encoder (E_d). With detailed model and loss designed are described in our work, we then fit such disentangled domain-specific features to Gaussian distribution priors, allowing the learning of multi-modality in image translation. As discussed, the domain code (l) in our model serves as supervision, which guides our unified generator to synthesize the output image in the domain of interest. In contrast, UFDN [26] can only perform one-to-one image translation without diversity. In Fig. 4 and Table II, we present qualitative and quantitative comparisons respectively to confirm the capability of our M^{2}RD to translate images across multiple domains with sufficient diversity.

Second, we consider to exploit both *inter*-domain and *intra*-domain variation during image translation, while UFDN [26] only observes inter-domain variation. As shown and compared in Table II, the lack of the ability in modeling intra-domain diversity would lead to a discernible drop in visual quality. Take Fig. 5 for examples, the domain change in seasons would be viewed as *inter*-domain variations, while the day/night lighting, etc. condition changes are modeled as *intra*-domain variations. Without our derivation of domain-specific features z^d, one cannot produce translated image outputs with satisfactory quality, generating winter scenes with irrational or unrealistic lighting conditions (and thus poor user study results, as shown and compared in Table II).

Third, we employ cycle-consistency loss in our model for feature consistency guarantees, while UFDN [26] does not include such constraints and thus suffers from drops in visual quality in performing image translation. To be more precise, we utilize *content* consistency to preserve content information during the generation process, instead of directly applying *pixel-level* consistency as used in DRIT [22]. Throughout our experiments, we observe that adding data recovery constraints over pixel levels would be overly restrictive and limit the diversity of the image outputs. With the above observations and as summarized in Table II, we show that our model achieved higher LPIPS (O2O) score than DRIT [22] did, which supports the effectiveness of our model in preserving content consistency during image translation. With the above remarks,

		MUNIT	DRIT	UFDN	M^{2}RD (Ours)
Realism	User Study (↑)	21.17%	18.17%	19.33%	**41.33%**
	FID (↓)	85.09 ± 0.77	68.44 ± 0.75	87.69 ± 0.70	$\mathbf{57.76 \pm 0.23}$
	LPIPS (I2O) (↓)	0.417 ± 0.003	0.385 ± 0.002	0.758 ± 0.002	$\mathbf{0.339 \pm 0.003}$
Diversity	LPIPS (O2O) (↑)	$\mathbf{0.225 \pm 0.002}$	0.173 ± 0.002	0.040 ± 0.001	0.196 ± 0.003

TABLE II
QUANTITATIVE COMPARISONS FOR VISUAL REALISM AND DIVERSITY WITH MUNIT, DRIT, UFDN, AND OUR M^{2}RD ON SUMMER-TO-WINTER TRANSLATION.

we believe the technical contributions of this work would be sufficiently unique, which makes our work very different from UFDN [26].

IV. EXPERIMENTS[1]

A. Implementation Details

We utilize PyTorch [38] to implement our model and choose ADAM [39] as the optimizer to train our network, with the learning rate, β_1, and β_2 set as 10^{-4}, 0.5, and 0.999, respectively. In our all experiments, we set the hyperparameters as follows: $\lambda_1 = 1$, $\lambda_2 = 1$, $\lambda_{rec} = 10$, $\lambda_{con} = 1$, $\lambda_{KL} = 10^{-3}$, $\lambda_{sty} = 10$, and $\lambda_{cla} = 1$.

More details about the network architecture for Summer ↔ Winter and Photo ↔ Art datasets are described in the following.

For content encoder E_c, we apply convolutional architecture composing of three convolution layers and four residual blocks. For domain encoder E_d, we implement it by utilizing four convolution layers followed by a fully-connected layer. Also, we use four residual blocks, followed by three deconvolution layers to realize generator G. For content discriminator D_c, it consists of five fully-connected layers. For domain discriminator D_{dom}, we utilize the architecture of PatchGANs [6] that contains six convolution layers, and add two convolution layers for outputting real/fake and domain code prediction respectively.

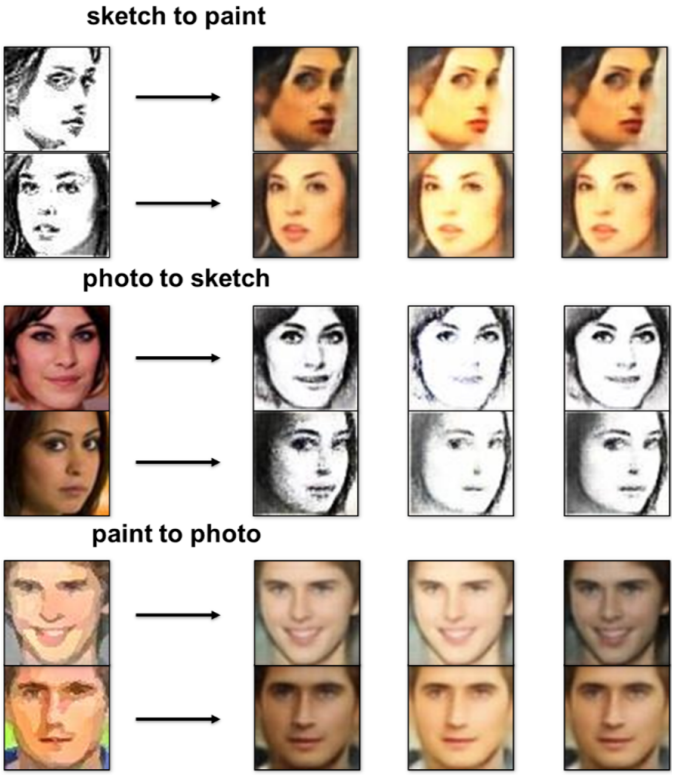


Fig. 6. Example results of multi-modal image translation for *face* images across multiple domains.

B. Datasets

We consider four different categories of image datasets, i.e., digit, face, seasons, and art paint, for performance evaluation:

Digits. The image datasets of *MNIST*, *USPS* and *Street View House Number (SVHN)* are hand-written digit image datasets, which are viewed as images observed in different domains. MNIST contains 60,000/10,000 images for training/testing, and USPS consists of 7,291/2,007 images for training/testing. SVHN is composed of colored digits images with complex background and contains 73,257 training images, 26,032 testing images, and 531,131 extra images. All images are converted to RGB images with the size of 32 × 32 × 3 pixels for our experiments.

Faces. We consider facial *photo*, *sketch*, and *paint* images as data in different domains. For facial photo images, we consider the *CelebFaces Attributes dataset (CelebA)* [40], which is a large-scale face image dataset including more than $200K$ celebrity photos annotated with 40 facial attributes. Following the settings of [6], [19], [26], we randomly transfer half of the photos to sketch, then convert the remaining photos into paint images.

Summer ↔ Winter. The Summer ↔ Winter dataset [8] contains natural scene images categorized into summer or winter. The size of all images is 256 × 256 × 3 pixels, and the numbers of images are 1273 and 854 for summer and winter, respectively.

Photo ↔ Art. We choose the photo from Yosemite [8] and the *Art* dataset [41] which collected from Wikiart containing 14 different artists. We conduct our experiments on Monet, Van Gogh, and Ukiyo-e, and also resize all images into 256 × 256 × 3 pixels.

It is worth noting that, while image data across multiple domains are presented during the training stage, we do not observe any cross-domain image pairs when learning our proposed model. This is different from recent translation models like [6], [24] with such requirements.

[1]The authors from National Taiwan University completed the experiments on the datasets.

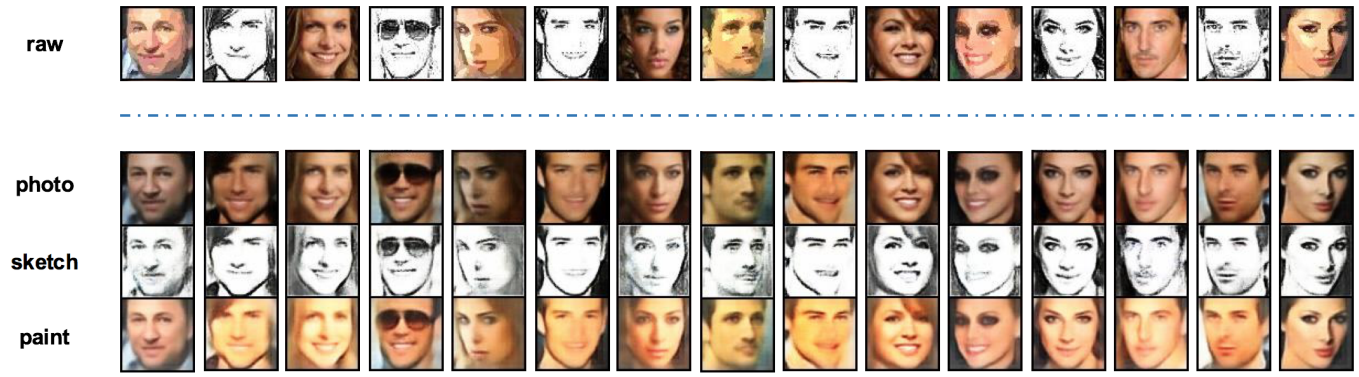


Fig. 7. Example results of image translation across multiple domains among *photo/sketch/paint*.

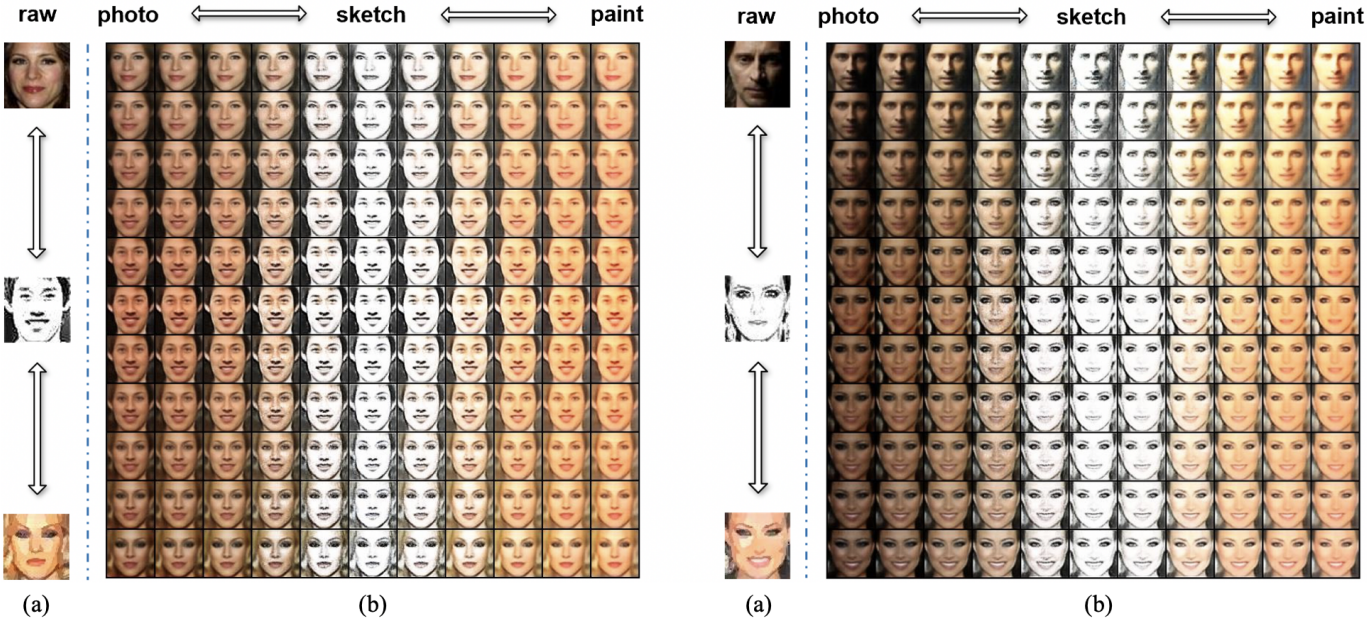


Fig. 8. Example results of our multi-domain image translations and manipulations. (a) Selected images from three different domains. (b) The horizontal axis shows the cross-domain style interpolation for facial *photo/sketch/paint*, while the vertical axis verifies that the domain-invariant content feature space is continuous.

C. Multi-domain and Multi-modal Image Translation and Manipulation

1) Multi-modal image manipulation: In order to provide diversity in the produced image outputs, we first manipulate the latent feature space by sampling the domain-specific feature from a prior Gaussian distribution, concatenated by a desired one-hot domain code. Example results are shown in Fig. 4 on *summer* $\leftrightarrow$ *winter* dataset and Fig. 6 on *face* dataset, in which multiple outputs in each domain can be produced based on the same input image. Specifically, in Fig. 4, we compare our M^{2}RD with the state-of-the-art image-to-image translation methods, showing that our M^{2}RD is capable of synthesizing high-quality output images with diversity. We observe that only injecting noise vectors to the generator of CycleGAN [8], which originally focuses on one-to-one image translation, cannot produce diverse outputs. While UFDN [26] translates images across multiple domains, the generated images mainly belong to one mode and fail to synthesize multi-modal images. Comparing with DRIT [22] and MUNIT [21], DRIT also generates plausible results, and MUNIT produces images with unrealistic style. We also demonstrate that our model without content discriminator (D_c) cannot preserve domain-invariant information well, causing unrealistic and ill-quality results. From the above experiments, the use of our proposed M^{2}RD for multi-modal image translation can be successfully verified.

In addition to qualitative results and comparisons, we further provide additional quantitative comparisons with MUNIT [21], DRIT [22], and UFDN [26], which are known as the state-of-the-art models on image translation.

To assess the visual quality and realism of the synthesized images, we adopt *Frechet Inception Distance* (FID) [42] and *Learned Perceptual Image Patch Similarity* (LPIPS) [43] as the metrics for quantitative evaluation. We compute FID to measure the distance between the generated distribution and the real image input, and we also calculate average Input-to-Output LPIPS, denoted as LPIPS (I2O), to measure the

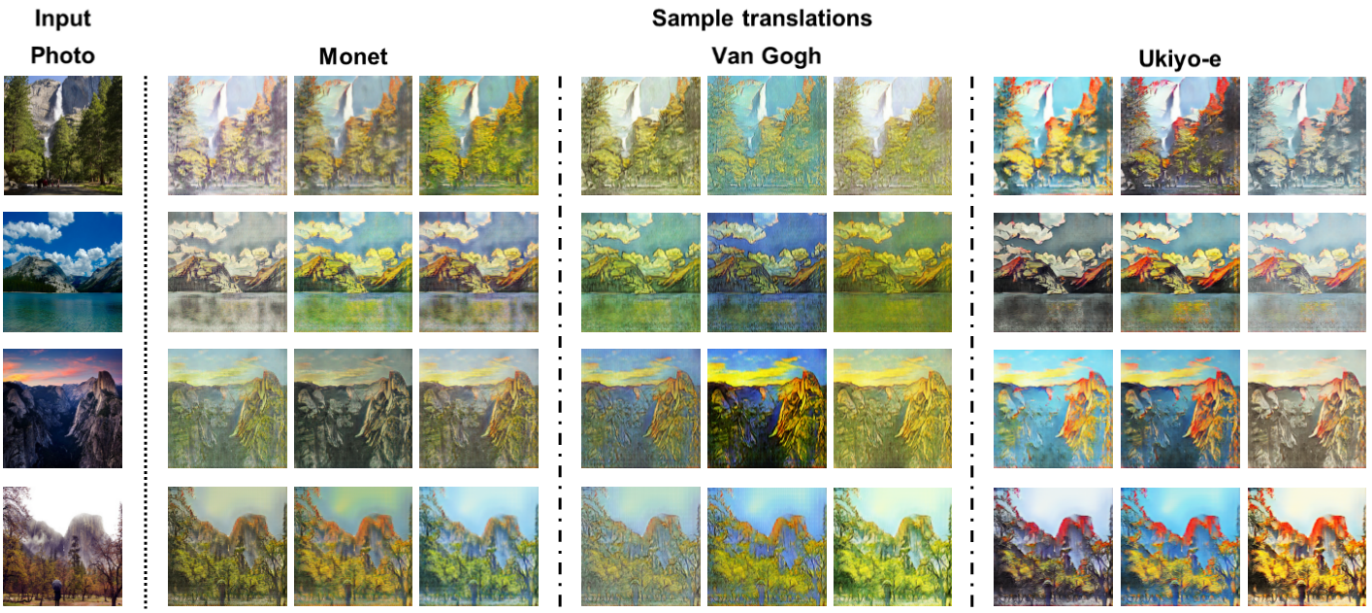


Fig. 9. Example results of our multi-domain and multi-modal image translations. We translate the input photos into another *Paint style* by manipulating the domain code. Further, by randomly sampling distinct noise vectors, we are able to synthesize output images in the domain of interest with multi-modality.

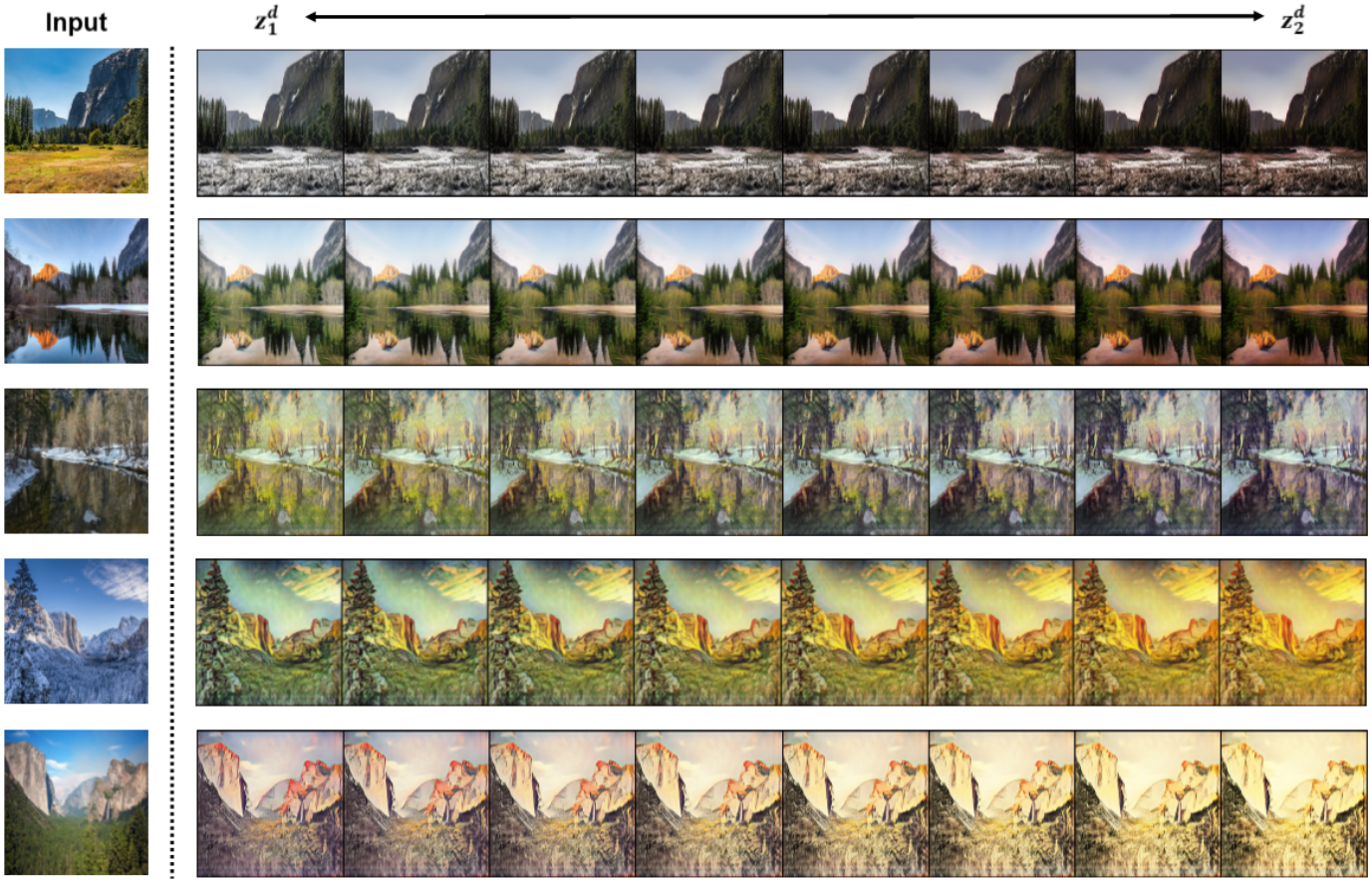


Fig. 10. Example results of linear interpolation between two sampled random vectors both on *Simmer* ↔ *Winter* and *Photo* ↔ *Art* dataset.

distance between the input image and its corresponding translated outputs (note that lower scores indicate outputs with better visual quality). In addition, we conduct studies by asking 30 users with diverse backgrounds and knowledge with 20 questions, each contains a given input image and four translated images generated by the above models (including

		w/o D_c	w/o D_{dom}	w/o $\mathcal{L}_{con}$	w/o $\mathcal{L}_{sty}$	w/o $\mathcal{L}_{KL}$	M^{2}RD (Ours)
Realism	FID (↓)	60.35 ± 0.56	444.48 ± 3.71	73.76 ± 0.97	68.65 ± 0.82	99.24 ± 1.37	$\mathbf{57.76 \pm 0.23}$
	LPIPS (I2O) (↓)	0.354 ± 0.002	0.976 ± 0.003	0.364 ± 0.002	0.347 ± 0.002	0.397 ± 0.003	$\mathbf{0.339 \pm 0.003}$
Diversity	LPIPS (O2O) (↑)	0.136 ± 0.002	0.067 ± 0.004	0.187 ± 0.002	0.107 ± 0.001	0.158 ± 0.002	$\mathbf{0.196 \pm 0.003}$

TABLE III
ABLATION STUDIES ON SUMMER TO WINTER TRANSLATION.

ours), and the user is asked to select the one which he/she feels to be most appropriate/realistic. In Table II, we show that our M^{2}RD outperformed the aforementioned state-of-the-art multi-modal or multi-domain image translation models in all categories. With this experiment, we confirm that our model is capable of producing output images with satisfactory visual quality.

In addition to visual realism, we provide quantitative comparison for visual diversity by calculating average Output-to-Output LPIPS, denoted as LPIPS (O2O), to measure the distance between the outputs translated from the same input image (note that larger distance values represent output images with more diversity). As shown in Table II, we see that despite UFDN [26] is capable of translating images across multiple domains, it cannot achieve *multi-modal* image translation (with the lowest LPIPS score). More importantly, our model was shown to perform favorably against DRIT [22] and MUNIT [21], which support the ability of our model in synthesizing plausible outputs with sufficient multi-modal diversity. With the above quantitative comparisons, the robustness and superiority of our model can be successfully verified.

2) Multi-domain image manipulation: We demonstrate the ability of our model in realizing image translation across multiple domains using *face* dataset. Given images from an arbitrary domain (i.e., top row in Fig. 7), we extract their domain-invariant and domain-specific features, respectively. For translation purposes, we assign and concatenate the above features with different domain codes of interest (e.g., [1, 0, 0] for *photo*, [0, 1, 0] for *sketch*, and [0, 0, 1] for *paint*) for image reconstruction. The translated results were shown in each corresponding column in the bottom row of Fig. 7.

Then, given images from different domains (i.e., *photo*, *sketch*, and *paint* in Fig. 8a), we extract their domain-invariant (content) features and domain-specific (style) features. Then, we perform feature interpolation within the same feature type. Using the resulting content/style features with an interpolated domain code, we are able to produce cross-domain image translation outputs. As shown in Fig. 8b, outputs in vertical and horizontal axes represent image variants in (domain-invariant) content and (domain-specific) style with the associated domain code, respectively. Observing the diagonal entries of Fig. 8b, which shows the extremely translation case, and fully exhibit the effectiveness and robustness in the derived feature representations for multi-domain image manipulation.

In addition to faces, we also demonstrate the use of our model for manipulating hand-written digit images. As shown in Fig. 11a and b, by manipulating the domain-specific feature with the desirable domain code (e.g., [1, 0] for USPS/SVHN, and [0, 1] for MNIST), our model are able to convert the USPS and SVHN images into MNIST ones. The above experiments of the use of our proposed M^{2}RD for multi-domain image manipulation are supportive.

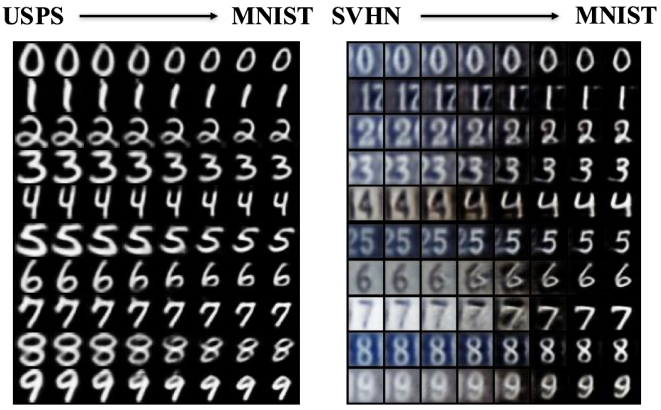


Fig. 11. Cross-domain continuous image manipulation for (a) USPS → MNIST and (b) SVHN → MNIST.

3) Multi-modal translation across multiple domains: As shown in Fig. 9, we conduct the experiment of multi-modal image translation across multiple domains on *Photo* ↔ *Art* dataset. By manipulating the domain code (e.g., [0, 0, 0, 1] for *Photo*, [0, 0, 1, 0] for *Monet*, [0 , 1, 0, 0] for *Van Gogh*, and [1, 0, 0, 0] for *Ukiyo-e*), our M^{2}RD is capable of translating given images to the domain of interest. We show that our model successfully captures different *Painting style* and presents clearly distinct results. Furthermore, by randomly sample different noise vectors from prior Gaussian distribution, we are able to model the intra-domain variation and perform multi-modal diversity.

For further evaluate the domain-specific (style) latent space derived by M^{2}RD, we perform linear interpolation between two sampled style feature as shown in Fig. 10. The corresponding results both on *Summer* ↔ *Winter* and *Photo* ↔ *Art* dataset change smoothly and continuously along with the variations of style latent feature.

4) Quantitative Ablation Study: In addition to the qualitative ablation study (i.e., Fig. 4) which partially performs such ablation studies (i.e., our model with and without D_c), we now present additional quantitative ablation studies in Table III to verify the technical contributions of our work.

As shown in Table III, our model surpassed others in terms

	MNIST → USPS	USPS → MNIST	SVHN → MNIST
DANN [11]	-	-	73.85
Associative DA [14]	-	-	93.71
DSN [13]	-	-	82.70
DTN [7]	-	-	84.88
PixelDA [44]	-	95.9	-
DRCN [45]	91.80	73.70	82.00
CoGAN [9]	95.65	93.15	-
ADDA [46]	89.40	90.10	76.00
UNIT [10]	95.97	93.58	90.53
CyCADA [15]	-	-	90.08
ADGAN [47]	92.80	90.80	92.40
CDRD [19]	95.05	94.35	-
SBADA-GAN [48]	97.6	95.0	76.1
UFDN [26]	97.13	93.77	**95.01**
M^{2}RD (Ours)	**98.54**	**98.49**	94.03

TABLE IV
A CLASSIFICATION ACCURACY (%) FOR TARGET DOMAIN IMAGES. FOR EXAMPLE, USPS → MNIST DENOTES USPS AND MNIST AS SOURCE AND TARGET DOMAIN IMAGES, RESPECTIVELY.

of all metrics of FID and LPIPS scores, which confirms the visual quality and diversity achieved by the full model of our M^{2}RD. We observe that, without content discriminator D_c, all scores became inferior since the derived features from content encoder E_c will not be domain-invariant and would carry the domain-specific information, even with the presence of domain-specific feature z^d and domain code l. This supports our network/loss designs for representation disentanglement. Moreover, without domain discriminator D_{dom}, all scores were degenerate significantly due to image details of the outputs across different domains cannot be properly preserved. Next, when the content consistency loss $\mathcal{L}_{con}$ was disabled, the content information would not be preserved well, resulting in poor visual quality and inferior FID/LPIPS scores. If the style regression loss $\mathcal{L}_{sty}$ was removed, we were not able to ensure the style information could be contained well in domain-specific features z^d, and thus lead to lower LPIPS (O2O) scores (i.e., images with poor diversity). Without $\mathcal{L}_{KL}$, we were not able to enforce the encoded domain-specific features to fit the prior Gaussian distribution, and thus failed to exhibit the multi-modal ability in cross-domain image translation. As a result, all the scores based on image realism and diversity dropped drastically. With the above quantitative ablation studies, we confirm the effectiveness and robustness of our M^{2}RD in performing multi-modal image translation across multiple domains.

D. Unsupervised Domain Adaptation

Finally, we apply our model for cross-domain classification. More specifically, we consider the challenging task of unsupervised domain adaptation (UDA), which aims at classifying images in the target domain while the labels are only available in the source domain during training. We conduct the UDA experiments using the handwritten digit datasets. For instance, MNIST → USPS indicates the use of MNIST as source-domain labeled data, while USPS is in the target domain without any categorical information. As mentioned in Section III-B, UDA can be achieved by our model by adding an extra classifier to recognize the disentangled content features. This classifier is jointly trained with our M^{2}RD.

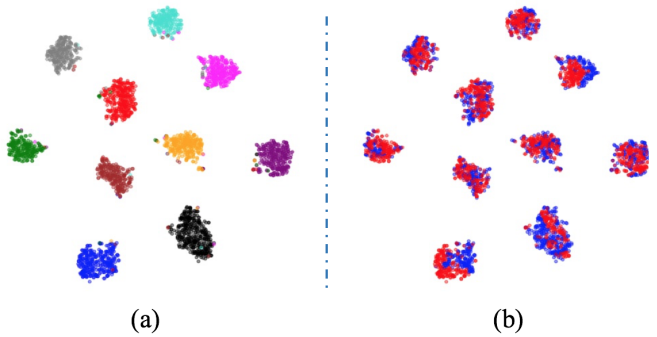


Fig. 12. t-SNE visualization of the handwritten digit data for USPS → MNIST. Note that different colors indicate data of (a) different digits classes **0-9** and (b) different domains (**source/target**).

Table IV compares the results of our model with recent translation-based UDA approaches. For MNIST → USPS, we achieved improved performances over the state-of-the-art methods, and our model performed favorably against others in USPS → MNIST. As for SVHN → MNIST, which is considered to be a more difficult scenario due to significant differences in background, stroke, and illumination, very promising results were reported by our proposed model as well.

In addition to quantitative evaluation, we further provide visualization results to further assess the UDA ability using our derived features. As shown in Fig. 12, we visualize domain-invariant representations of USPS → MNIST using t-SNE. To be more precise, Fig. 12a illustrates the image data of 10 categories which were properly separated, while Fig. 12b shows the same data associated with different domains (which are close to each other with reduced domain differences).

V. CONCLUSIONS

In this paper, we proposed a unified deep learning model of Multi-domain and Multi-modal Representation Disentangler (M^{2}RD). This unique network architecture addresses image manipulation and recognition across multiple domains by properly disentangling feature representation of interest. As a unique characteristics, multi-modal diversity is introduced into our proposed model, which realizes multi-modal image translation during the image manipulation process. In our experiments, we successfully verified that our model produced promising multi-domain and multi-modal image manipulation results using face, seasons, paints, and handwritten digit data, and can be applied to solve unsupervised domain adaptation with satisfactory accuracy.

ACKNOWLEDGMENT

This work is supported by the Ministry of Science and Technology of Taiwan under grant MOST 108-2634-F-002-018, and is funded in part by Qualcomm through a Taiwan University Research Collaboration Project.

REFERENCES

[1] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," *arXiv preprint arXiv:1508.06576*, 2015.

[2] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision*. Springer, 2016, pp. 694–711.

[3] X. Huang and S. J. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization." in *ICCV*, 2017, pp. 1510–1519.

[4] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M.-H. Yang, "Universal style transfer via feature transforms," in *Advances in Neural Information Processing Systems*, 2017, pp. 386–396.

[5] A. Sanakoyeu, D. Kotovenko, S. Lang, and B. Ommer, "A style-aware content loss for real-time hd style transfer," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 698–714.

[6] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial nets," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[7] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised cross-domain image generation," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

[8] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *arXiv preprint*, 2017.

[9] M.-Y. Liu and O. Tuzel, "Coupled generative adversarial networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[10] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[11] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2015.

[12] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko, "Adversarial discriminative domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[13] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan, "Domain separation networks," in *Advances in Neural Information Processing Systems*, 2016, pp. 343–351.

[14] P. Haeusser, T. Frerix, A. Mordvintsev, and D. Cremers, "Associative domain adaptation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2765–2773.

[15] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. A. Efros, and T. Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," *arXiv preprint arXiv:1711.03213*, 2017.

[16] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," in *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[17] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "β-VAE: Learning basic visual concepts with a constrained variational framework," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

[18] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier GANs," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2017.

[19] Y.-C. Liu, Y.-Y. Yeh, T.-C. Fu, S.-D. Wang, W.-C. Chiu, and Y.-C. F. Wang, "Detach and adapt: Learning cross-domain disentangled deep representation," *arXiv preprint arXiv:1705.01314*, 2017.

[20] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[21] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," *arXiv preprint arXiv:1804.04732*, 2018.

[22] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, "Diverse image-to-image translation via disentangled representations," *arXiv preprint arXiv:1808.00948*, 2018.

[23] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," *arXiv preprint*, vol. 1711, 2017.

[24] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman, "Toward multimodal image-to-image translation," in *Advances in Neural Information Processing Systems*, 2017, pp. 465–476.

[25] A. Gonzalez-Garcia, J. van de Weijer, and Y. Bengio, "Image-to-image translation for cross-domain disentanglement," in *Advances in Neural Information Processing Systems*, 2018, pp. 1287–1298.

[26] A. Liu, Y.-C. Liu, Y.-Y. Yeh, and Y.-C. F. Wang, "A unified feature disentangler for multi-domain image translation and manipulation," *arXiv preprint arXiv:1809.01361*, 2018.

[27] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz, "Mocogan: Decomposing motion and content for video generation," *arXiv preprint arXiv:1707.04993*, 2017.

[28] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning gan for pose-invariant face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1415–1424.

[29] Y. Liu, Z. Wang, H. Jin, and I. Wassell, "Multi-task adversarial network for disentangled feature learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3743–3751.

[30] X. Peng, X. Yu, K. Sohn, D. N. Metaxas, and M. Chandraker, "Reconstruction-based disentanglement for pose-invariant face recognition," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1623–1632.

[31] Y. Tian, X. Peng, L. Zhao, S. Zhang, and D. N. Metaxas, "Cr-gan: learning complete representations for multi-view generation," *arXiv preprint arXiv:1806.11191*, 2018.

[32] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[33] T. Kim, M. Cha, H. Kim, J. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2017.

[34] Z. Yi, H. Zhang, P. T. Gong *et al.*, "Dualgan: Unsupervised dual learning for image-to-image translation," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[35] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen, "Attgan: Facial attribute editing by only changing what you want," *arXiv preprint arXiv:1711.10678*, 2017.

[36] Z. Ding, S. Li, M. Shao, and Y. Fu, "Graph adaptive knowledge transfer for unsupervised domain adaptation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 37–52.

[37] Z. Ding and Y. Fu, "Deep transfer low-rank coding for cross-domain learning," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 6, pp. 1768–1779, 2018.

[38] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[40] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[41] A. Anoosheh, E. Agustsson, R. Timofte, and L. Van Gool, "Combogan: Unrestrained scalability for image domain translation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 783–790.

[42] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems*, 2017, pp. 6626–6637.

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

[44] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, "Unsupervised pixel-level domain adaptation with generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[45] M. Ghifary, W. B. Kleijn, M. Zhang, D. Balduzzi, and W. Li, "Deep reconstruction-classification networks for unsupervised domain adaptation," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 597–613.

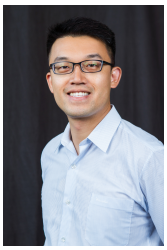
[46] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Computer Vision and Pattern Recognition (CVPR)*, vol. 1, no. 2, 2017, p. 4.

[47] S. Sankaranarayanan, Y. Balaji, C. D. Castillo, and R. Chellappa, "Generate to adapt: Aligning domains using generative adversarial networks," *arXiv preprint arXiv:1704.01705*, 2017.

[48] P. Russo, F. M. Carlucci, T. Tommasi, and B. Caputo, "From source to target and back: symmetric bi-directional adaptive gan," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8099–8108.

Fu-En Yang received the B.S. degree in Department of Electrical Engineering from National Taiwan University, Taipei, Taiwan, in 2018. He is currently a M.S. student of Graduate Institute of Communication Engineering at National Taiwan University, Taipei, Taiwan. His current research interests include computer vision, deep learning, and image processing.

Jing-Cheng Chang received the B.S. degree in Department of Electrical Engineering from National Taiwan University, Taipei, Taiwan, in 2019. He is currently a research assistant of Graduate Institute of Communication Engineering at National Taiwan University, Taipei, Taiwan. His current research interests include computer vision, deep learning, and image processing.



Chung-Chi "Charles" Tsai received the B.S. degree from National Tsing-Hua University, Hsinchu, Taiwan, and M.S. degree from University of California at Santa Barbara, Santa Barbara, CA, USA, and the Ph.D. degree from Texas A&M University, College Station, TX, USA, in 2009, 2012 and 2018, respectively and all in Electrical Engineering. He attended a one-year exchange program, at the University of New Mexico, Albuquerque, NM, USA, in 2007, and also participated in the summer internship with MediaTek in the summer of 2013/2015/2016. He is currently a senior system engineer for camera ISP in Qualcomm Technologies, Inc, San Diego, CA, USA. His research interests include image processing, computational photography, and computer vision.

Yu-Chiang Frank Wang received the B.S. degree in Electrical Engineering from the National Taiwan University, Taipei, Taiwan in 2001. He obtained the M.S. and Ph.D. degrees in electrical and computer engineering from Carnegie Mellon University, Pittsburgh, USA, in 2004 and 2009, respectively. Dr. Wang joined the Research Center for Information Technology Innovation (CITI), Academia Sinica, Taiwan, in 2009 as an assistant research fellow, and was later promoted as an associate research fellow in 2013. From 2015 to 2017, Dr. Wang also served as a Deputy Director of CITI at Academia Sinica.

In 2017, Dr. Wang joined the Graduate Institute of Communication Engineering and Department of Electrical Engineering at National Taiwan University, Taipei, Taiwan, as an associate professor, and is promoted to professor in 2019. He leads the Vision and Learning Lab at NTU, and focuses on research topics of computer vision and machine learning. He serves as Program Committee Members and Area Chairs at multiple international conferences or activities, and several of his papers were nominated for the Best Paper Awards at related international conferences such as IEEE ICIP, IEEE ICME and IAPR MVA. In 2013 and 2015, he was twice selected among the Outstanding Young Researchers by the Ministry of Science and Technology of Taiwan.