Multivariate-Information Adversarial Ensemble for Scalable Joint Distribution Matching

Ziliang Chen*, Zhanfu Yang*, Xiaoxi Wang*, Xiaodan Liang, Xiaopeng Yan,
Guanbin Li, Liang Lin
Sun Yat-sen University, Purdue University

Motivation

Implicit Generative Models (IGM), e.g., GAN [1] and CycleGAN [2], boil down to an m-domain joint distribution matching (JDM):

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

[1] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, Generative Adversarial Nets. Proceedings Neural Information Processing Systems Conference, 2014

[2] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. ArXiv, 2017

Motivation

Implicit Generative Models (IGM), e.g., GAN [1] and CycleGAN [2], boil down to an m-domain joint distribution matching (JDM):

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

However, if $m > 2$,

[1] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, Generative Adversarial Nets. Proceedings Neural Information Processing Systems Conference, 2014

[2] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. ArXiv, 2017

Motivation

Implicit Generative Models (IGM), e.g., GAN [1] and CycleGAN [2], boil down to an m-domain joint distribution matching (JDM):

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

However, if $m > 2$,

Solution 1: CycleGAN, JointGAN[3]

They suffer combinatorial explosion in their parameters

Solution 2: StarGAN [4] and its variants

The domain-shared model lacks theoretical support, fragile in model collapse

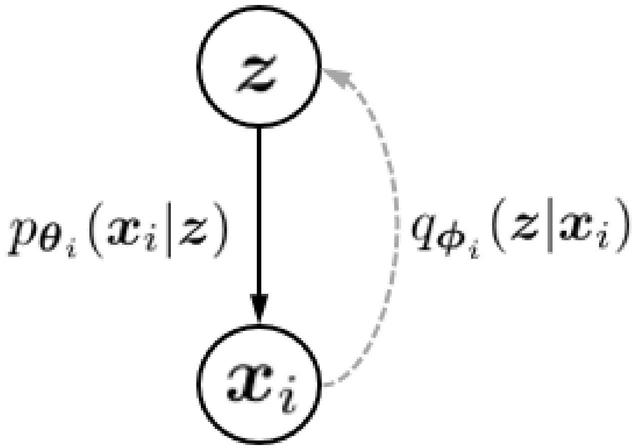
[1] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, Generative Adversarial Nets. Proceedings Neural Information Processing Systems Conference, 2014

[2] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. ArXiv, 2017

[3] Pu Y, Dai S, Gan Z, et al. JointGAN: Multi-Domain Joint Distribution Learning with Generative Adversarial Nets ICML. 2018.

[4] Choi Y, Choi M, Kim M, et al. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. CVPR 2018.

Solution 3: ALI Ensemble



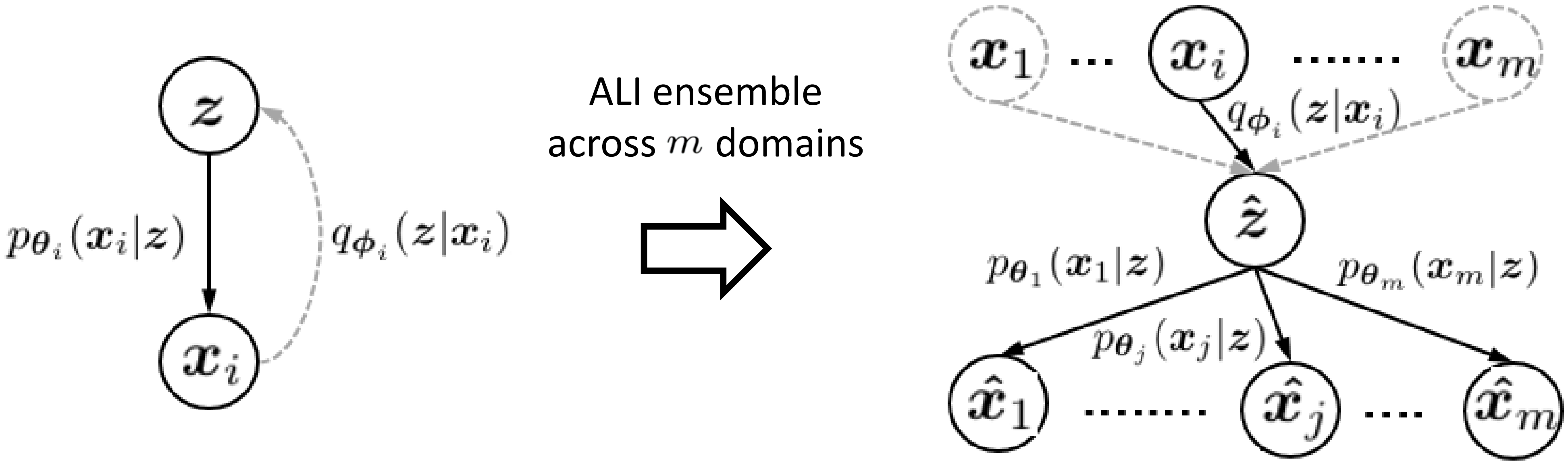
Adversarially **L**earned **I**nference (**ALI**) Model [5]

$$\min_{\boldsymbol{\theta_i},\boldsymbol{\phi}_i} \max_{\boldsymbol{\omega_i}} \mathcal{L}^{(i)}_{\mathrm{ALI}}(\boldsymbol{\theta}_i, \boldsymbol{\phi}_i, \boldsymbol{\omega}_i) = \mathbb{E}_{\boldsymbol{x}_i \sim p(\boldsymbol{x}_i), \hat{\boldsymbol{z}} \sim q_{\boldsymbol{\phi}_i}(\hat{\boldsymbol{z}}|\boldsymbol{x}_i)}\big[\log f_{\boldsymbol{\omega}_i}(\boldsymbol{x}_i, \hat{\boldsymbol{z}})\big] + \mathbb{E}_{\hat{\boldsymbol{x}}_i \sim p_{\boldsymbol{\theta}_i}(\hat{\boldsymbol{x}}_i|\boldsymbol{z}), z \sim q(\boldsymbol{z})}\big[\log\big(1 - f_{\boldsymbol{\omega}_i}(\hat{\boldsymbol{x}}_i, \boldsymbol{z})\big)\big]$$

Lemma 1 ((Dumoulin et al., 2016))**.** *The optimal generation, inference and critic nets w.r.t.,* $\{\boldsymbol{\theta}^*_i, \boldsymbol{\phi}^*_i, \boldsymbol{\omega}^*_i\}$ $(\forall i \in [m])$ *refer to a saddle point in Eq.2* $\iff$ $p_{\boldsymbol{\theta}^*_i}(\boldsymbol{x}_i|\boldsymbol{z})q(\boldsymbol{z}) = q_{\boldsymbol{\phi}^*_i}(\boldsymbol{z}|\boldsymbol{x}_i)p_i(\boldsymbol{x}_i)$.

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

Solution 3: ALI Ensemble

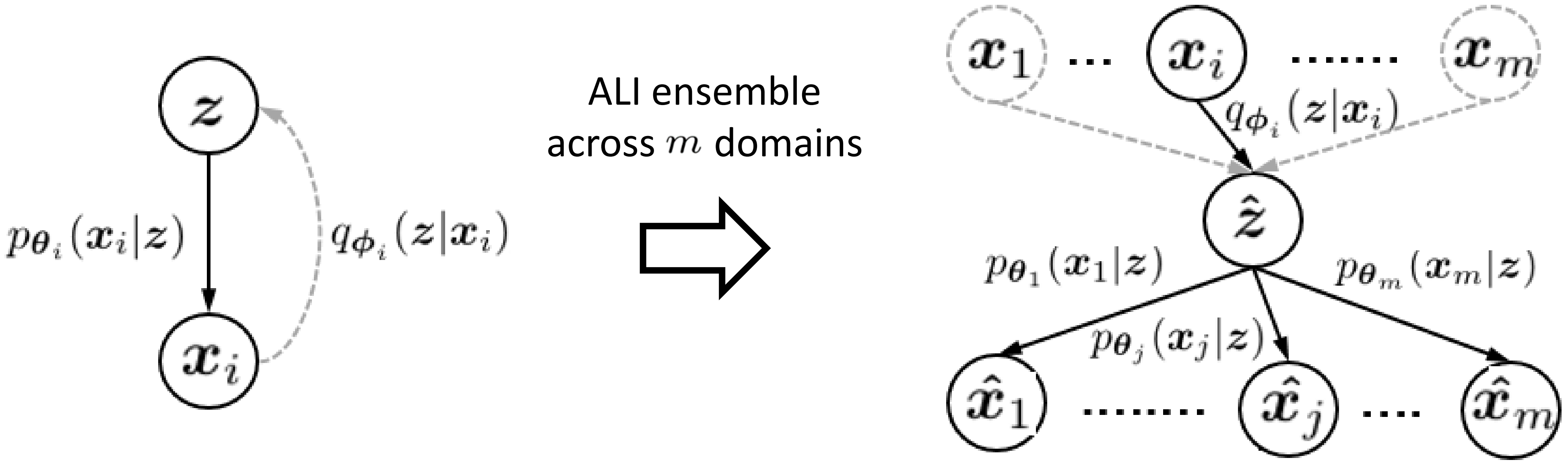


Advantages: (1). Linear-parameter scalability as m increases.

(2). Generative model capability: **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*

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

Solution 3: ALI Ensemble



Advantages: (1). Linear-parameter scalability as m increases.

(2). Generative model capability:

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 $p_{\Phi,\Theta}(\hat{x}_j|x_i)$ that satisfies*

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

Marginal matching do not imply joint density matching !

JDM Criteria

In supervised learning, data are drawn from $p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_m)$, each of them presenting as m-tuple. So the criterion can be written as

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

In unsupervised learning, no access is provided to draw m-tuple from $p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_m)$. Extending the observation from [6], the criterion is to minimize the conditional entropy

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

[6] 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.

Multivariate Mutual Information

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

Multivariate Mutual Information:
$$I(\boldsymbol{y}_1;\cdots;\boldsymbol{y}_n) \\ := I(\boldsymbol{y}_1;\cdots;\boldsymbol{y}_{n-1}) - I(\boldsymbol{y}_1;\cdots;\boldsymbol{y}_{n-1}|\boldsymbol{y}_n)$$

Our method aims to achieve:
$$\min_{\boldsymbol{\Phi},\boldsymbol{\Theta}} \quad -\sum_{i,j\in[m],i\neq j} I_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\boldsymbol{x}_i;\boldsymbol{x}_j;\boldsymbol{z})$$

Multivariate Mutual Information Leads To JDM

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

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

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

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

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

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

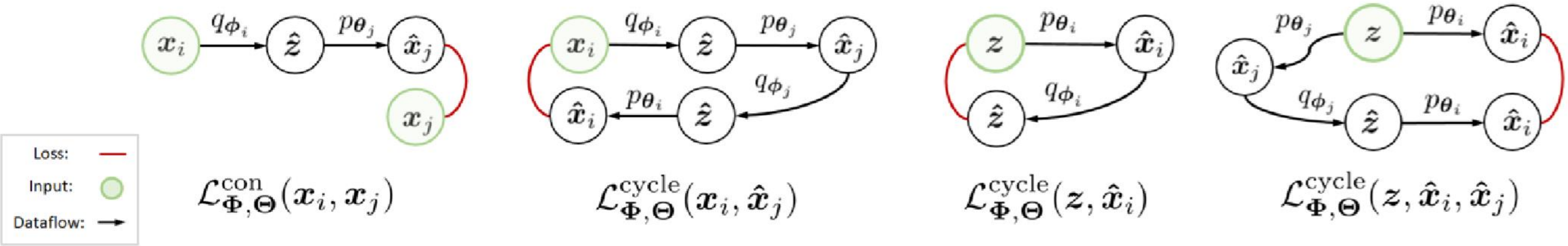
$\mathcal{L}^{\mathrm{cycle}}_{\boldsymbol{\Phi},\boldsymbol{\Theta}}(\boldsymbol{x}_i,\hat{\boldsymbol{x}}_j)$ is constructed as illustrated in Fig.2.b.

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

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

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

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



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

Multivariate Mutual Information Leads To JDM

$$\mathcal{R}_{\mathrm{SL}}(\mathbf{\Theta}, \mathbf{\Phi}) = \sum_{i,j \in [m], i \neq j} \mathcal{L}^{\mathrm{con}}_{\mathbf{\Phi},\mathbf{\Theta}}(\boldsymbol{x}_i, \boldsymbol{x}_j) + \mathcal{L}^{\mathrm{cycle}}_{\mathbf{\Phi},\mathbf{\Theta}}(\boldsymbol{z}, \hat{\boldsymbol{x}}_i) + \mathcal{L}^{\mathrm{cycle}}_{\mathbf{\Phi},\mathbf{\Theta}}(\boldsymbol{z}, \hat{\boldsymbol{x}}_i, \hat{\boldsymbol{x}}_j)$$

$$\mathcal{R}_{\mathrm{UL}}(\mathbf{\Theta}, \mathbf{\Phi}) = \sum_{i,j \in [m], i \neq j} \mathcal{L}^{\mathrm{cycle}}_{\mathbf{\Phi},\mathbf{\Theta}}(\boldsymbol{x}_i, \hat{\boldsymbol{x}}_j) + \mathcal{L}^{\mathrm{cycle}}_{\mathbf{\Phi},\mathbf{\Theta}}(\boldsymbol{z}, \hat{\boldsymbol{x}}_i) + \mathcal{L}^{\mathrm{cycle}}_{\mathbf{\Phi},\mathbf{\Theta}}(\boldsymbol{z}, \hat{\boldsymbol{x}}_i, \hat{\boldsymbol{x}}_j)$$

Theorem 1. *Suppose that true and parameterized domain marginal distributions maintain a high likelihood to domain variables,* $\mathcal{R}_{\mathrm{SL}} \to 0$ *leads to the optima in* $\min_{\mathbf{\Phi},\mathbf{\Theta}} \; -\mathbb{E}_p\big[\log p_{\mathbf{\Phi},\mathbf{\Theta}}(\{\boldsymbol{x}_i\}_{i=1}^m)\big]$

$\mathcal{R}_{\mathrm{UL}} \to 0$ *leads to the optima in* $\min_{\mathbf{\Phi},\mathbf{\Theta}} H\big(\boldsymbol{x}_i | \{\hat{\boldsymbol{x}}_j\}_{j \in [m] \& j \neq i}\big)$

Adversarial Ensemble Learning

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

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

where $\mathcal{R}_{\text{SL}}/\mathcal{R}_{\text{UL}}$ are switched by supervised/unsupervised learning and $\beta > 0$ denotes the loss-balance factor.

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

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

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

Experiments

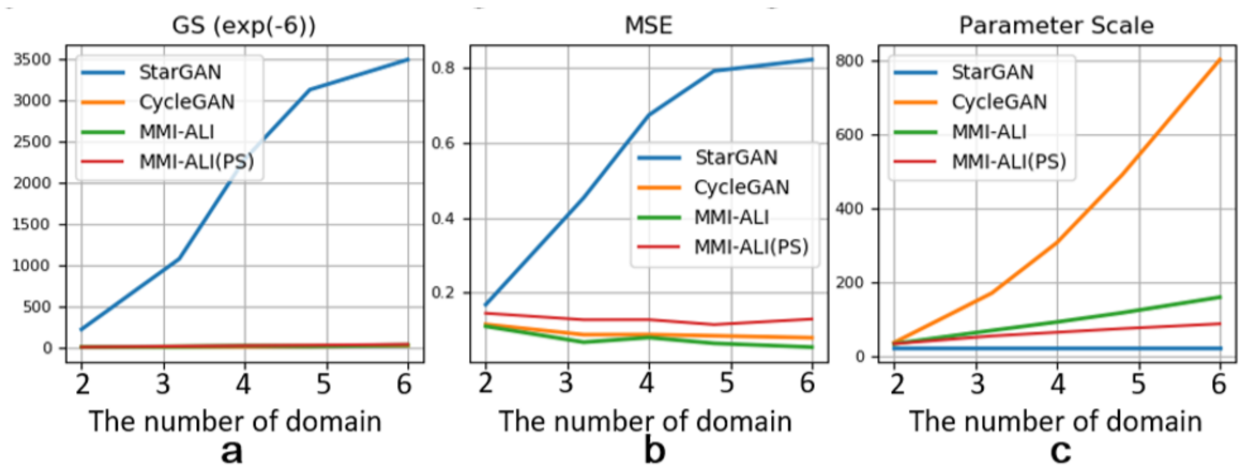
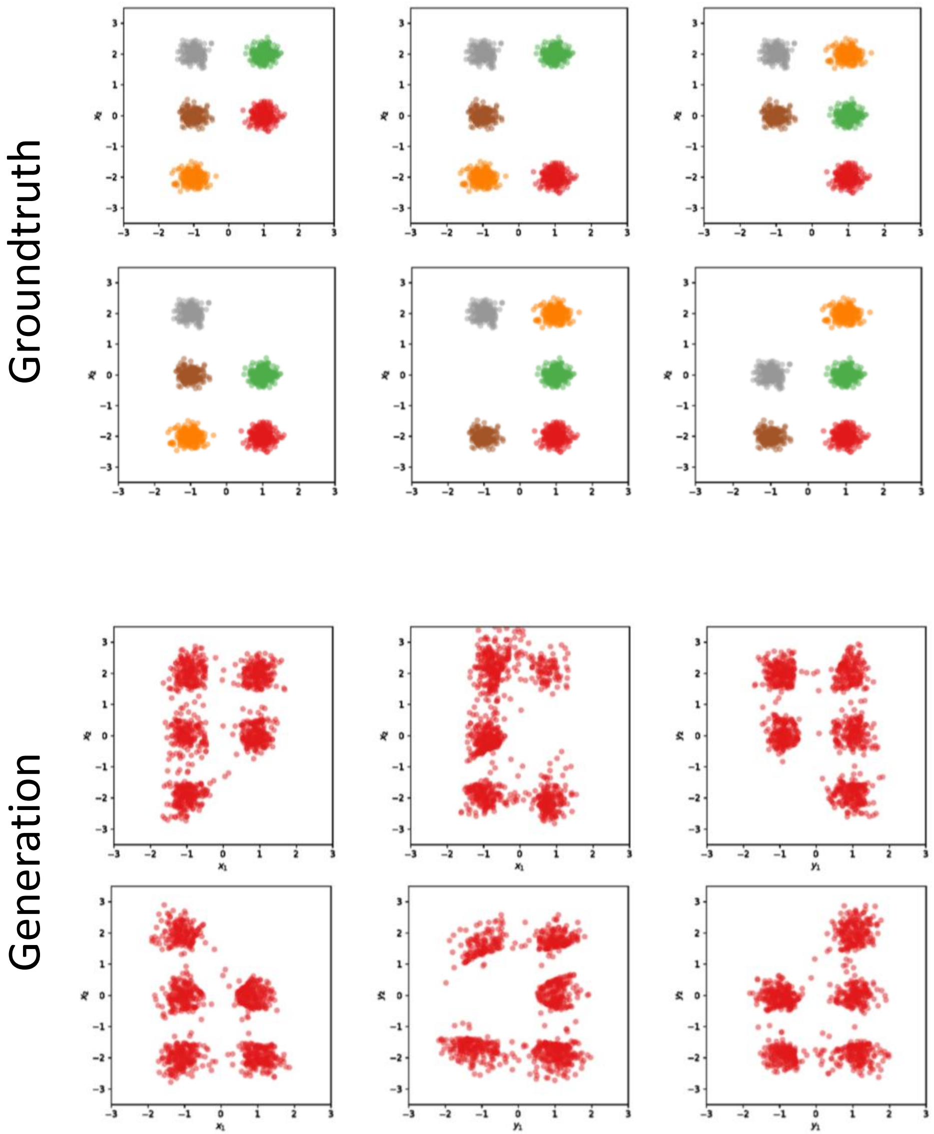
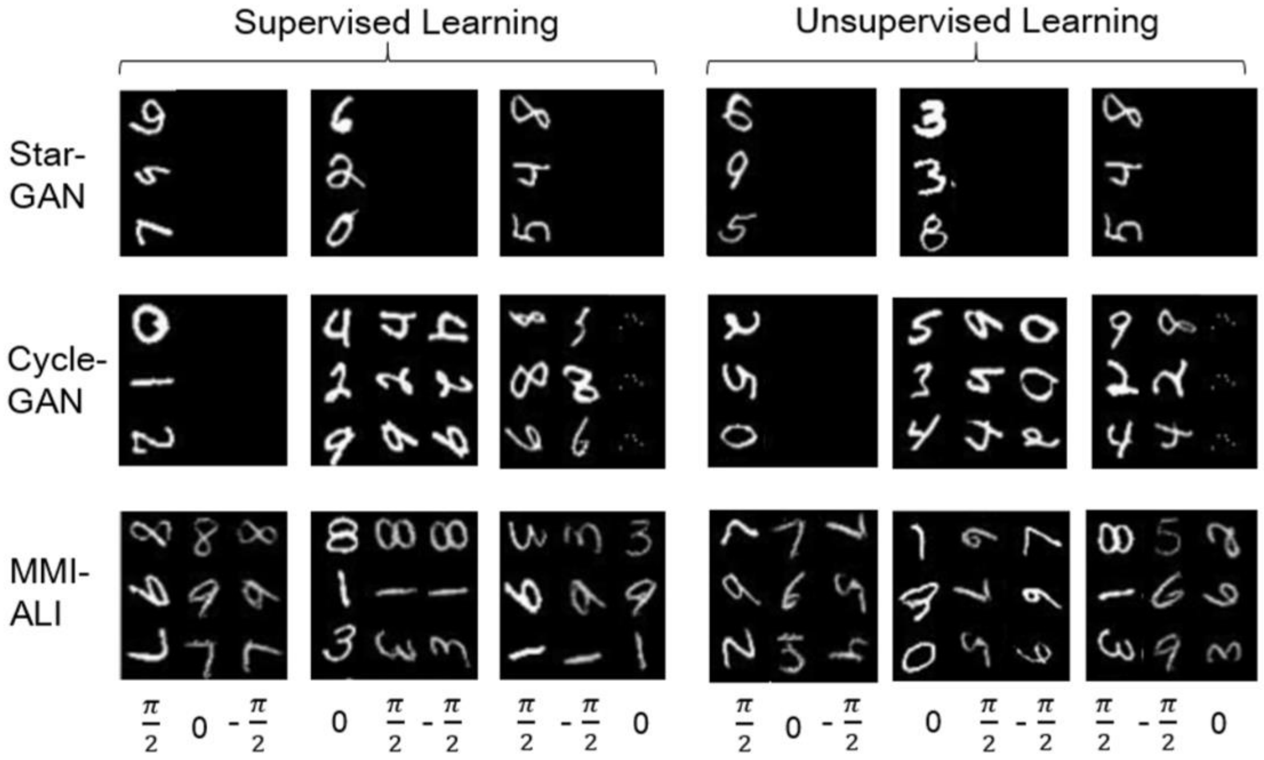


Figure 4. Transfer evaluations with 2∼6 synthetic domains: (a). Geometric Score (GS, lower is better); (b). Mean Square Error (MSE, lower is better); (c). Parameter Scale (lower is better).

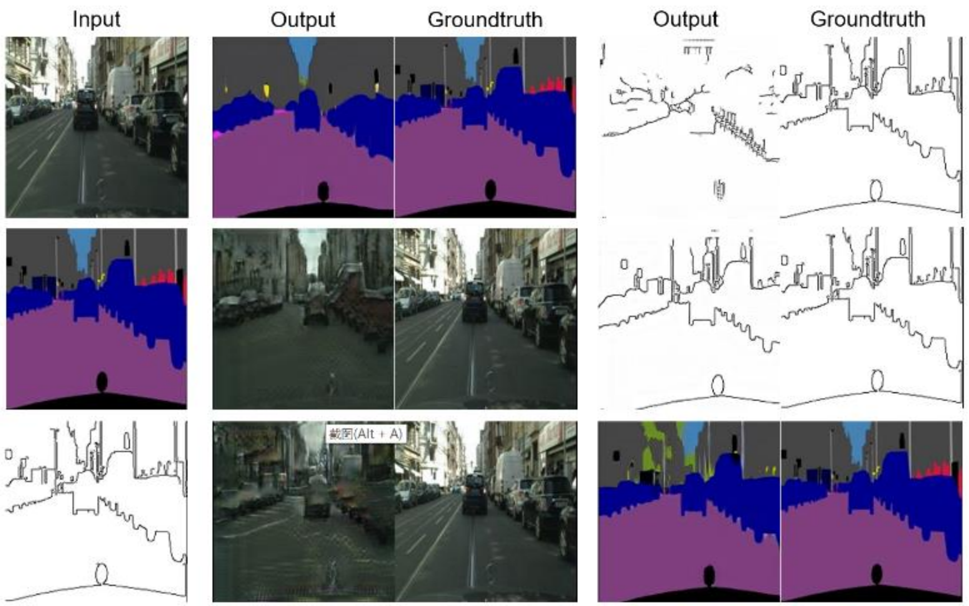


Experiments

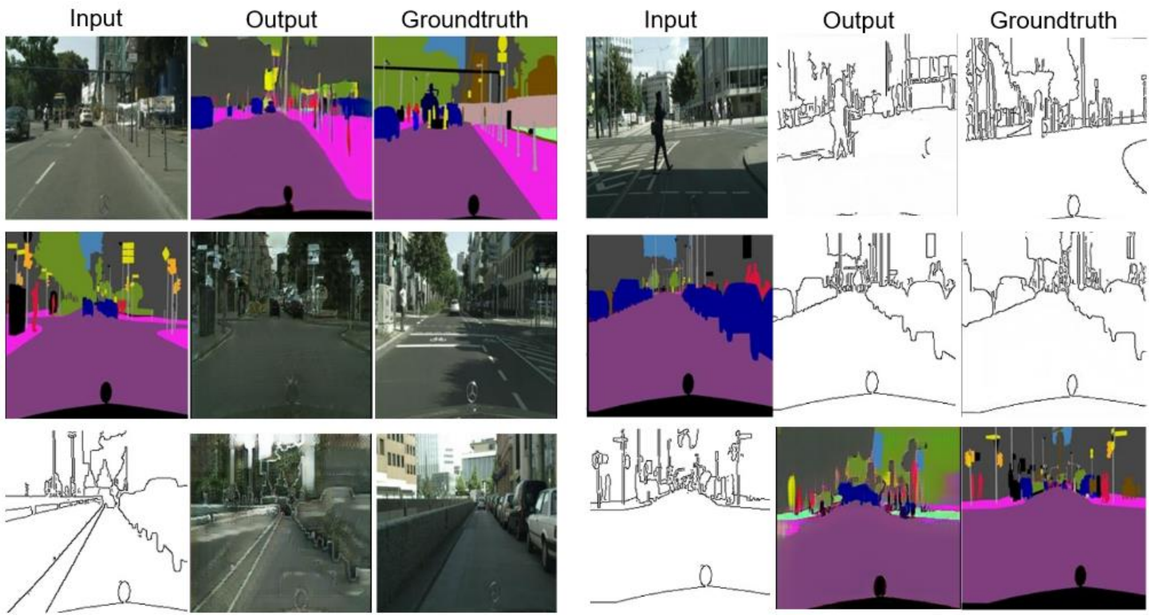


	$-\frac{\pi}{2}\to 0$	$\frac{\pi}{2}\to 0$	$0\to\frac{\pi}{2}$	$-\frac{\pi}{2}\to\frac{\pi}{2}$	$-\frac{\pi}{2}\to 0$	$\frac{\pi}{2}\to-\frac{\pi}{2}$
StarGAN	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
CycleGAN	8.34±0.12	6.13±0.37	2.25±0.02	2.38±0.06	1.71±0.03	1.04±0.01
MMI-ALI (wo MMI)	7.48±0.05	6.06±0.10	3.19±0.04	2.90±0.05	2.73±0.09	2.47±0.12
MMI-ALI ($\gamma = 0$)	8.34±0.20	8.27±0.10	**3.26**±0.06	3.06±0.10	3.15±0.11	2.92±0.12
MMI-ALI	**8.99**±0.06	**9.01**±0.00	2.95±0.08	**3.86**±0.12	**3.31**±0.12	**3.08**±0.05

Experiments



Supervised Learning

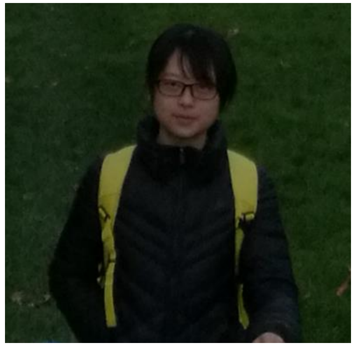


Unsupervised Learning

		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**

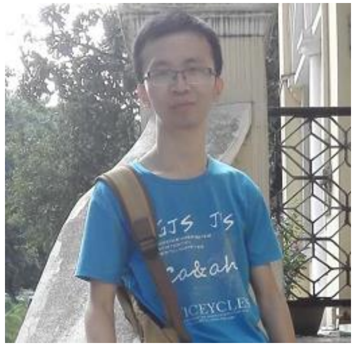
Collaborators

Zhanfu Yang



Xiaoxi Wang

Xiaodan Liang



Xiaopeng Yan

Guanbin Li

Liang Lin

Thank You!