Multivariate-Information Adversarial Ensemble for Scalable Joint Distribution Matching

Ziliang Chen [*1] Zhanfu Yang [*2] Xiaoxi Wang [1] Xiaodan Liang [1] Xiaopeng Yan [1] Guanbin Li [1] Liang Lin [1]

Abstract

A broad range of cross-m-domain generation researches boil down to matching a joint distribution by deep generative models (DGMs). Hitherto algorithms excel in pairwise domains while as m increases, remain struggling to scale themselves to fit a joint distribution. In this paper, we propose a domain-scalable DGM, *i.e.*, MMI-ALI for m-domain joint distribution matching. As an m-domain ensemble model of ALIs (Dumoulin et al., 2016), MMI-ALI is adversarially trained with maximizing *Multivariate Mutual Information* (MMI) *w.r.t.* joint variables of each pair of domains and their shared feature. The negative MMIs are upper bounded by a series of feasible losses that provably lead to matching m-domain joint distributions. MMI-ALI linearly scales as m increases and thus, strikes a right balance between efficacy and scalability. We evaluate MMI-ALI in diverse challenging m-domain scenarios and verify its superiority.

1. Introduction

Remarkable advances of Deep Generative Models (DGMs), *e.g.*, *Generative Adversarial Net* (GAN) (Goodfellow et al., 2014), give rise to a variety of cross-domain generation and transfer tasks, *e.g.*, label-to-image translation (Isola et al., 2017; Wang et al., 2018), visual / text style transfers (Shen et al., 2017; Zhu et al., 2017), *etc*. In these scenarios, examples drawn from one domain transform their appearances via DGMs to synthesize the data patterns that belong to the other domains. This magic is formally interpreted as learning a joint distribution *w.r.t.* multi-domain random variables. Specifically, suppose that m ($\forall m \in \mathbb{N}_+$) domains underly marginal distributions $\{p_1, \cdots, p_m\}$. Given an example $\boldsymbol{x}_i \sim p_i$ ($\forall i \in [m] = \{1, \cdots, m\}$), DGMs generate $\boldsymbol{x}_j$ ($\forall j \in [m], j \neq i$) to satisfy the equation:

$$\begin{aligned} p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_m) : &= p(\{\boldsymbol{x}_j\}_{j\in[m]\&j\neq i}|\boldsymbol{x}_i)p(\boldsymbol{x}_i) \\ &= p_\Theta(\{\boldsymbol{x}_j\}_{j\in[m]\&j\neq i}|\boldsymbol{x}_i)p(\boldsymbol{x}_i) \end{aligned} \tag{1}$$

where $p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_m)$ denotes the joint distribution on m-domain random variables. $p(\{\boldsymbol{x}_j\}_{j\in[m]\&j\neq i}|\boldsymbol{x}_i)$ is the conditional distribution *w.r.t.* $\boldsymbol{x}_i$, and $p_\Theta(\{\boldsymbol{x}_j\}_{j\in[m]\&j\neq i}|\boldsymbol{x}_i)$ is parametrized from DGMs to match the m-domain joint distribution (Θ indicates the parameters of those DGMs). Eq.1 is connected with a broad set of GAN-based DGMs. Particularly when $m = 2$, (1) refers to finding a pair of generation nets to model $p(\boldsymbol{x}_2|\boldsymbol{x}_1)$ and $p(\boldsymbol{x}_1|\boldsymbol{x}_2)$, exactly the learning goal shared by c-GAN (Isola et al., 2017), CycleGAN (Zhu et al., 2017; Kim et al., 2017; Yi et al.) and other DGM methods (Dumoulin et al., 2016; Li et al., 2017).

Despite rapid progresses in learning paired-domain joint distribution, existing DGMs seldom prepare for the challenges as m>2, notably, the balance between model efficacy and scalability. On one hand, to cover $m(m-1)$ cross-domain transfer cases, most DGMs, *e.g.*, CycleGAN and JointGAN (Pu et al., 2018), have to deploy the same amount of (or even more) generation nets to learn m-domain joint distributions. It lacks efficiency in parameters and in turn, hinders them to capture richer information to improve their performances. On the other hand, recent heuristic methods, *i.e.*, StarGAN (Choi et al., 2017), attempt to suit all the transfer tasks by a single pipeline where each domain is treated as a class. Their pipelines are indeed scalable but the algorithms do not promise them to learn joint distributions. In fact, this line of methods can be technically fragile: If the supports of $\{p_i\}_{i=1}^m$ tend to intersect, treating domains as classes will fail and arouse serious model collapse.

In this paper, we focus on matching a m-domain joint distribution in a scalable and effective way. Instead of hacking a complex DGM pipeline, we revisit a famous *Adversarially Learned Inference* (ALI) (Dumoulin et al., 2016) model from a prospective of ensemble (Polikar, 2009).We assign m ALIs (allowed to share some of parameters) to each domain for learning m domain marginals by sharing their feature variables. By this mutual feature variable, each sample from domain i can be encoded to a feature by the inference net in the i^{th} ALI, then mapped into the j^{th} domain ($j \neq i$) by the generation net in the j^{th} ALI. This m inference-

[*] Equal contribution [1] Sun Yat-sen University, China [2] Purdue University, USA. Correspondence to: Liang Lin <linliang@ieee.org>.

Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

generation ensemble enable $m(m-1)$ transfer cases and more importantly, may lead to m-domain joint distribution by appropriately regulating cross-domain dependency.

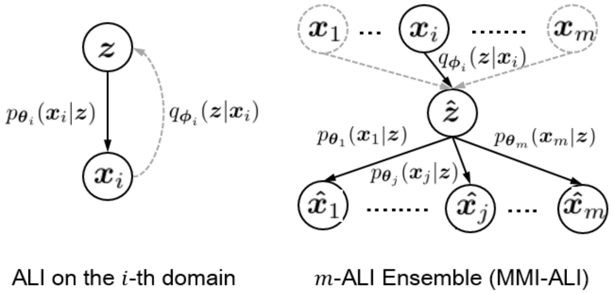
Specifically, we reframe this m-ALI ensemble trained with maximizing *multivariate mutual information* (MMI) (Bell, 2003; Mcgill, 2003). The MMIs act on arbitrary joint variables originating from each pair of domains and the domain-shared feature, which implies that m-domain information flow may exchange via their mutual feature. This observation nails down to a series of upper bounds that indicates conditional generation (Isola et al., 2017) and cycle consistency (Zhu et al., 2017). They are provably connected with matching a m-domain joint distribution and make the m-ALI ensemble our final model, *i.e.*, MMI-ALI.

MMI-ALI mainly contributes as:

1). MMI-ALI is linearly-scalable with m and more importantly, holds a series of loss upper bounds for provable joint distribution matching.
2). MMI-ALI revisit classical ALI from a view of ensemble model and learn with a adversarial ensemble loss (Sect.2.5), which are powerful for cross-domain generative modeling
3). A variety of m-domain experiments ($m \geq 2$) are placed in diverse scenarios, *e.g.*, 6-domain setup, visual / text style transfer, *etc*. The evaluation in supervised and unsupervised learning demonstrate the superiority of MMI-ALI.

Related work. Joint distribution matching has been considerably discussed in pairwise domain setups. Relevant researches based on GANs are classed into two lines. Models in the first line present as bidirectional DGMs associated with sample generation and feature inference, (Dumoulin et al., 2016; Donahue et al., 2016; Tolstikhin et al., 2017; Belghazi et al., 2018), real-real domain translations, *e.g.*, CycleGANs (Zhu et al., 2017; Kim et al., 2017; Yi et al.), the variants (Hoffman et al., 2017; Gan et al., 2017) and other adversarial dual learning models (Ulyanov et al., 2017; Deng et al., 2017). When cross-real-domain data are given in pairs, the second branch is connected with c-GAN (Isola et al., 2017) and other conditional adversarial DGMs (Reed et al., 2016b;a; Pathak et al., 2016; Wang et al., 2018) . (Li et al., 2017) shows their relationships by conditional entropy (CE). Our paper extends it into m-domain scenarios.

In m-domain setup, joint distribution becomes more cumbersome to learn and a few of recent DGMs refer to this problem. To the best of our knowledge, JointGAN (Pu et al., 2018) is the only existing research that promises (1) when $m > 2$. JointGAN chases for fully learning joint distribution, but ignores the scalability when m increases and requires C_m^3 generative modules to attain $m(m-1)$ cross-domain transformations. StarGAN (Choi et al., 2017) and its variants (Zhao et al., 2018; Kameoka et al., 2018) use a domain-shared backbone where each domain is viewed as a class. They cast m-domain transfer to a category generation problem and do not aim to learn a joint distribution.



ALI on the i-th domain

m-ALI Ensemble (MMI-ALI)

Figure 1. The overviews of ALI and m-ALI ensemble. MMI-ALI is learned from m-ALI ensemble with MMI constraints (Sect.2.4).

2. Multivariate Mutual Information Adversarially Learned Inference

In this section, we elaborate MMI-ALI in the following routine: **1).** We introduce ALI (Sect.2.1) and how it leads to an ensemble to achieve $m(m-1)$ cross-domain transfer tasks (Sect.2.2); **2).** We show the limitation of the m-ALI ensemble in cross-domain transfer (Sect.2.3) and how MMI induces a feasible regulation for the m-ALI ensemble to learn a joint distribution (Sect.2.4). **3).** We provide the adversarial ensembel learning algorithm of MMI-ALI (Sect.2.5). All proofs are deferred in our Appendix.A.

2.1. Preliminary: Adversarially Learned Inference

ALI is a bidirectional DGM derived from GAN, as it additionally incorporates an inference net trained with a generation net by playing against a discriminator. More specifically, in our context, suppose that a ALI model refers to generating a fake domain-i example $\hat{x}_i$ ($\forall i \in [m]$). Without loss of generality, we employ a distribution $q(z)$ as a prior on feature space $\mathbb{R}^d$, e.g. $q(z) = \mathcal{N}(\mathbf{0}^d, \mathbf{I}^{d\times d})$. Under the nonparametric assumption, we present the generation and inference nets by conditional distributions $p_{\theta_i}(\hat{x}_i|z)$ and $q_{\phi_i}(\hat{z}|x_i)$, where θ_i, ϕ_i denote their parameters and their inputs z, x_i are treated as the conditions. In this manner, ALI casts an adversarial game between p_{θ_i}, q_{ϕ_i} and a ω_i-parameterized critic net (discriminator) f_{ω_i} in

$$\begin{aligned}&\min_{\theta_i,\phi_i}\max_{\omega_i}\mathcal{L}_{\mathrm{ALI}}^{(i)}(\theta_i,\phi_i,\omega_i) = \\ &\mathbb{E}_{x_i\sim p(x_i),\hat{z}\sim q_{\phi_i}(\hat{z}|x_i)}\big[\log f_{\omega_i}(x_i,\hat{z})\big] \\ &\qquad + \mathbb{E}_{\hat{x}_i\sim p_{\theta_i}(\hat{x}_i|z),z\sim q(z)}\big[\log\big(1-f_{\omega_i}(\hat{x}_i,z)\big)\big]\end{aligned} \tag{2}$$

where $(x_i, \hat{z})$ denotes a real domain-i example x_i with its corresponding feature $\hat{z}$ inferred by q_{ϕ_i} and $(\hat{x}_i, z)$ denotes a fake domain-i sample $\hat{x}_i$ generated from $z \sim q(z)$ via p_{θ_i}. $f_{\omega_i}(\cdot,\cdot)$ is a binary classifier that distinguishes each sample-feature joint pair drawn from either $q_{\phi_i}(x_i, \hat{z})$ or $p_{\theta_i}(\hat{x}_i, z)$. The minimax objective (2) encourages the iterative update between ω_i and θ_i, ϕ_i. Similar to GAN, their resulting saddle point promises marginal matching on $p(x_i)$, $q(z)$.

Lemma 1 ((Dumoulin et al., 2016))**.** *The optimal generation, inference and critic nets w.r.t.,*$\{\theta_i^*, \phi_i^*, \omega_i^*\}$ ($\forall i \in [m]$)

refer to a saddle point in Eq.2 $\iff$ $p_{\boldsymbol{\theta}_i^*}(\boldsymbol{x}_i|\boldsymbol{z})q(\boldsymbol{z}) = q_{\boldsymbol{\phi}_i^*}(\boldsymbol{z}|\boldsymbol{x}_i)p_i(\boldsymbol{x}_i)$.

2.2. m-ALI Ensemble

With regards to m domains, there can be m ALIs that share the feature variable $\boldsymbol{z}$ to make marginal matchings on their own. It inspires an ensemble that associates m domains to enable $m(m-1)$ cross-domain data transformations. As illustrated in Fig.1.Right, suppose that $\forall \boldsymbol{x}_i \sim p_i$ is demanded to transform to the other j^{th} domain ($\forall i, j \in [m]$, $j \neq i$). By the aid of inference net $q_{\boldsymbol{\phi}_i}$ in the i^{th} ALI, it is able to encode $\boldsymbol{x}_i$ into a domain-agnostic feature $\hat{\boldsymbol{z}}$, and then use the generation net $p_{\boldsymbol{\theta}_j}$ in the j^{th} ALI to decode $\hat{\boldsymbol{z}}$ into $\hat{\boldsymbol{x}}_j$. This cross-domain generative process can be formulated as:

$$\begin{aligned} & p_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\{\hat{\boldsymbol{x}}_j\}_{j\in[m]\&j\neq i}|\boldsymbol{x}_i) \\ & \quad = \int p_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\{\hat{\boldsymbol{x}}_j\}_{j\in[m]\&j\neq i}|\hat{\boldsymbol{z}}, \boldsymbol{x}_i) p_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\hat{\boldsymbol{z}}|\boldsymbol{x}_i) d\hat{\boldsymbol{z}} \\ & \quad = \int \underbrace{\Big(\prod_{j\in[m]\&j\neq i} p_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\hat{\boldsymbol{x}}_j|\hat{\boldsymbol{z}}) \Big)}_{\substack{\text{Given } \hat{\boldsymbol{z}},\ \{\hat{\boldsymbol{x}}_j\}_{j\in[m]\&j\neq i} \text{ and} \\ \boldsymbol{x}_i \text{ are independent}}} p_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\hat{\boldsymbol{z}}|\boldsymbol{x}_i) d\boldsymbol{z} \\ & = \int \prod_{j\in[m], j\neq i} p_{\boldsymbol{\theta}_j}(\hat{\boldsymbol{x}}_j|\hat{\boldsymbol{z}}) q_{\boldsymbol{\phi}_i}(\hat{\boldsymbol{z}}|\boldsymbol{x}_i) d\hat{\boldsymbol{z}},\ s.t. \forall i \in [m] \end{aligned} \tag{3}$$

where we summarize the parameters of m-domain generation, inference, critic nets by $\boldsymbol{\Phi} = \{\boldsymbol{\phi}_i\}_{i=1}^m$, $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_i\}_{i=1}^m$, $\boldsymbol{\Omega} = \{\boldsymbol{\omega}_i\}_{i=1}^m$. As a cross-$m$-domain generative model, the m-ALI ensemble in (3) presents two advantages.

- **Scalability**: (3) is linearly-scalable with m. For subnets $\{q_{\boldsymbol{\phi}_i}\}_{i=1}^m$ and $\{p_{\boldsymbol{\theta}_i}\}_{i=1}^m$, it is possible to share their high-level layers across domains, as m-domain ALIs share their feature variable $\boldsymbol{z}$.
- **Generative model capability**: According to Lemma.1, (3) with $\boldsymbol{\phi}_i^*$ and $\boldsymbol{\theta}_j^*$ promises the transformed item $\hat{\boldsymbol{x}}_j$ following the true domain marginal p_j:

Proposition 1. *Given a pair of domains $\forall i, j \in [m]$, $i \neq j$, their well-trained ALIs (in Lemma.1) construct a cross-domain transfer process $p_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\hat{\boldsymbol{x}}_j|\boldsymbol{x}_i)$ that satisfies*

$$p_{\boldsymbol{\Phi}^*,\boldsymbol{\Theta}^*}(\hat{\boldsymbol{x}}_j) = \int p_{\boldsymbol{\Phi}^*,\boldsymbol{\Theta}^*}(\hat{\boldsymbol{x}}_j|\boldsymbol{x}_i) p_i(\boldsymbol{x}_i) d\boldsymbol{x}_i = p_j(\hat{\boldsymbol{x}}_j)$$

where $p_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\hat{\boldsymbol{x}}_j|\boldsymbol{x}_i)$ is the parameterized marginal of (3).

2.3. MMI-ALI: Motivation

How to learn m-ALI ensemble. As we previously discuss, m-ALI ensemble is a promising non-parametric model to achieve $m(m-1)$ cross-domain transfer, as the scalability and generative model capability have verified its potential. But the vital problem is, how to encourage the m-ALI ensemble to learn a m-domain joint distribution. Obvisouly, since each ALI model in m-ALI ensemble is independently trained, no cross-domain dependencies enforce $p_{\boldsymbol{\Phi},\boldsymbol{\Theta}}$ to approximate the joint distribution $p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_m)$. As long as generated data can match domain marginals (Proposition.1), (3) may tolerate all erratic cross-domain transfer. To tackle this problem, we first need to understand how to match a joint distribution in the m-domain scenario.

Criterion for m-domain joint distribution matching. In terms of supervised and unsupervised learning, joint distribution matching presents as satisfying different criterion. **1).** In *supervised learning*, we have access to draw samples from the true joint density $p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_m)$ and each of them presents as a m-tuple. Hence $p(\{\boldsymbol{x}_i\}_{i=1}^m)$ can be learned by minimizing the log-likelihood estimator:

$$\min\nolimits_{\boldsymbol{\Phi},\boldsymbol{\Theta}} \ -\mathbb{E}_p \big[\log p_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\{\boldsymbol{x}_i\}_{i=1}^m) \big] \tag{4}$$

2). In *unsupervised learning*, data across domains are unparalleledly aligned so that no access is provided to draw m-tuple from $p(\boldsymbol{x}_i, \cdots, \boldsymbol{x}_m)$. In the pairwise domain setup (Zhu et al., 2017), the unsupervised learning is typically considered as a cross-domain data reproduction problem that decreasing their conditional entropy (CE) theoretically helps to solve (see more in Li et al. 2017):

$$\min\nolimits_{\boldsymbol{\Phi},\boldsymbol{\Theta}} H\big(\boldsymbol{x}_i|\hat{\boldsymbol{x}}_j\big) = -\mathbb{E}_{p_{\boldsymbol{\Phi},\boldsymbol{\Theta}}} \big[\log p_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\boldsymbol{x}_i|\hat{\boldsymbol{x}}_j) \big] \tag{5}$$

where $H\big(\boldsymbol{x}_i|\hat{\boldsymbol{x}}_j\big)$ measures the input reproduction uncertainty *w.r.t.* $\boldsymbol{x}_i$ in the condition of $\hat{\boldsymbol{x}}_j$, *i.e.*, what the input has produced. In our scenario, we develop (5) to incorporate m-domain variables

$$\begin{aligned} & \min\nolimits_{\boldsymbol{\Phi},\boldsymbol{\Theta}} H\big(\boldsymbol{x}_i|\{\hat{\boldsymbol{x}}_j\}_{j\in[m]\&j\neq i}\big) \\ & = -\mathbb{E}_{p_{\boldsymbol{\Phi},\boldsymbol{\Theta}}} \big[\log p_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\boldsymbol{x}_i|\{\hat{\boldsymbol{x}}_j\}_{j\in[m]\&j\neq i}) \big] \end{aligned} \tag{6}$$

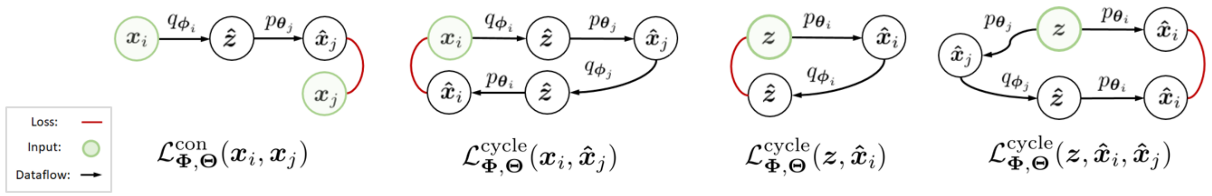
where $\forall i \in [m]$, $\boldsymbol{x}_i$ denotes an empirical draw from p_i; $\{\hat{\boldsymbol{x}}_j\}_{j=1\&j\neq i}^m$ denote fake items generated from $\boldsymbol{x}_i$ via (3).

It is worth noting that, (4) (6) with $m{=}2$ refer to condition (Isola et al., 2017) and cycle-consistency loss (Zhu et al., 2017) that have been widely-used in GAN-based DGM. But in general cases ($m \geq 2$), they are typically intractable and disconnected with the learning algorithm of ALI.

Rather than directly optimizing (4) (6), we prefer exploring the information-theoretic meaning behind m-domain joint distribution. In the next subsection, we introduce *Multivariate Mutual Information* (MMI) and explain it in the m-ALI ensemble context. We derive feasible MMIs *w.r.t.* each pair of domains and feature. They refer to a series of upper bounds that can also be interpreted as condition and cycle losses. They result in (4) (6) to promise m-ALI ensemble learn for joint distribution matching.

2.4. MMI-Induced Regularization

Before diving into further technical analysis, let's quickly go through MMI, the pivotal ingredient of our regularization.


Loss:
Input:
Dataflow:
$\boldsymbol{x}_i$
$q_{\boldsymbol{\phi}_i}$
$\hat{\boldsymbol{z}}$
$p_{\boldsymbol{\theta}_j}$
$\hat{\boldsymbol{x}}_j$
$\boldsymbol{x}_j$
$\mathcal{L}^{\text{con}}_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\boldsymbol{x}_i, \boldsymbol{x}_j)$
$\boldsymbol{x}_i$
$q_{\boldsymbol{\phi}_i}$
$\hat{\boldsymbol{z}}$
$p_{\boldsymbol{\theta}_j}$
$\hat{\boldsymbol{x}}_j$
$q_{\boldsymbol{\phi}_j}$
$\hat{\boldsymbol{x}}_i$
$p_{\boldsymbol{\theta}_i}$
$\hat{\boldsymbol{z}}$
$\mathcal{L}^{\text{cycle}}_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\boldsymbol{x}_i, \hat{\boldsymbol{x}}_j)$
$\boldsymbol{z}$
$p_{\boldsymbol{\theta}_i}$
$\hat{\boldsymbol{x}}_i$
$q_{\boldsymbol{\phi}_i}$
$\hat{\boldsymbol{z}}$
$\mathcal{L}^{\text{cycle}}_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\boldsymbol{z}, \hat{\boldsymbol{x}}_i)$
$p_{\boldsymbol{\theta}_j}$
$\boldsymbol{z}$
$p_{\boldsymbol{\theta}_i}$
$\hat{\boldsymbol{x}}_j$
$\hat{\boldsymbol{x}}_i$
$q_{\boldsymbol{\phi}_j}$
$\hat{\boldsymbol{z}}$
$p_{\boldsymbol{\theta}_i}$
$\hat{\boldsymbol{x}}_i$
$\mathcal{L}^{\text{cycle}}_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\boldsymbol{z}, \hat{\boldsymbol{x}}_i, \hat{\boldsymbol{x}}_j)$

(a).Observation.1 (b).Observation.2 (c).Observation.3

Figure 2. The diagram of constructing MMI-induced regularizations by generation and inference nets in m ALIs. Best viewed in color.

Multivariate Mutual Information (MMI). Given a pair of random variables $\boldsymbol{x}$, $\boldsymbol{y}$, *Mutual Information* (MI) $I(\boldsymbol{x};\boldsymbol{y})$ quantifies the amount of information one of them contains about the other, *i.e.*,

$$I(\boldsymbol{x};\boldsymbol{y}) = I(\boldsymbol{y};\boldsymbol{x}) := H(\boldsymbol{y}) - H(\boldsymbol{y}|\boldsymbol{x}) \tag{7}$$

. Maximizing $I(\boldsymbol{x};\boldsymbol{y})$ relates to an invertible function that knowing one of $\boldsymbol{x}$, $\boldsymbol{y}$ almost reveals the other. MMI extends MI by including n random variables $\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n$ ($\forall n \in \mathbb{N}_+$). It can be recursively defined as

$$\begin{aligned} &I(\boldsymbol{y}_1;\cdots;\boldsymbol{y}_n) \\ &\quad := I(\boldsymbol{y}_1;\cdots;\boldsymbol{y}_{n-1}) - I(\boldsymbol{y}_1;\cdots;\boldsymbol{y}_{n-1}|\boldsymbol{y}_n) \end{aligned} \tag{8}$$

where $I(\boldsymbol{y}_1;\cdots;\boldsymbol{y}_{n-1}|\boldsymbol{y}_n)$ denotes *Conditional Mutual Information* (CMI), the expectation of $I(\boldsymbol{y}_1;\cdots;\boldsymbol{y}_{n-1})$ when its value is conditioned on $\boldsymbol{y}_n$.

MMI for joint distribution matching. MMI resembles the information-theoretic sense of MI. Maximizing m-domain MMI with respect to densities parameterized by $\boldsymbol{\Phi}$, $\boldsymbol{\Theta}$, *i.e.*, $I_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\boldsymbol{x}_1;\cdots;\boldsymbol{x}_m)$, intuitively encourages discovering an identical information flow from one domain to the others. It corresponds to the cross-domain transfer $p_{\boldsymbol{\Phi},\boldsymbol{\Theta}}$ under m-domain joint distribution matching. However, on the basis of the recursive routine in (8), m-variable MMI is comprised of $\mathcal{O}(2^m)$ entropy terms that can be positive or negative. It makes $I_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\boldsymbol{x}_1;\cdots;\boldsymbol{x}_m)$ intractable and formidable to extend with m. Besides, it probably arouses unstable optimization, as $I_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\boldsymbol{x}_1;\cdots;\boldsymbol{x}_m)$ may be unbounded.

Instead of simultaneously considering m-domain variables, we tend to explore the linear combination of MMIs on each pair of domain variables $\boldsymbol{x}_i$, $\boldsymbol{x}_j$ with the m-domain-shared feature variable $\boldsymbol{z}$. In this principle, MMI $I_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\boldsymbol{x}_i;\boldsymbol{x}_j;\boldsymbol{z})$ has been covered $m(m-1)$ transfer cases and their maximizations are understood as

$$\min_{\boldsymbol{\Phi},\boldsymbol{\Theta}} \quad - \sum_{i,j\in[m], i\neq j} I_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\boldsymbol{x}_i;\boldsymbol{x}_j;\boldsymbol{z}) \tag{9}$$

which implies the m-domain information flows exchange via their features. $I_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\boldsymbol{x}_i;\boldsymbol{x}_j;\boldsymbol{z})$ conceives two technical merits. First, three-variable MMI is always non-positive and thus, the minimization $-I_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\boldsymbol{x}_i;\boldsymbol{x}_j;\boldsymbol{z})$ is lower bounded by 0, which substantially stabilizes the optimization process. Second, $-I_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\boldsymbol{x}_i;\boldsymbol{x}_j;\boldsymbol{z})$ can be pushed into a line of upper bounds that serve as condition and cycle-consistency losses. Their minimization results in (4) (6) that encourages $p_{\boldsymbol{\Phi},\boldsymbol{\Theta}}$ to learn the m-domain joint distribution. We are going to elaborate them.

Upper bounds. Derived from ALIs, $-I_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\boldsymbol{x}_i;\boldsymbol{x}_j;\boldsymbol{z})$ consists of generation and inference nets. Hence inputs underlie true distributions and may be drawn from either m domain marginals $\{p_i\}_{i=1}^m$ or feature density $q(\boldsymbol{z})$. Suppose that $\boldsymbol{x}_i$, $\boldsymbol{x}_j$, $\boldsymbol{z}$ denote the observed variables *w.r.t.* true distributions and $\hat{\boldsymbol{x}}_i$, $\hat{\boldsymbol{x}}_j$, $\hat{\boldsymbol{z}}$ denote the variables *w.r.t.* $\boldsymbol{\Phi}, \boldsymbol{\Theta}$-parameterized distributions. The upper bounds derived from $-I_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\boldsymbol{x}_i;\boldsymbol{x}_j;\boldsymbol{z})$ can be interpreted in three aspects.

In the supervised case, training instances are m-tuples and for each domain-i empirical draw, it is able to search its corresponding domain-j empirical draw as the transformation groundtruth. In this scenario, $-I_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\boldsymbol{x}_i;\boldsymbol{x}_j;\boldsymbol{z})$ is bounded by the condition loss $\mathcal{L}^{\text{con}}_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\boldsymbol{x}_i,\boldsymbol{x}_j)$ as below

Observation 1. *Given empirical draws from p_i ($\forall i \in [m]$), in supervised learning,*

$$\begin{aligned} &- I_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\boldsymbol{x}_i;\boldsymbol{x}_j;\hat{\boldsymbol{z}}) \leq H_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\boldsymbol{x}_i|\boldsymbol{x}_j) \\ &\leq \mathop{\mathbb{E}}_{\boldsymbol{x}_i,\boldsymbol{x}_j\sim p_{i,j}} - \Big[\log \int p_{\boldsymbol{\theta}_i}(\boldsymbol{x}_i|\hat{\boldsymbol{z}}) q_{\boldsymbol{\phi}_j}(\hat{\boldsymbol{z}}|\boldsymbol{x}_j) d\hat{\boldsymbol{z}}\Big] \triangleq \mathcal{L}^{\text{con}}_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\boldsymbol{x}_i,\boldsymbol{x}_j) \end{aligned} \tag{10}$$

where $p_{i,j} = p(\boldsymbol{x}_i, \boldsymbol{x}_j)$.

In Fig.2.a., we show how to build $\mathcal{L}^{\text{con}}_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\boldsymbol{x}_i,\boldsymbol{x}_j)$. The loss can be implemented by l_1/l_2 norms.

In the unsupervised case, each empirical draw is separately given, therefore we have no access to $\boldsymbol{x}_j$. Distinct from (10), the MMI turns into $I_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\boldsymbol{x}_i;\hat{\boldsymbol{x}}_j;\hat{\boldsymbol{z}})$ where $\hat{\boldsymbol{x}}_j$ implies that domain-j samples are counterfeits and the bound constitutes a cross-domain cycle-consistency loss by means of $\hat{\boldsymbol{z}}$:

Observation 2. *Given empirical draws from p_i ($\forall i \in [m]$), in unsupervised learning,*

$$\begin{aligned} &- I_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\boldsymbol{x}_i;\hat{\boldsymbol{x}}_j;\hat{\boldsymbol{z}}) \leq H_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\boldsymbol{x}_i|\hat{\boldsymbol{x}}_j) \\ &\leq \mathop{\mathbb{E}}_{\boldsymbol{x}_i,\hat{\boldsymbol{x}}_j\sim p_{\boldsymbol{\theta}_j,\boldsymbol{\phi}_i}} - \Big[\log \int p_{\boldsymbol{\theta}_i}(\boldsymbol{x}_i|\hat{\boldsymbol{z}}) q_{\boldsymbol{\phi}_j}(\hat{\boldsymbol{z}}|\boldsymbol{x}_j) d\hat{\boldsymbol{z}}\Big] \triangleq \mathcal{L}^{\text{cycle}}_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\boldsymbol{x}_i,\hat{\boldsymbol{x}}_j) \end{aligned} \tag{11}$$

where $p_{\boldsymbol{\theta}_j,\boldsymbol{\phi}_i} = p(\boldsymbol{x}_i) \int_{\hat{\boldsymbol{z}}} p_{\boldsymbol{\theta}_j}(\hat{\boldsymbol{x}}_j|\hat{\boldsymbol{z}}) q_{\boldsymbol{\phi}_i}(\hat{\boldsymbol{z}}|\boldsymbol{x}_i) d\hat{\boldsymbol{z}}$.

$\mathcal{L}^{\text{cycle}}_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\boldsymbol{x}_i,\hat{\boldsymbol{x}}_j)$ is constructed as illustrated in Fig.2.b.

The observations above presumed inputs drawn from the domain marginals $\{p_i\}_{i=1}^m$. If inputs are drawn from the feature distribution $q(\boldsymbol{z})$, $\hat{\boldsymbol{x}}_i$, $\hat{\boldsymbol{x}}_j$ would be generated from $\boldsymbol{z}$, and $-I_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\hat{\boldsymbol{x}}_i;\hat{\boldsymbol{x}}_j;\boldsymbol{z})$ is upper bounded by the conditional entropies $H_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\boldsymbol{z}|\hat{\boldsymbol{x}}_i)$ and $H_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\hat{\boldsymbol{x}}_j|\hat{\boldsymbol{x}}_i)$. They are equivalent to the cycle losses $\mathcal{L}^{\text{cycle}}_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\boldsymbol{z},\hat{\boldsymbol{x}}_i)$ and $\mathcal{L}^{\text{cycle}}_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\boldsymbol{z},\hat{\boldsymbol{x}}_i,\hat{\boldsymbol{x}}_j)$, which are revealed in Fig.2.c.

Observation 3. *Given empirical draws from $q(\boldsymbol{z})$,*

$$-I_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\hat{\boldsymbol{x}}_i;\hat{\boldsymbol{x}}_j;\boldsymbol{z}) \leq H_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\boldsymbol{z}|\hat{\boldsymbol{x}}_i) + H_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\hat{\boldsymbol{x}}_j|\hat{\boldsymbol{x}}_i) \quad (12)$$

$$H_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\boldsymbol{z}|\hat{\boldsymbol{x}}_i) = \mathop{\mathbb{E}}_{\hat{\boldsymbol{x}}_i\sim p_{\boldsymbol{\theta}_i},\boldsymbol{z}\sim q(\boldsymbol{z})} -\log q_{\boldsymbol{\phi}_i}(\boldsymbol{z}|\hat{\boldsymbol{x}}_i) \triangleq \mathcal{L}^{\text{cycle}}_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\boldsymbol{z},\hat{\boldsymbol{x}}_i)$$

$$H_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\hat{\boldsymbol{x}}_j|\hat{\boldsymbol{x}}_i) = \mathop{\mathbb{E}}_{\substack{\boldsymbol{z}\,\sim\, q(\boldsymbol{z})\\ \hat{\boldsymbol{x}}_i\,\sim\, p_{\theta_i},\hat{\boldsymbol{x}}_j\,\sim\, p_{\theta_j}}} -\Big[\log \int_{\boldsymbol{z}} p_{\boldsymbol{\theta}_j}(\hat{\boldsymbol{x}}_j|\boldsymbol{z})q_{\boldsymbol{\phi}_i}(\boldsymbol{z}|\hat{\boldsymbol{x}}_i)d\boldsymbol{z}\Big]$$
$$\triangleq \mathcal{L}^{\text{cycle}}_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\boldsymbol{z},\hat{\boldsymbol{x}}_i,\hat{\boldsymbol{x}}_j)$$

Associate Observations (1-3) and we impose cross-domain structure dependencies on $\boldsymbol{\Phi}$, $\boldsymbol{\Theta}$ by

$$\begin{aligned}\mathcal{R}_{\text{SL}}(\boldsymbol{\Theta},\boldsymbol{\Phi}) &= \sum_{i,j\in[m],i\neq j} \mathcal{L}^{\text{con}}_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\boldsymbol{x}_i,\boldsymbol{x}_j) + \mathcal{L}^{\text{cycle}}_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\boldsymbol{z},\hat{\boldsymbol{x}}_i)\\ &\qquad + \mathcal{L}^{\text{cycle}}_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\boldsymbol{z},\hat{\boldsymbol{x}}_i,\hat{\boldsymbol{x}}_j)\\ \mathcal{R}_{\text{UL}}(\boldsymbol{\Theta},\boldsymbol{\Phi}) &= \sum_{i,j\in[m],i\neq j} \mathcal{L}^{\text{cycle}}_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\boldsymbol{x}_i,\hat{\boldsymbol{x}}_j) + \mathcal{L}^{\text{cycle}}_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\boldsymbol{z},\hat{\boldsymbol{x}}_i)\\ &\qquad + \mathcal{L}^{\text{cycle}}_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\boldsymbol{z},\hat{\boldsymbol{x}}_i,\hat{\boldsymbol{x}}_j)\end{aligned} \quad (13)$$

where $\mathcal{R}_{\text{SL}}$ / $\mathcal{R}_{\text{UL}}$ respectively regulate the supervised / unsupervised learning and upper bound (9). It implies that the minimization of $\mathcal{R}_{\text{SL}}$, $\mathcal{R}_{\text{UL}}$ equal to maximizing the MMIs. By Proposition.1, desire that adversarial learning (2) encourages $\{p_i\}_{i=1}^m$ and parameterized domain marginals agree with a high likelihood to domain variables $\big($*i.e.*, $\boldsymbol{x}_i = \hat{\boldsymbol{x}}_i$ in (13)$\big)$, then the minimization of $\mathcal{R}_{\text{SL}}$, $\mathcal{R}_{\text{UL}}$ leads to the joint distribution matching criterion (4),(6).

Theorem 1. *Suppose that true and parameterized domain marginal distributions maintain a high likelihood to domain variables, $\mathcal{R}_{\text{SL}} \to 0$ leads to the optima in (4); $\mathcal{R}_{\text{UL}} \to 0$ leads to the optima in (6).*

2.5. Adversarial Ensemble Learning

Learning m-ALI ensemble by (13) is able to capture the m-domain joint density. But it can be problematic as samples directly generated from $q(\boldsymbol{z})$ can be of low quality, *e.g.*, due to the poorly-efficient sampling in a high-dimensional feature space. To overcome this issue, we invent a *domain mixture adversarial ensemble (DMAE) loss* to refine (2) :

$$\begin{aligned}\mathcal{L}^{(i)}_{\text{DMAE}}(\boldsymbol{\Phi},\boldsymbol{\Theta},\boldsymbol{\Omega}) &= \mathbb{E}_{\boldsymbol{x}_i,\hat{\boldsymbol{z}}\sim q_{\boldsymbol{\phi}_i}(\boldsymbol{x}_i,\hat{\boldsymbol{z}})}\big[\log f_{\boldsymbol{\omega}_i}(\boldsymbol{x}_i,\hat{\boldsymbol{z}})\big]\\ &+\sum_{j=1}^{m}\pi_j\Big(\mathbb{E}_{\hat{\boldsymbol{x}}_i\sim p_{\boldsymbol{\theta}_i}(\hat{\boldsymbol{x}}_i|\boldsymbol{z}),\boldsymbol{z}\sim q_{\boldsymbol{\phi}_j}}\left[\log\left(1-f_{\boldsymbol{\omega}_i}(\hat{\boldsymbol{x}}_i,\boldsymbol{z})\right)\right]\Big)\end{aligned} \quad (14)$$

where $\sum_{j=1}^m \pi_j = 1$ indicates the proportion of the domain mixture for adversary. Compared with (2) whose fake samples are solely generated from $q(\boldsymbol{z})$, $\mathcal{L}^{(i)}_{\text{DMAE}}(\boldsymbol{\Phi},\boldsymbol{\Theta},\boldsymbol{\Omega})$ consider fake samples generated from the domain-encoded features, which are derived from the real samples that belong to the other domains, *i.e.*, $\boldsymbol{z} \sim \int q_{\boldsymbol{\phi}_j}(\boldsymbol{z},\boldsymbol{x}_j)d\boldsymbol{x}_j$ $(\forall j \in [m])$. These fake samples converted from different domains are unified into the DMAE loss (14) to cheat the domain-i critic net $f_{\boldsymbol{\omega}_i}$. It can be provably verified that, the adversarial ensemble learning retains the theoretical property of (2):

Proposition 2. *The optimum of the generation, inference and critic networks in*

$$\min_{\boldsymbol{\Theta},\boldsymbol{\Phi}}\max_{\boldsymbol{\Omega}}\ (1-\gamma)\sum_{i=1}^{m}\mathcal{L}^{(i)}_{\text{ALI}} + \gamma\sum_{i=1}^{m}\mathcal{L}^{(i)}_{\text{DMAE}} \quad (15)$$

refer to their saddle points in Lemma.1 if and only if $\forall i \in [m]$, there exist $p_{\boldsymbol{\theta}^*_i}(\boldsymbol{x}|\boldsymbol{z})q(\boldsymbol{z}) = q_{\boldsymbol{\phi}^*_i}(\boldsymbol{z}|\boldsymbol{x})p(\boldsymbol{x})$.

where γ denotes the trade-off between (2) and DAME loss. Proposition.2 demonstrates that, even if we change the learning objective (2), Lemma.1 and the other analysis based on (2) can be completely followed by the new objective (15).

Combining (13) and (15), we formalize MMI-ALI as

$$\min_{\boldsymbol{\Theta},\boldsymbol{\Phi}}\max_{\boldsymbol{\Omega}}\ (1-\gamma)\sum_{i=1}^{m}\mathcal{L}^{(i)}_{\text{ALI}} + \gamma\sum_{i=1}^{m}\mathcal{L}^{(i)}_{\text{DMAE}} + \beta\,\mathcal{R}_{\text{SL}}/\mathcal{R}_{\text{UL}} \quad (16)$$

where $\mathcal{R}_{\text{SL}}/\mathcal{R}_{\text{UL}}$ are switched by supervised/unsupervised learning and $\beta > 0$ denotes the loss-balance factor. Normally, we set $\beta = 1$ in our implementation.

3. Experiments

In this section, we propose diverse cross-m-domain experiments to evaluate our MMI-ALI in generative modeling and show the primal empirical results. More experiments (*e.g.*, ablation) and visualization are founded in Appendix.B.

3.1. Balance between efficacy and scalability

Compared with existing methods, MMI-ALI strikes a right balance between model capacity and scalability. To highlight this merit, we design the first experiment on synthetic data domains with m ranged in 2∼6. We choose $q(\boldsymbol{z})$ as an isotropic Gaussian $\mathcal{N}(\mathbf{0},\mathbf{I})$, then each density in $\{p_i\}_{i=1}^m$ is a 2D Gaussian Mixture Model (GMM) with 5 components $\mathcal{N}(\mathbf{0}, 0.2\mathbf{I})$. (As illustrated in Fig.4) Due to the simplicity of synthetic data, we only consider unsupervised learning across them. We evaluate MMI-ALI and its parameter-shared version termed "MMI-ALI (PS)", with CycleGAN and StarGAN. All of them are trained on 2048 with vanilla GAN loss and tested on 1024 examples drawn from each of $\{p_i\}_{i=1}^m$. For a fair comparison, all baselines use two-layered fully-connected nets with ReLU to generate data and make critics. l_2-norm is chosen as the cycle-consistency loss for all baseline during training.

Evaluation. Two measures have been introduced. The first is *geometric score* (GS) (Khrulkov & Oseledets, 2018) that

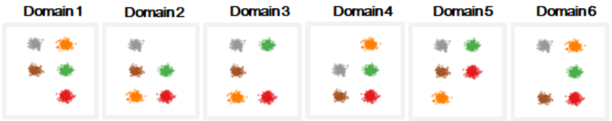


Figure 3. Synthetic domains used in our first experiments. As m increases, they are proceedingly incorporated for multi-domain joint distribution leanring from left to right.

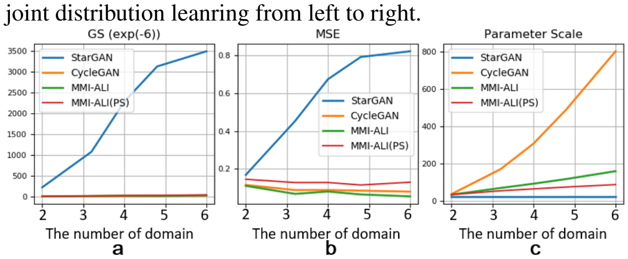


Figure 4. Transfer evaluations with 2∼6 synthetic domains: (a). Geometric Score (GS, lower is better); (b). Mean Square Error (MSE, lower is better); (c). Parameter Scale (lower is better).

evaluates generation quality by comparing the topological properties of the supports behind the generated and true domain marginals. The other is *mean squared error* (MSE) broadly used to measure the conditional density modeling via sample reconstruction quality across domains. Each baseline is performed in average of $m(m-1)$ transfer cases on two measures to thoroughly reflect the learned joint distribution. The results and parameters are shown in Fig.4.(a-b) and (c), respectively. Note that, StarGAN uses a domain-shared pipeline so that its parameter scale is almost consistent as m increases. However, StarGAN's GS, MSE heavily suffer even in toy domains, due to its intrinsic vulnerability as we have discussed. Particularly, when there exists an overlap across domains, the examples drawn from the overlap (or close to the overlap) can belong to all of these domains. This phenomena is general (see our empirical results in real data) and StarGANs can do nothing to help. On the other hand, MMI-ALI and CycleGAN are close in GS and MSE, yet CycleGAN requires exponentially-increasing parameters. They demonstrate that MIM-ALIs remain convincing performances as they scale to the scenarios with more domains. We show more visualization results in SM.

3.2. Geometry-varying m domains.

Geometry-varying information is difficult to capture in generative modeling (Sabour et al., 2017). Based on this challenge, our second experiment considers cross-m-domain generation where the m-domain samples present significant variation in geometry. We evaluate whether this information can be captured by MMI-ALI and the other baselines.

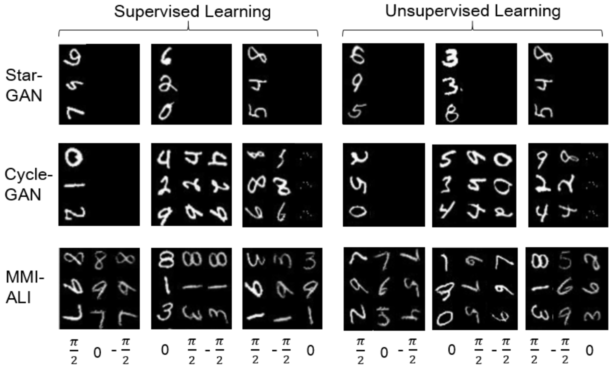

Supervised Learning
Unsupervised Learning
Star-
GAN
Cycle-
GAN
MMI-
ALI
$\frac{\pi}{2}$ 0 -$\frac{\pi}{2}$
0 $\frac{\pi}{2}$ -$\frac{\pi}{2}$
$\frac{\pi}{2}$ -$\frac{\pi}{2}$ 0
$\frac{\pi}{2}$ 0 -$\frac{\pi}{2}$
0 $\frac{\pi}{2}$ -$\frac{\pi}{2}$
$\frac{\pi}{2}$ -$\frac{\pi}{2}$ 0

Figure 5. Cross-3-domain generation performed by StarGAN, CycleGAN and MMI-ALI (ours) in supervised and unsupervised learning setups. ***For each sub-picture, the left column indicates inputs*** and the rest indicate the cross-domain transformed results.

Table 1. SSIM of StarGAN (ST), CycleGAN (CG) and MMI-ALI(MA) in supervised cross-domain generation case.

	1%	5%	10%
ST	0.00	0.00	0.00
CG	0.32	0.31	0.35
MA	**0.57**	**0.68**	**0.72**

Table 2. IS of StarGAN (ST), CycleGAN (CG) and MMI-ALI(MA) in unsupervised cross-domain generation case.

	$-\frac{\pi}{2}\to 0$	$\frac{\pi}{2}\to 0$	$0\to\frac{\pi}{2}$	$-\frac{\pi}{2}\to\frac{\pi}{2}$	$-\frac{\pi}{2}\to 0$	$\frac{\pi}{2}\to-\frac{\pi}{2}$
ST	1.00	1.00	1.00	1.00	1.00	1.00
CG	8.34	6.13	2.25	2.38	1.71	1.04
Ours	8.99	9.01	2.95	3.86	3.31	3.08

Specifically, we choose MNIST as the base domain, then rotate the images by $-\frac{\pi}{2}^{\circ}$, $\frac{\pi}{2}^{\circ}$ to create two other domains. Then MMI-ALI, CycleGAN and StarGAN are demanded to learn pattern transfer across the three domains in supervised and unsupervised learning setups. In supervised setup, data present as triplets so that each example from one domain has its corresponding groundturth in other domains. This information is not provided in unsupervised cases. In supervised case, we compare (supervised) MMI-ALI with CycleGAN and StarGAN augmented with condition loss used by c-GAN. In unsupervised case, we compare (unsupervised) MMI-ALI with ordinary CycleGAN and StarGAN. For a fair comparison, we standardize backbone behind the baselines in DCGAN (Dumoulin et al., 2016), and they are trained with vanilla GAN and l_1-norm cycle losses.

Evaluation. In supervised learning setup, we measure transformed results by Structured SIMilarity (SSIM) (Zhou et al., 2004). The visualization and quantitative results are shown in Fig.4 and Table.2, respectively. MMI-ALI is the *only baseline* that can produce all transfer patterns. StarGAN collapses during training and create nothing for transfer. CycleGAN performs better than MMI-ALI in $0 \to -\frac{\pi}{2}, \frac{\pi}{2}$, however, fails in capturing larger rotation (*e.g.*, $-\frac{\pi}{2} \to \frac{\pi}{2}$). It demonstrates a weakness of CycleGAN, which merely learns a pairwise joint distribution per time. In other word, it can not leverage m-domain knowledge to enhance the cross-domain generation performance. MMI-ALI avert this issue due to modeling m-domain joint distribution by ensemble. For more concrete evaluation, we provide different proportion of supervised data, *i.e.*, 1%, 5%, 10%, to check how much the model can benefit from supervision. We find that in 3-domain Rotated MNIST, cross-domain align-

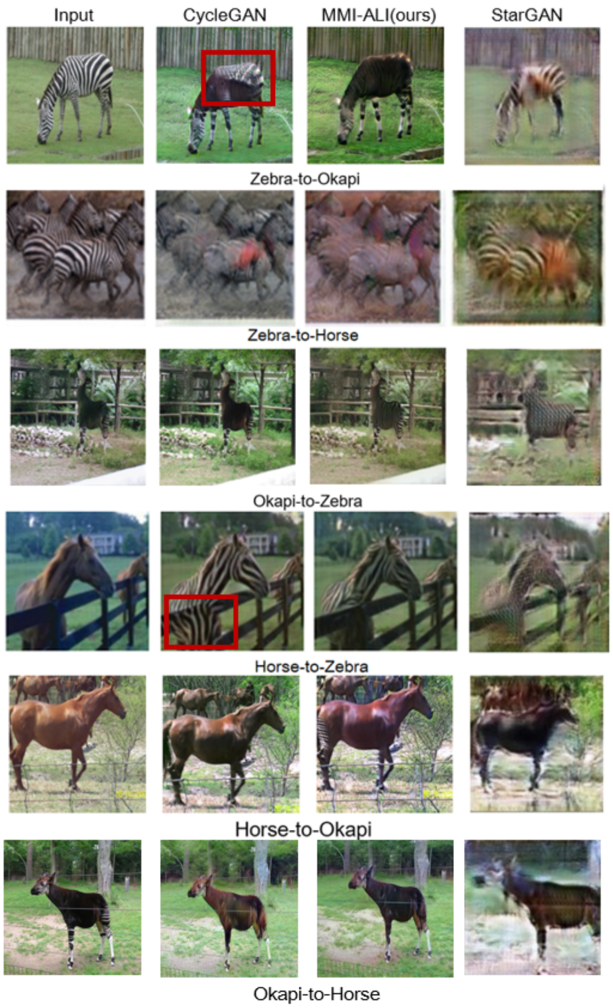


Figure 6. Style transfer on Zebra&Horse&Okapi.

ment can not significantly help StarGAN and CycleGAN to improve their joint distribution learning performance. But MMI-ALI can benefit from small amount of supervision. Cross-domain digit transformation conceives structure variation, thus, the patterns are difficult to capture without supervisions. This statement is verified in unsupervised results shown in Fig.4. Even so, our MMI-ALI is still powerful in generative modeling. To be specific, we evaluate the unsupervised generation by *Inception Score* (Salimans et al., 2016). MMI-ALI consistently outperform the other baselines across 6 cross-domain generation scenarios.

3.3. Cross-m-domain visual style transfer.

In this experiment, we consider 3-domain object transfiguration and 3-heterogeneous-domain style transfer.

In object transfiguration, evaluated DGMs are required to transform a specific part of an object to some target pattern whereas the other parts remain the same. One example is to translate a sort of animals (*e.g.*, 1000 classes in ImageNET) to become another kind with visual similarity. In our experiment, we consider the 3-object transfiguration in $\mathrm{Zebra} \leftrightarrow \mathrm{Horse} \leftrightarrow \mathrm{Okapi}$, where Zebra and Horse share their shapes while differ from the strip; then Okapi is *"zebra- striped" on its legs with a "horse-like" torso*. The experiment is conducted by reconfiguring the state-of-the-art residual-block-based (He et al., 2015) CycleGAN into MMI-ALI. For a fair comparison with CycleGAN, we depart the generator of CycleGAN as a pair of inference and generation net for our MMI-ALI, and follow the identical training tricks. Instead of using a non-informative prior, we apply $\boldsymbol{z} = \mu(\boldsymbol{z}) + \epsilon$ to provide features. As for StarGAN, we employ the official code reported in their original paper where their models are also built on ResNet.

In 3-heterogeneous-domain transfer, we consider Cityscape (Cordts et al., 2016) as the base benchmark, then employ the real data and their segmentation labels to construct two domains (R and Seg). We further applied the pretrained sketch detector (Xie & Tu, 2015) to generate the third domain (Ske). To this we are able to evaluate all baselines in unsupervised and supervised learning manners (Condition loss is used in the supervised case). We resemble the similar configuration and training strategy in object transfiguration.

Evaluation. Amazon Mechanical Turk (AMT) is employed to evaluate the object transfiguration experiment. We follow the perceptual evaluation from (Dong et al., 2018), where workers are provided with a pair of generated image (ours and the other baseline), and given unlimited time to select the one more likely as a target domain image. In Cityscape, we take *Frechet Inception Distance* (FID)(Heusel et al., 2018) and MSE as the metrics (MSE deferred in SM).

Table 3. Pairwise comparison of MMI-ALI with other baselines. Chance is at 50%. Each cell indicates the percentage where our result is preferred over the other method. MMI-ALI overwhelmingly outperforms StarGAN and stay ahead of CycleGAN.

	Okapi2Zebra	Okapi2Horse	Zebra2Okapi	Horse2Okapi
StarGAN	100.0%	100.0%	97.6%	100.0%
CycleGAN	57.2%	52.1%	56.5%	67.2%

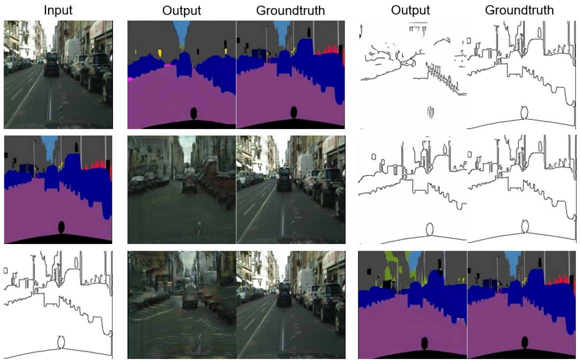


Figure 7. Cross-3-domain supervised transfer in Cityscape.

The visualization of object transfiguration are illustrated in Fig.6. First of all, StarGAN takes a mild effect. Due to the

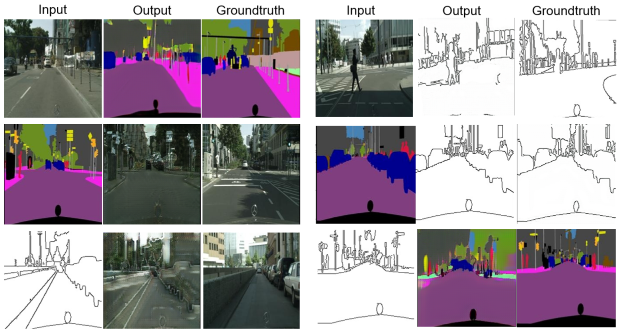


Figure 8. Cross-3-domain unsupervised transfer in Cityscape.

its category-generative pipeline, cross-domain style knowledge is hardly disentangled and thus, drives the produced images lack of fidelity in details. In a comparison, CycleGAN performs so aggressive that some details in the original images have been undesirably modified (Such negative effect is highlighted in red boxes). MMI-ALI successfully avoids the problem CycleGAN and StarGAN suffer from. Table.3 shows the consistent quantitative results.

Table 4. FID in cross-3-domain transfer in Cityscape

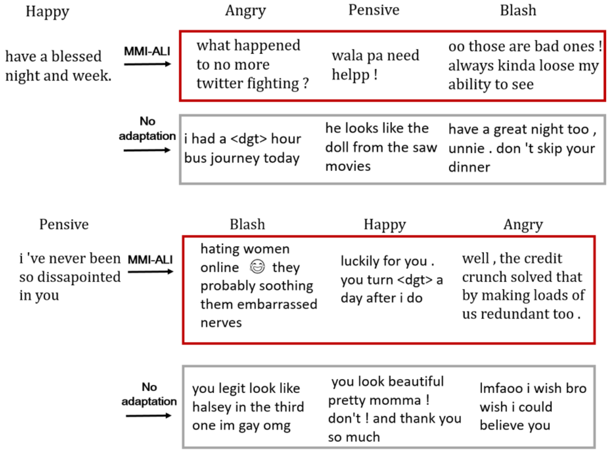
		R→Seg	Seg→R	R→Ske	Ske→R	Seg→Ske	Ske→Seg
Unsuper	ST	405.16	372.59	385.08	388.97	357.19	417.39
	CG	224.04	**213.43**	164.65	**222.24**	**60.20**	**144.07**
	Ours	**202.93**	254.41	**150.98**	246.04	101.30	192.13
Super	ST	382.90	440.53	419.11	383.72	400.70	299.82
	CG	**217.28**	260.41	**171.04**	**223.43**	65.18	228.61
	Ours	250.48	**246.01**	196.06	229.45	**55.76**	**143.20**

In Cityscape, MMI-ALI achieved the leg-and-leg performances with CycleGAN in FID in supervised and unsupervised learning (Table 4). But CycleGAN gets less benefits from supervision. They significantly outperformed StarGAN. As observed in Fig 7 8, when MMI-ALI is compared with the target generation groundtruth, it has achieved superior transfers so that avoided modeling C_m^2 generators.

3.4. Cross-m-emotion text style transfer.

In final experiment, we conduct a emotion style transfer in a text semantic embedding space. Specifically, we employ MojiTalk dataset (Zhou & Wang, 2017) that contains 64 emojis, and we collect a part of them to construct 4 domains related to 'Happy' (40000 entries), 'Angry' (29000 entries), 'Pensive' (14000 entries) and 'Abash' (6261 entries), respectively. In this scenario, the goal of MMI-ALI is to transform the emotional text embeddings (we choose skip-thought (Kiros et al., 2015) as our language model to extract the representation of each text in the domains) from one domain to the others.

Evaluation. Due to the embedding space is substantially discrete, the aforementioned metrics are not appropriate to evaluate the transfer efficiency. In this way, we employ a famous MRR (Mean Reciprocal Rank, (Craswell, 2009)), to measure the emotion transfer quality. For instance, when MMI-ALI transfer "happy" into "angry", we sort all sentences' embeddings based on their cosine distance to the embeddings generated from MMI-ALI. Then we calculate the rank of the nearest "angry" embedding and use its average of all transfer score. We use a simple fully-connected network with ReLU as the base backbone of MMI-ALI and train it with Batch normalization (BN). We compare MMI-ALI with the no-adaptation groundtruth results and the state-of-the-art unaligned text style transfer model (Shen et al., 2017) that trained by the official code .The results are shown in Table.5. We provide more visualization by retrieving the nearest neighbor of each target domain, for the embeddings before (*no adaptation*) and after MMI-ALI transform (Fig.9). As can be observed, the transferred embeddings (outputs of MMI-ALI) leads to the neighbor embeddings with the texts containing more significant emotion.



have a blessed night and week.

what happened to no more twitter fighting ?

wala pa need helpp !

oo those are bad ones ! always kinda loose my ability to see

i had a <dgt> hour bus journey today

he looks like the doll from the saw movies

have a great night too , unnie . don 't skip your dinner

i 've never been so dissapointed in you

hating women online 😊 they probably soothing them embarrassed nerves

luckily for you . you turn <dgt> a day after i do

well , the credit crunch solved that by making loads of us redundant too .

you legit look like halsey in the third one im gay omg

you look beautiful pretty momma ! don't ! and thank you so much

lmfaoo i wish bro wish i could believe you

Figure 9. The illustration of emotion style transfer in skipthough embedding space. We compare our MMI-ALI with no adaptation.

Table 5. MRR for each domain transfer evaluation. Higher is better. As can be seen, MRRs in "Happy" and "Abush" are even higher than the original domain, indicating the effectiness of MMI-ALI.

	Happy	Angry	Pensive	Abash
groundtruth	0.71	0.41	0.53	0.21
(Shen et al., 2017)	0.52	0.17	0.31	0.07
MMI-ALI	1.0	0.40	0.27	0.24

4. Conclusion

In this paper, we have delved into the problem of multiple domain joint distribution matching that summarized a variety of cross-domain generation tasks. Instead of hacking a complex DGM pipeline, we propose MMI-ALI, which reshapes classical ALI from the perspective of model integration and is linearly-scalable with the domain number. It learns with an adversarial ensemble loss and can be applied in both supervised and unsupervised learning schemes. Extensive evaluation results on diverse m-domain scenarios have demonstrated the superiority of the proposed framework to the existing DGMs feasible for cross-m-domain generation, e.g., CycleGAN and Star-GAN.

Acknowledgement

This work was supported in part by the National Key Research and Development Program of China under Grant No. 2018YFC0830103 and Grant No.2016YFB1001004, in part by National High Level Talents Special Support Plan (Ten Thousand Talents Program), and in part by National Natural Science Foundation of China (NSFC) under Grant No. 61622214, 61836012, and 61876224. Also, we thank Pengxu Wei for her valuable comments of this manuscript.

References

Belghazi, M. I., Rajeswar, S., Mastropietro, O., Rostamzadeh, N., Mitrovic, J., and Courville, A. Hierarchical adversarially learned inference. 2018.

Bell, A. J. The co-information lattice. In *Proceedings of the Fifth International Workshop on Independent Component Analysis and Blind Signal Separation: ICA*, volume 2003, 2003.

Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., and Choo, J. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. *arXiv preprint arXiv:1711.09020*, 2017.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Craswell, N. Mean reciprocal rank. 2009.

Deng, Z., Zhang, H., Liang, X., Yang, L., Xu, S., Zhu, J., and Xing, E. P. Structured generative adversarial networks. In *Advances in Neural Information Processing Systems*, pp. 3902–3912, 2017.

Donahue, J., Krähenbühl, P., and Darrell, T. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*, 2016.

Dong, H., Liang, X., Gong, K., Lai, H., Zhu, J., and Yin, J. Soft-gated warping-gan for pose-guided person image synthesis. 2018.

Dumoulin, V., Belghazi, I., Poole, B., Mastropietro, O., Lamb, A., Arjovsky, M., and Courville, A. Adversarially learned inference. *arXiv preprint arXiv:1606.00704*, 2016.

Gan, Z., Chen, L., Wang, W., Pu, Y., Zhang, Y., Liu, H., Li, C., and Carin, L. Triangle generative adversarial networks. In *Advances in Neural Information Processing Systems*, pp. 5253–5262, 2017.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. 2015.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. 2018.

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

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. Image-to-image translation with conditional adversarial networks. *arXiv preprint*, 2017.

Kameoka, H., Kaneko, T., Kou, T., and Hojo, N. Stargan-vc: Non-parallel many-to-many voice conversion with star generative adversarial networks. 2018.

Khrulkov, V. and Oseledets, I. Geometry score: A method for comparing generative adversarial networks. 2018.

Kim, T., Cha, M., Kim, H., Lee, J., and Kim, J. Learning to discover cross-domain relations with generative adversarial networks. *arXiv preprint arXiv:1703.05192*, 2017.

Kiros, R., Zhu, Y., Salakhutdinov, R., Zemel, R. S., Torralba, A., Urtasun, R., and Fidler, S. Skip-thought vectors. *Advances in Neural Information Processing Systems*, 28, 2015.

Li, C., Liu, H., Chen, C., Pu, Y., Chen, L., Henao, R., and Carin, L. Alice: Towards understanding adversarial learning for joint distribution matching. In *Advances in Neural Information Processing Systems*, pp. 5501–5509, 2017.

Mcgill, W. J. Multivariate information transmission. *Transactions of the Ire Professional Group on Information Theory*, 4(4):93–111, 2003.

Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., and Efros, A. A. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2536–2544, 2016.

Polikar, R. Ensemble learning. *Scholarpedia*, 4(1):1–34, 2009.

Pu, Y., Dai, S., Gan, Z., Wang, W., Wang, G., Zhang, Y., Henao, R., and Carin, L. Jointgan: Multi-domain joint distribution learning with generative adversarial nets. 2018.

Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., and Lee, H. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, 2016a.

Reed, S. E., Akata, Z., Mohan, S., Tenka, S., Schiele, B., and Lee, H. Learning what and where to draw. In *Advances in Neural Information Processing Systems*, pp. 217–225, 2016b.

Sabour, S., Frosst, N., and Hinton, G. E. Dynamic routing between capsules. 2017.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training gans. 2016.

Shen, T., Lei, T., Barzilay, R., and Jaakkola, T. Style transfer from non-parallel text by cross-alignment. 2017.

Tolstikhin, I., Bousquet, O., Gelly, S., and Schoelkopf, B. Wasserstein auto-encoders. 2017.

Ulyanov, D., Vedaldi, A., and Lempitsky, V. It takes (only) two: Adversarial generator-encoder networks. 2017.

Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Liu, G., Tao, A., Kautz, J., and Catanzaro, B. Video-to-video synthesis. *arXiv preprint arXiv:1808.06601*, 2018.

Xie, S. and Tu, Z. Holistically-nested edge detection. *International Journal of Computer Vision*, 125(1-3):3–18, 2015.

Yi, Z., Zhang, H., Tan, P., and Gong, M. Dualgan: Unsupervised dual learning for image-to-image translation.

Zhao, B., Chang, B., Jie, Z., and Sigal, L. Modular generative adversarial networks. 2018.

Zhou, W., Alan Conrad, B., Hamid Rahim, S., and Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. 13(4):600–612, 2004.

Zhou, X. and Wang, W. Y. Mojitalk: Generating emotional responses at scale. 2017.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593*, 2017.