A Geometrical Perspective on Image Style Transfer with Adversarial Learning

Xudong Pan, Mi Zhang, Daizong Ding and Min Yang

Abstract—Recent years witness the booming trend of applying Generative Adversarial Nets (GAN) and its variants to *image style transfer*. Although many reported results strongly demonstrate the power of GAN on this task, there is still little known about neither the interpretations of several fundamental phenomenons of image style transfer by generative adversarial learning, nor its underlying mechanism. To bridge this gap, this paper presents a general framework for analyzing style transfer with adversarial learning through the lens of differential geometry. To demonstrate the utility of our proposed framework, we provide an in-depth analysis of Isola et al.'s pioneering style transfer model *pix2pix* [1] and reach a comprehensive interpretation on their major experimental phenomena. Furthermore, we extend the notion of generalization to conditional GAN and derive a condition to control the generalization capability of the pix2pix model. From a higher viewpoint, we further prove a learning-free condition to guarantee the existence of infinitely many perfect style transfer mappings. Besides, we also provide a number of practical suggestions on model design and dataset construction based on these derived theoretical results to facilitate further researches.

Index Terms—Generative Adversarial Learning, Unsupervised Learning Theory, Generalization Theory, Machine Learning

1 Introduction

Generative Adversarial Nets (GAN), from its first proposal by Goodfellow et al. [2], remain one of the most popular paradigms in generative modeling of unknown distributions, fostering a wide range of applications in image generation [3], text generation [4], speech synthesis [5] and numerously more [6], [7], [8]. Roughly speaking, the general idea behind GAN and its variants is intuitive: It aims at learning a mapping from a source distribution, e.g., a Gaussian distribution in the vanilla GAN [2] or an unknown distribution of *labels* in its conditional variant [9], to a target data distribution (e.g., a collection of facial images). By solving the min-max game [10] between a *generator* (i.e., a learning model which maps data from the source domain to the target one) and a *discriminator* (i.e., a learning model which distinguishes a generated sample from the target distribution), the adversarial learning process finally learns a realistic distribution for downstream generative tasks [11].

As a major use case of GAN and its variants, *image style transfer* has been intensively studied in the recent few years under the adversarial learning paradigm [1], [12], [13], [14], [15]. Image style transfer, as a generic name for various specific tasks in image processing, includes tasks such as facial expression transfer (e.g., poker face $\rightarrow$ smiley face), artistic style transfer (e.g., realism $\rightarrow$ impressionism) and many more. In general, image style transfer aims at processing an image from a source collection to make it indistinguishable among a target collection of images. Although a number of models and methods exist in previous literature of image processing [16], [17], [18], [19], the first successful attempt to leverage the power of GAN on image style transfer, ought to be attributed to the pioneering work of Isola et al. [1] (aka. *pix2pix*), which aroused a surge of research interests on this topic (e.g., [12], [13]). Formally, most of the existing adversarial learning based methods solve the optimization problem below

$$\overbrace{\min_{G} \underbrace{\mathcal{L}_{\text{adv}}(G)}_{\text{adversarial loss}} + \lambda \underbrace{\mathcal{L}_{\text{id}}(G)}_{\text{identity loss}}}^{\text{style transfer with adversarial learning}} \tag{1}$$

Here, the former term $\mathcal{L}_{\text{adv}}(G)$ is called the *adversarial loss*, which follows the general idea of conditional GAN [9] and is usually implemented with the Wasserstein variant in practice [20] to alleviate the instability of the vanilla GAN loss [21]. Formally, the term displays as

$$\mathcal{L}_{\text{adv}}(G) = \max_{\|D\|_L \leq 1} \mathbb{E}_{q\sim p_r}[D(q)] - \mathbb{E}_{p\sim p_g}[D(G(p))] \tag{2}$$

where p_r, p_g denote respectively the distribution of images over the source and the target collections, G is the generator, and the scalar-valued mapping D (i.e., the discriminator) is required to satisfy Lipschitz continuity with constant 1.

The latter term $\mathcal{L}_{\text{id}}(G)$ is called the *identity loss*, an original creation in the pix2pix model for utilizing the available paired images as additional supervision for the adversarial learning process. Formally, the identity loss was implemented as ℓ_1-loss in [1] between the transferred image and the ground-truth

$$\mathcal{L}_{\text{id}}(G) = \mathbb{E}_{x,y\sim p_r(x,y)}[\|y - G(x)\|_1] \tag{3}$$

where $p_r(x, y)$ denotes the distribution of *paired images* (e.g., in facial expression transfer, one's poker face and the ground-truth smiley face). Noticeably, several alternative implementations were also proposed in the following works, such as the cycle consistency loss [12].

Despite the empirical success of the above adversarial learning paradigm for image style transfer, the following

• *The authors are with the School of Computer Science, Fudan University, Shanghai 200433, China.*
Corresponding authors: Mi Zhang and Min Yang
E-mail: {xdpan18, mi_zhang, dzding17, m_yang}@fudan.edu.cn

two experimental phenomena of pix2pix model, which we suggest to be highly connected with the foundations of GAN and adversarial learning, are still far from well-studied and fully-interpreted in the existing literature.

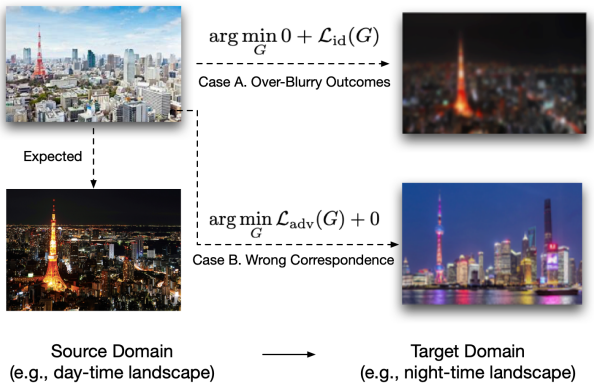

Fig. 1. An illustration of two noteworthy yet theoretically unclear phenomena first reported in the original work of pix2pix.

- **Case A: Blurry versus Sharp.** Omitting the adversarial loss, i.e., solving $\mathcal{L}_{\text{id}}(G)$ alone, *will lead to reasonable but blurry results*. In this case, although the learned style transfer mapping G generates relevant target images, most details are hard to be recognized. As illustrated in the upper part of Fig. 1, the generated night-time counterpart seems correct in correspondence to the source image but is overly blurry.
- **Case B: Source of Artifacts.** Omitting the identity loss by setting the regularization factor λ to 0 would *produce much sharper results yet cause some artifacts*. As illustrated in the lower part of Fig. 1, in this case, although the learned style transfer mapping G generates realistic images with recognizable details, the content of the generated image however is kind of irrelevant with the given source image.

Although there do exist theoretical studies on analyzing and improving the training dynamics and generalization capability of GAN [11], [21], [22], [23], there is rarely applicable theoretical results for analyzing conditional GAN, thus for pix2pix and its empirical results. The inappropriateness mainly comes from Eq. (2), where the model generates fake images directly from a given image of intensively high dimension [24], instead of a low-dimensional Gaussian noise in GAN. In fact, the simple violation of the low-dimensional assumption would immediately invalidate most of the previously obtained theoretical results for GAN. Considering the worthiness of obtaining reasonable theoretical interpretations as guidance for further researches, we formulate this non-standard model from a geometrical perspective, propose an extended definition of generalization for conditional GAN and eventually reach a number of inspiring theoretical results.

In this paper, to better understand the adversarial learning paradigm on style transfer, we first present a general framework for analyzing style transfer with adversarial learning with the aid of differential geometry [25]. Basically, we propose to equip the source and the target collections of images with Riemannian manifold structures and extend the notion of the discriminator and the generator to this setting. From this geometric perspective, the adversarial learning approach to style transfer can be viewed as learning an optimal transform from the source manifold to the target manifold with oracles on distributional and point-wise correspondence, respectively provided by the adversarial loss and the identity loss. To demonstrate the effectiveness of our proposed theoretical framework, we successfully **a)** attain a full picture on fundamental yet unclear experimental phenomena reported in [1], **b)** derive a quantitative condition on the generalization capacity of pix2pix model and **c)** prove a model-free condition to guarantee the existence of infinitely many perfect generators at large. Here, by *perfect*, we mean the generator G has its range covering the whole target domain, which therefore eliminates the potential mode collapse [26].

Noticeably, this work makes substantial extensions over our earlier work [27] in the following aspects.

- Our theoretical results presented in Section 6 are unpublished and, to the best of our knowledge, novel in the sense that it observes for the first time that, the C^∞-diffeomorphism between the source and the target image manifolds as a necessary and sufficient condition to guarantee the existence of infinitely many perfect generators.
- Additional discussions with well-presented illustrations are added in the first two parts for better readability.
- Previously omitted technical details are clarified to further strengthen the rigorousness of the statements.

In summary, we mainly make the following contributions.

- We present a geometric formulation of the standard paradigm for image style transfer with adversarial learning in the language of differential geometry (§2).
- We derive the equivalence between the adversarial loss part in the pix2pix model with a set of independent learning tasks between paired charts of the source and the target image manifolds (Thm. 3.1) and therefore provide full interpretations on several unclear empirical phenomena reported in previous works (§3.3 & §5.1)
- We extend the notion of generalization to conditional GAN (Def. 4.2) and derive a quantitative condition on the generalization of the pix2pix model (Thm. 5.1).
- As a substantial improvement over our earlier work [27], we establish the C^∞-diffeomorphism of the source and target image manifolds as a sufficient and necessary condition to the existence of infinitely many perfect generators for style transfer (Thm. 6.1 & 6.3).

2 Preliminaries

2.1 Image Spaces as Smooth Manifolds

In this paper, we mainly study the WGAN adversarial loss in its primal form [28] other than the Kantorovich-Rubinstein dual form in Eq. (2), which writes

$$\mathcal{L}_{\text{adv}}(G) = \inf_{\gamma \in \Pi(p_r, p_g)} \mathbb{E}_{(x,y)\sim\gamma}[\|G(x) - y\|] \tag{4}$$

where $\Pi(p_r, p_g)$ is the set of joint distributions for pairs of images (x, y) such that the marginal distributions are equal to p_r, p_g. Intuitively, the explicit term of discriminator in its dual form (Eq. (2)) can be considered to be replaced by the inner optimal transport task [28] implicitly in Eq. (4).

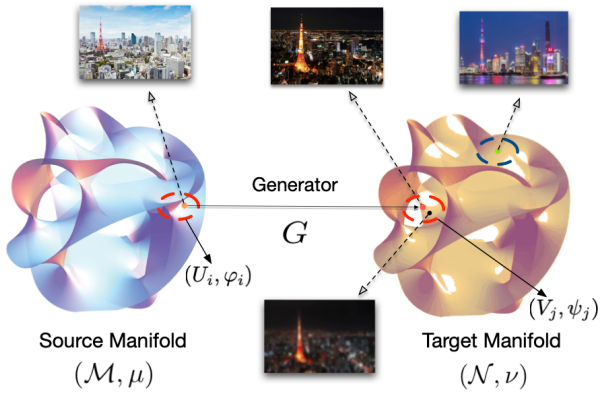

Fig. 2. Proposed geometric viewpoint on image style transfer with adversarial learning.

Without loss of generality, we focus on the image style transfer task from a source set of images $\mathcal{I}_S$ to a target set $\mathcal{I}_T$, with images of the same resolution $w \times h$. As is well-known, images can always be viewed as elements in certain ambient Euclidean space (e.g., for images encoded in the RGB scheme, it is $\mathbb{R}^{3\times w\times h}$). In fact, there also exists an intrinsic structure over the image set alongside the ambient space, as is validated by various empirical works previously [24], [29]. Such an intrinsic structure is usually formulated as a smooth manifold [21], [30]. For the basics of differential geometry, we recommend the readers to refer to standard texts (e.g., [31]).

In this work, we make a similar assumption as follows.

Assumption 2.1. *There exist d-dimensional regular* [1] *Riemannian manifolds $\mathcal{M}$, $\mathcal{N}$ embedded in $\mathbb{R}^{w\times h}$, with the constructed atlas as $\{(U_i, \varphi_i)\}_{i=1}^K$, $\{(V_j, \psi_j)\}_{j=1}^K$, respecting the pairwise disjointness property, i.e., $\forall i, j \in [K], U_i \cap U_j = \varnothing, V_i \cap V_j = \varnothing$ if $i \neq j$, such that $\mathcal{I}_S \subset \mathcal{M}$, $\mathcal{I}_T \subset \mathcal{N}$. ($[K]$ denotes the set $\{1, 2, \ldots, K\}$ and K a natural number)*

Intuitively, if we take the example of facial expression transfer (e.g., smiley face → poker face), each chart $U_i(V_i)$ of the underlying image manifold $\mathcal{M}(\mathcal{N})$ can be viewed intuitively as a set of images belonging to the same person. In this scenario, the pairwise disjointness condition can be naturally valid. Moreover, the assumption of equal dimensions contained above is only for the convenience of notation simplification. Results presented in the remainder of this paper can be directly extended to the situation when source and target image manifolds are of different dimensions.

2.2 Induced Probability Measures on Image Manifolds

With Assumption 2.1, we partition the image sets $I_S \subset \mathcal{M}$, $I_T \subset \mathcal{N}$ into finer subsets, which formally writes $\mathcal{I}_S = \cup_{k=1}^K \mathcal{I}_S^k$, $\mathcal{I}_T = \cup_{k=1}^K \mathcal{I}_T^k$, where $\forall k \in [K]$, $\mathcal{I}_S^k \doteq \{s_k^i\}_{i=1}^{m_k} \subset U_k$ and $\mathcal{I}_T^k \doteq \{t_k^i\}_{i=1}^{n_k} \subset V_k$.

In order to describe the relatedness of images from the same chart, the following assumption is imposed.

1. Throughout this paper, the following regularity conditions on Riemannian manifolds are required, (1) compactness (2) finiteness in sectional curvature. Note those assumptions are common in many Riemannian geometry texts (e.g. [32]) and realistic in practice.

Assumption 2.2. *For each $k \in [K]$, there exist absolutely continuous probability measures $\mu_k, \nu_k : \mathcal{B}(\mathbb{R}^d) \to [0, 1]$, supported on $\varphi_k(U_k)$ and $\psi_k(V_k)$ respectively, such that $\{\varphi_k(s_k^i)\}_{i=1}^{m_k} \overset{i.i.d.}{\sim} \mu_k$, $\{\psi_k(t_k^i)\}_{i=1}^{n_k} \overset{i.i.d.}{\sim} \nu_k$, where $\mathcal{B}(\mathbb{R}^d)$ denotes the Borel set over $\mathbb{R}^d$.*

Noteworthily, the probabilistic structure introduced in Assumption 2.2 also enhances the flexibility of Assumption 2.1. When the original images are not obviously divided to different objects as in the facial expression transfer case, one can indeed construct their own different atlas structures under the restriction of the pairwise disjointness, by, e.g., clustering similar images into one chart. In this case, it only requires us to redefine the imposed probabilistic measure μ_i to correspondingly adapt to the new atlas structure.

With probability measures defined on each chart [2], we utilize the following proposition to "glue" them together to induce a unified probability measure globally respectively over the underlying manifold structures $\mathcal{M}, \mathcal{N}$, denoted as μ, ν.

Proposition 2.1. *Given a smooth manifold $\mathcal{M} = \{(U_i, \varphi_i)\}_{i=1}^K$ with pairwise disjointness and $\{\mu_i\}_{i=1}^K$ as the probability measures supported on $\{\varphi_i(U_i)\}_{i=1}^K$ correspondingly, a function $\mu : \mathcal{B}(\mathcal{M}) \to [0, 1]$ is defined by*

$$d\mu(s) = \frac{1}{K} \sum_{i=1}^{K} \mathbf{1}_{s\in U_i} d\varphi_{i\#}\mu_i(s) \tag{5}$$

Then μ is a probability measure defined on $\mathcal{M}$. Here, $\varphi_{i\#}\mu_i := \mu_i \circ \varphi_i$ denotes the pull-back *of the probability measure μ_i onto the manifold $\mathcal{M}$, which is a common notion used in measure theory [33].*

Proof. Please see Appendix A.1. □

As a remark, although Eq. (5) contains a slight abuse of notations (consider if $s \notin U_i$, $\varphi_i(s)$ is not defined), it can be naturally resolved according to the pairwise disjointness in Assumption 2.1, that is, all except one $\mathbf{1}\{s \in U_i\}$ is non-vanishing for any $s \in \mathcal{M}$.

2.3 A Geometrical Formulation of Style Transfer with Adversarial Learning

As long as additional geometrical structure are imposed on image sets, the definition of generator and discriminator in the adversarial learning paradigm ought to vary correspondingly.

Generator. In our context, the generator should be redefined as a mapping between manifolds instead of between flat Euclidean spaces. Formally, we require the generator $G \in C^\infty(\mathcal{M}; \mathbb{R}^{w\times h})$, the set of smooth open mappings from Riemannian manifold $\mathcal{M}$ to $\mathbb{R}^{w\times h}$, which generally includes common implementations of generators due to the negligible measure of the set of discontinuity [34].

Discriminator. Within the manifold settings, the norm $\|\cdot\|$ ought to be imposed on the ambient Euclidean space. Specifically, we equip the ambient space $\mathbb{R}^{w\times h}$ of the target image manifold $\mathcal{N}$ with a general metric d (e.g., the Euclidean distance or general L_p metrics). Finally, we correspondingly

2. More precisely, probability measures are defined on its homeomorphism as $\mathbb{R}^d$.

reformulate the adversarial loss and the corresponding identity loss (cf. Eq. (4) & (3)), which serve as a departure for the subsequent analysis in the rest of this paper.

$$\mathcal{L}_{\text{adv}}(G) = \inf_{\gamma\in\Pi(\mu,\nu)} \mathbb{E}_{(s,t)\sim\gamma} d(G(s),t) \tag{6}$$

$$\mathcal{L}_{\text{id}}(G) = \mathbb{E}_{s,t\sim p_r(s,t)} d(G(s),t) \tag{7}$$

where $\forall\gamma \in \Pi(\mu,\nu)$, the following constraints are required,

$$\int_{\mathcal{M}} \gamma(s,\cdot)d\mu = \nu, \int_{\mathcal{N}} \gamma(\cdot,t)d\nu = \mu \tag{8}$$

For compatibility with inner-relatedness in Assumption 2.2, we further assume $\forall i \neq j \in [K]$, $\forall x,y \in V_i, z \in V_j$, $d(x,y) \leq d(x,z)$. Intuitively, it requires images in one chart are more similar to each other than to outsiders, which was recently validated in [7]. Fig. 2 illustrates our proposed geometric viewpoint above.

2.4 Diffeomorphism of Riemannian Manifolds

Illustratively, C^∞-diffeomorphism condition is described as source image manifold $\mathcal{M}$ can be smoothly shaped into the target image manifold $\mathcal{N}$. Formally, we provide its rigorous definition as follows.

Definition 2.1 (C^∞-Diffeomorphism [35])**.** *Let $\mathcal{M}$ and $\mathcal{N}$ be Riemannian manifolds. A map $f : \mathcal{M} \to \mathcal{N}$ is a* homeomorphism *if it is continuous and has an inverse $f^{-1} : \mathcal{N} \to \mathcal{M}$ which is also continuous. Furthermore, we call f a C^∞-diffeomorphism between $\mathcal{M}$ and $\mathcal{N}$ if f, f^{-1} are smooth w.r.t. $\mathcal{M},\mathcal{N}$ respectively.*

Exchangeably, we call a Riemannian manifold $\mathcal{M}$ is C^∞-diffeomorphic with $\mathcal{N}$ if there exists a C^∞-diffeomorphism between $\mathcal{M}$ and $\mathcal{N}$, denoted as $\mathcal{M} \equiv \mathcal{N}$.

In the proof of our main results, the following theorem from Riemannian geometry is indispensable.

Theorem 2.1 (Gromov's Convergence Theorem [32])**.** *If $\{\mathcal{M}_i\}_{i=1}^\infty$, $\mathcal{N}$ are d-dimensional Riemannian manifolds s.t. $\lim_{i\to\infty} d_H(\mathcal{M}_i,\mathcal{N}) = 0$, then for a sufficiently large i, $\mathcal{M}_i \equiv \mathcal{N}$. Here, d_H is called Hausdorff distance between Riemannian manifolds, defined as*

$$d_H(\mathcal{M},\mathcal{N}) = \max\{\sup_{p\in\mathcal{M}} \inf_{q\in\mathcal{N}} d(p,q), \sup_{q\in\mathcal{N}} \inf_{p\in\mathcal{M}} d(p,q)\} \tag{9}$$

where d is an arbitrary metric function defined on the ambient space $\mathbb{R}^n$.

Intuitively, Gromov's theorem states, if a sequence of Riemannian manifolds converges in the sense of Hausdorff distance, then they are indeed C^∞-diffeomorphic to the limit manifold asymptotically.

3 Analysis and Interpretations of Pix2Pix

A widely recognized difficulty on analyzing adversarial learning process lies in the bilevel optimization problem [2] (here, specifically $\min_G \inf_{\gamma\in\Pi(\mu,\nu)}$). To resolve this obstacle, we prove in this section that, in our proposed framework above, the infimum term in Eq. (6) can be solved in a closed form, when non-trivial constraints are posed on the candidate set of generator G (Thm. 3.1). Furthermore, we observe that the derived closed-form solution decomposes the original learning task as a set of independent learning tasks on *paired charts* (i.e., a tuple of charts respectively of $\mathcal{M},\mathcal{N}$, such as (U_i,V_j)) and the pairing relations are uniquely determined by the candidate sets. Based on this result, we provide comprehensive interpretations fully for Source of Artifacts and partially for Blurry versus Sharp (§3.3).

3.1 An Equivalent Form of $\mathcal{L}_{\text{adv}}$

As a preparation, we give the explicit form of the probability measures μ, ν on manifolds as

$$d\mu(s) = \frac{1}{K}\sum_{i=1}^{K} \mathbf{1}_{s\in U_i} d\varphi_{i\#}\mu_i(s) \tag{10}$$

$$d\nu(t) = \frac{1}{K}\sum_{j=1}^{K} \mathbf{1}_{t\in V_j} d\psi_{j\#}\nu_j(t) \tag{11}$$

For simplicity, we write $d\tilde{\mu}_i := d\varphi_{i\#}\mu_i$ and $d\tilde{\nu}_i := d\psi_{i\#}\nu_i$, $\forall i \in [K]$.

We then expand $\mathcal{L}_{\text{adv}}$ in Eq. (6) with the pairwise disjointness property and obtain

$$\inf_{\gamma\in\Pi(\mu,\nu)} \mathbb{E}_{(s,t)\sim\gamma} \sum_{i=1}^{K}\sum_{j=1}^{K} \mathbf{1}_{s\in U_i}\mathbf{1}_{t\in V_j} d(G(s),t) \tag{12}$$

By exchanging the expectation operator with summations according to Fubini's theorem [36] and writing the expectation directly in the integral form, we have

$$\inf_{\gamma\in\Pi(\mu,\nu)} \sum_{i=1}^{K}\sum_{j=1}^{K} \int_{U_i}\int_{V_j} d(G(s),t)d\gamma(s,t) \tag{13}$$

With a similar technique adopted in Dai et al. [37], for every $\gamma \in \Pi(\mu,\nu)$, there exist a function $\Delta : \mathcal{N}\times\mathcal{N} \to \mathbb{R}^+ \cup \{0\}$ and $f_\gamma : \mathcal{M} \to \mathcal{N}$, satisfying

$$d\gamma(s,t) = d\gamma(t|s)d\mu(s) = \Delta(f_\gamma(s),t)d\mu(s)d\nu(t) \tag{14}$$

where Δ has an intuitive interpretation as a metric of dissimilarity between elements on the manifold $\mathcal{N}$, which is independent from the choice of path and compatible with inner-relatedness. In fact, with the following observations, we suggest it is proper to absorb the term $\Delta(f_\gamma(s),t)$ into $d(G(s),t)$.

a) Equivalence of optimization problems (without boundary condition) [38]
   - $\min_G \min_{f_\gamma} \Delta(f_\gamma(s),t)d(G(s),t)$
   - $\min_G \Delta(G(s),t)d(G(s),t)$

   considering the relatively large learning capacity of G, usually implemented as a neural network [39].
b) It is possible to reparametrize the target manifold $\mathcal{N}$ by altering the choice of its Riemannian metric τ so that the preset metric function d is invariant, which is asserted by the following proposition.

   Proposition 3.1. *Consider a regular Riemannian manifold $\mathcal{N}$ with its metric $\tau \in C^\infty$ and its induced distance function $d_{\mathcal{N}}$, then for any path-independent function $f : \mathcal{N}\times\mathcal{N} \to \mathbb{R}^+ \cup \{0\}$, there exists a Riemannian metric τ' on $\mathcal{N}$, induced by the distance function*

$$d'_{\mathcal{N}}(x,y) = f(x,y)d_{\mathcal{N}}(x,y) \quad \forall x,y \in \mathcal{N} \tag{15}$$

Proof. Please see Appendix A.2. □

It is worth to notice, reparametrization of an image manifold in practice can be done with different configurations on the standard manifold learning algorithms [40], [41], [42].

After we replace the original Riemannian metric, the boundary condition $\int_{\mathcal{M}}\int_{\mathcal{N}} d\gamma = 1$ requires renormalization. By introducing an additional matrix $A \in \mathbf{H}(K)$ s.t. $\mathbf{H}(K) \doteq \{A \in \mathbb{R}^{K\times K} | \forall j \in K, \sum_i A_{ij} = K;\ \forall i,j \in [K], A_{ij} \geq 0\}$, the adversarial loss $\min_G \mathcal{L}_{\text{adv}}(G)$ can be reformulated as

$$\min_G \min_{A\in\mathbf{H}(K)} \sum_{i=1}^{K}\sum_{j=1}^{K} \int_{U_i}\int_{V_j} A_{ij} d_{\mathcal{N}}(G(s),t) d\tilde{\mu}_i(s) d\tilde{\nu}_j(t) \tag{16}$$

Specifically, we start from the following form

$$\min_G \inf_{\gamma\in\Pi(\mu,\nu)} \sum_{i=1}^{K}\sum_{j=1}^{K} \int_{U_i}\int_{V_j} d(G(s),t) d\gamma(s,t) \tag{17}$$

By expanding the form with

$$d\gamma(s,t) = d\gamma(t|s)d\mu(s) = \Delta(f_\gamma(s),t)d\mu(s)d\nu(t)$$

we obtain

$$\min_G \min_{f_\gamma} \sum_{i=1}^{K}\sum_{j=1}^{K} \int_{U_i}\int_{V_j} \Delta(f_\gamma(s),t) d(G(s),t) d\mu(s) d\nu(t) \tag{18}$$

By utilizing the observed equivalence between $\min_G \min_{f_\gamma}$ and $\min_G$, we notice the boundary condition $\int_{\mathcal{M}}\int_{\mathcal{N}} d\gamma = 1$ may be broken. Thus we introduce additional variable $A \in \mathbf{H}(K)$ to maintain the normalization condition, which can be checked by

$$\int_M d\gamma = \frac{1}{K^2}\sum_{j=1}^{K}\int_{U_i}\left(\int_{V_j}\sum_{i=1}^{K} A_{ij} d\nu_j\right) d\mu_i = 1 \tag{19}$$

Finally, by inserting the A_{ij} term into the original optimization problem above, we obtain the final form in Eq. (16).

3.2 A Closed-Form Solution as Learning Tasks on Paired Charts

The form of Eq. (16) basically comes from a re-choice of Riemannian metric on $\mathcal{N}$ and a reparametrization of $d\gamma(s,t)$ as $\sum_{i=1}^{K}\sum_{j=1}^{K} A_{ij} d\tilde{\mu}_i(s) d\tilde{\nu}_j(t)$, s.t. $A \in \mathbf{H}(K)$. Illustratively, the adversarial learning problem is depicted in the upper part of Fig. 3. Although it is almost infeasible to obtain a closed-form solution for arbitrary mapping G, we find it is indeed possible after imposing non-trivial constraints on the candidate set of G, namely by restricting G into one *PTI-family*, with its definition below.

Definition 3.1 (Pairwise Topological Immersion family (PTI-family))**.** *Given manifolds $\mathcal{M} = \{(U_i,\varphi_i)\}_{i=1}^K$ and $\mathcal{N} = \{(V_j,\psi_j)\}_{j=1}^K$, the set of mappings $F_p = \{G : \mathcal{M} \to \mathcal{N} | G(U_i) \subset V_{p(i)}, \forall i \in [K]\}$, where $p \in Sym(K)$ the symmetric group of $[K]$ [43], is called pairwise topological immersion mappings indexed by p, w.r.t. $\mathcal{M}$, $\mathcal{N}$.*

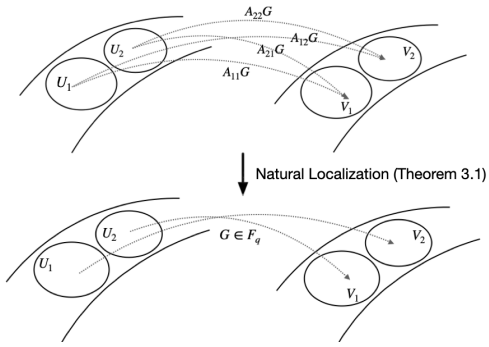

Fig. 3. An illustration of Theorem 3.1. By restricting the generator to certain PTI-family F_p, we prove the original adversarial loss is naturally decomposed as independent style transfer tasks between paired charts.

Preferring to delay intuitive remarks on this definition to Section 3.3.1, we present one of our main results below, which shows that, we can indeed obtain a meaningful closed-form solution for the inner minimization problem, by constraining the candidate set of G as any PTI-family (Def. 3.1).

Theorem 3.1. *[Natural Localization of Adversarial Loss] For any $p \in Sym(K)$, the optimization problem below*

$$\min_{G\in F_p} \min_{A\in\mathbf{H}(K)} \sum_{i=1}^{K}\sum_{j=1}^{K} \int_{U_i}\int_{V_j} A_{ij} d(G(s),t) d\tilde{\mu}_i(s) d\tilde{\nu}_j(t) \tag{20}$$

is equivalent to

$$\min_{G\in F_p} \sum_{i=1}^{K} \int_{U_i}\int_{V_{p(i)}} d(G(s),t) d\tilde{\mu}_i(s) d\tilde{\nu}_{p(i)}(t) \tag{21}$$

In other words, the optimal $A^* \in \mathbf{H}(K)$ has the closed form as $(A^*)^{ij} = K\delta_j^{p(i)}$, where $\delta_j^{p(i)}$ is the Kronecker delta function. This result is also illustrated in Fig. 3.

Proof. Fix $i,j \in [K]$, s.t. $j \neq p(i)$ and arbitrary $G \in F_p$. We first compare the following two terms

$$T_{\text{non-paired}} = \int_{U_i}\int_{V_j} d(G(s),t) d\tilde{\mu}_i(s) d\tilde{\nu}_j(t) \tag{22}$$

and

$$T_{\text{paired}} = \int_{U_i}\int_{V_{p(i)}} d(G(s),t) d\tilde{\mu}_i(s) d\tilde{\nu}_{p(i)}(t) \tag{23}$$

Notice any $s \in U_i$, $G(s) \in V_{p(i)} \cap V_j = \varnothing$, which comes from the assumption that $G \in F_p$ and $j \neq p(i)$, we have $\forall t \in V_{p(i)}, t' \in V_j$, $d(G(s),t) \leq d(G(s),t')$, according to the compatibility of distance function with inner-relatedness of charts.

And thus $T_{\text{non-paired}} \geq T_{\text{paired}}$. Then we relax the fixation of j. It is easy to see,

$$\sum_{j=1}^{K} A_{ij} \int_{U_i}\int_{V_j} d(G(s),t) d\tilde{\mu}_i(s) d\tilde{\nu}_{p(i)}(t) \geq K\int_{U_i}\int_{V_{p(i)}} d(G(s),t) d\tilde{\mu}_i(s) d\tilde{\nu}_{p(i)}(t) \tag{24}$$

which is equivalent to the statement that the optimal $(A_{ij})^* = K\delta^j_{p(i)}$ for each $j \in [K]$.

On the contrary direction, we have for each $A \in \mathbf{H}(K)$,

$$\sum_{i=1}^{K}\sum_{j=1}^{K}\int_{U_i}\int_{V_j} A_{ij}d(G(s),t)d\tilde{\mu}_i(s)d\tilde{\nu}_j(t) \geq \sum_{i=1}^{K}\int_{U_i}\int_{V_{p(i)}} d(G(s),t)d\tilde{\mu}_i(s)d\tilde{\nu}_{p(i)}(t) \quad (25)$$

which concludes the equivalence between optimization problems above. □

3.3 Remarks on Theoretical Results

3.3.1 Discussions with an Illustrative Example

Intuitively, we may consider each chart on $\mathcal{M}, \mathcal{N}$ as a cluster of images, which has inner-relatedness imposed by $\{\mu_i\}_{i=1}^K$, $\{\nu_i\}_{i=1}^K$. Take the example of facial expression transfer [13]. In Fig. 3, U_2 is said to contain a set of Bob's poker face, while V_1, V_2 are respectively sets of Alice's and Bob's smiley face. A PTI-family F_p exactly characterizes the generating tendency of a given generator G. Let us come back to the example. Assume $p, q \in Sym(K)$ with $p(2) = 1$ while $q(2) = 2$. Thus with the input as an image of Bob's poker face, generators from F_p tend to generate a sample of Alice's smiley face, while those from F_q prefer a sample of Bob's smiley face. Although it is clear to us the latter behavior is expected, the adversarial learning model itself however hardly has such a knowledge.

It comes to the significance of Thm. 3.1, which not only gives a closed form for further analysis, but, more essentially, also points out the role of $\{F_p\}_{p\in \text{Sym(K)}}$ as *attractors* during optimization. As we can see, only if the optimizer chooses some generator $G \in F_p$ at some epoch, the original optimization problem (Eq. (20)) will immediately degenerate to learning tasks on paired charts $\{(U_i, V_{p(i)})\}_{i=1}^K$ (Eq. (21)). The generator will thus be trapped in the subset F_p until the end of the training. In our example, that is to say, if the adversarial learning process accidentally lets the generator fall into a wrong PTI family F_p (which generates Alice's faces with the input as Bob's), the generator with no external supervision will never attain the correct correspondence no matter the duration of the training, according to Thm. 3.1. This theorem can be considered as a support to a recent result called *imaginary adversary*, which points out that in the WGAN setting, the minimax game between generator and discriminator can be resolved under some technical conditions [30].

3.3.2 Interpretations of Empirical Phenomena

Interpret Source of Artifacts. Although it brings sharper results with the adversarial loss alone, a non-negligible proportion of artifacts is observed in experiments [1], [13]. As a reasonable interpretation, we suggest it is tightly related with what we have discussed above. Since the adversarial learning model itself has no knowledge of the expected pairing relation, or formally the ground-truth $p \in \text{Sym}(K)$. Although the choice of G (thus F_p) can be guided by the empirical loss during the training phase, it still has a large probability of mistaking the pairing relation, probably due to, e.g., randomization in parameter initialization or sampling. Especially when the optimal pairing it observes is different from the expected one, a PTI-family as an attractor will let the choice *irrevocable*. A compensative approach is by imposing a pointwise correspondence oracle as a regularization term, such as the ℓ_1-loss in [1] or the cycle consistency loss in [12]. From our perspective, these oracles mainly play the role as a *rectifier* for the choice of p.

Interpret Blurry versus Sharp. In previous empirical studies, after learning with identity loss (Eq. (7)) alone, the final generator usually produces more blurry images compared with the generator after learning with the adversarial loss (Eq. (6)). When both of the losses are optimized w.r.t. the same hypothesis space, the identity loss needs to learn a global mapping $G^* : \mathcal{M} \to \mathcal{N}$, while, as a direct result of Thm. 3.1, learning with the adversarial loss theoretically only requires learning the independent local mappings $\{f_i : U_i \to V_{p(i)}\}_{i=1}^K$ first and then gluing them into a global mapping with a well-known theorem from general topology called *partition of unity* [36]. Intuitively, learning local mappings independently requires much smaller capacity of G, compared with learning a globally compatible one (for a theoretical justification, please see Proposition 5.1).

As a complement and a step further, we provide a formal analysis on the benefit of localization in Section 5.1 to complete our interpretations. Due to the indispensable notion of generalization in analyzing model's learning capability [44], we first present an extended definition of generalization for conditional GAN.

4 Generalization for Conditional GAN

4.1 Extension from Previous Definition

As generalization plays a central role in analyzing learning models from a theoretical aspect, Arora et al. proposed the following notion of generalization for GAN [11], by incorporating its difference from conventional discriminative models [44]. Below, we introduce their definition with our notations.

Definition 4.1 (Generalization w.r.t. Divergence [11])**.** *A divergence $D(\cdot,\cdot)$ is said to generalize with m training samples and error ϵ if for the learned generative distribution ν_G (i.e., the true distribution of the generated samples from the generator G), the following inequality holds with a high probability,*

$$|D(\hat{\nu}_{real}, \hat{\nu}_G) - D(\nu_{real}, \nu_G)| < \epsilon \quad (26)$$

where $\hat{\nu}_{real}, \hat{\nu}_G$ are respectively the empirical versions of the real and the generative distributions, and ν_{real} is the true distribution of the real samples.

Although their work marks the first attempt to study the generalization capability of GAN, such a definition has several limitations: **a)** generalization is defined w.r.t a specific divergence, instead of the generator itself. From our perspective, we suggest it is still the generator that holds the fundamental position in generative tasks. **b)** hard for extension to conditional GAN, which however plays an increasingly important role in empirical researches and applications. Such a limitation directly makes it improper to be applied to analyze the adversarial learning approach to style transfer.

As a complement to their notion of generalization, we propose the following extension for both GAN and its conditional variants, which characterizes the phenomenon of generalization with respect to a learned generator G.

Definition 4.2 (Generalization w.r.t. Generator)**.** *Given a divergence $D(\cdot,\cdot)$ and a generator $G : \mathcal{M} \to \mathcal{N}$, we call G generalizes with (m, n) training samples respectively from the source and the target distributions and error ϵ if the following inequality holds with a high probability,*

$$D(G(\hat{\mu}^m_{\mathcal{M}}), \nu_{\mathcal{N}}) - D(\hat{\nu}^n_{\mathcal{N}}, \nu_{\mathcal{N}}) < \epsilon \tag{27}$$

where $\hat{\mu}^m_{\mathcal{M}}, \hat{\nu}^n_{\mathcal{N}}$ are estimators of the source and target distributions, with $\mu_{\mathcal{M}}, \nu_{\mathcal{N}}$ the corresponding true distributions and $G(\hat{\mu}^m_{\mathcal{M}}) \doteq \hat{\mu}^m_{\mathcal{M}} \circ G^{-1}$ the induced distribution on $\mathcal{N}$ [33].

Compared with Def. 4.1, our extension explicitly contains the generator as an essential factor for generalization. Furthermore, instead of assuming the source distribution as a priorly known Gaussian, we view it with an empirical estimator from m observed samples. Notice our definition is actually an extension of Def. 4.1, since, by limiting m to infinity and assuming G of sufficient learning capability in the classical sense, Inequality (27) will directly degenerate to Inequality (26) in the previous definition.

4.2 Relations of Generalization in Different Senses

As an auxiliary theorem for further analysis in the next section, we derive the relation of the classical generalization bound and the generalization bound for adversarial learning based on Def. 4.2. For the sake of concreteness, we specify the divergence $D(\cdot,\cdot)$ in Def. 4.2 as the Lukaszyk-Karmowski metric [45], due to its similarity to the form we have derived in Eq. (21) (by taking d as $\|\cdot\|$, the Euclidean distance on $\mathbb{R}^{w\times h}$),

$$D_{\text{LK}}(\nu, \nu') = \int_{\mathbb{R}^{w\times h}} \int_{\mathbb{R}^{w\times h}} \|x - x'\| d\nu(x) d\nu'(x') \tag{28}$$

where ν, ν' are arbitrary probability measures with their supports in $\mathbb{R}^{w\times h}$ (cf. Eq. (21)).

Theorem 4.1. *Consider a generator $G : \mathbb{R}^{w\times h} \to \mathbb{R}^{w\times h}$ satisfying Lipschitz condition with constant M_G and μ_X, ν_Y are probability measures on $\mathbb{R}^{w\times h}$ respectively with $\{x_i\}_{i=1}^{n_X} \overset{i.i.d.}{\sim} \mu_X$ and $\{y_i\}_{i=1}^{n_Y} \overset{i.i.d.}{\sim} \nu_Y$.*

Assume the classical generalization bound satisfies the following inequality with a probability $1-\delta$

$$\mathbb{E}_{x\sim\mu_X, y\sim\nu_Y} \|G(x) - y\| - \sum_{i=1}^{n_X}\sum_{j=1}^{n_Y} \frac{\|G(x_i) - y_j\|}{n_X n_Y} < \epsilon_{classical} \tag{29}$$

where $\epsilon_{classical} \doteq \epsilon(n_X, n_Y, \mu_X, \nu_Y, \delta)$ is the upper bound and the empirical risk minimization (ERM) principle [44] is satisfied with η (i.e., $\frac{1}{n_X n_Y}\sum_{i=1}^{n_X}\sum_{j=1}^{n_Y}\|G(x_i) - y_j\| < \eta$), then G generalizes with (n_X, n_Y) training samples and with an ϵ_{adv} error, with a probability $1-\delta$, i.e.,

$$D_{LK}(G(\hat{\mu}^{n_X}_X), \nu_Y) - D_{LK}(\hat{\nu}^{n_Y}_Y, \nu_Y) < \epsilon_{adv} \tag{30}$$

if the following condition is satisfied

$$\epsilon_{classical} - \epsilon_{adv} + \eta < D_{LK}(\nu_Y, \hat{\nu}^{n_Y}_Y) - M_G D_{LK}(\mu_X, \hat{\mu}^{n_X}_X) \tag{31}$$

Proof. Please see Appendix A.3. □

As Theorem 4.1 indicates, unlike the classical generalization bound (especially in the Vapnik-Chervonenkis (VC) sense [44]), the generalization error in adversarial learning is also related with the variation of the source and the target distributions.

5 Conditions of Generalization for Pix2Pix

Based on our extended notion of generalization above, we are now able to fulfill our interpretations for Blurry versus Sharp (§5.1). As a step further, we also derive a quantitative condition to control the generalization capability of the pix2pix model (Thm. 5.1), which provides several practical implications on model design and dataset construction for practitioners.

For the sake of concreteness, we start by specifying some additional statistical settings. Recall in Assumption 2.2, we have imposed abstract probability measures $\{\mu_i\}_{i=1}^K$, $\{\nu_i\}_{i=1}^K$ on $\{\varphi_i(U_i)\}_{i=1}^K$ and $\{\psi_i(V_i)\}_{i=1}^K$ respectively. We further specify such an assumption with *local Gaussian settings*.

Assumption 5.1. *There exist unknown mean vectors in $\mathbb{R}^d$, denoted as $\{x_i\}_{i=1}^K$, $\{y_i\}_{i=1}^K$, and known covariance matrices $\Sigma_{\mathcal{M}}, \Sigma_{\mathcal{N}} \in \mathbb{R}^{d\times d}$, such that for each $i \in [K]$, $\mu_i = \mathcal{N}(\cdot; x_i, \Sigma_{\mathcal{M}})$, $\nu_i = \mathcal{N}(\cdot; y_i, \Sigma_{\mathcal{N}})$, where $\mathcal{N}(\cdot; x, \Sigma)$ denotes the normal distribution parametrized by (x, Σ). Additionally, we set the sample sizes on charts $\{U_i\}_{i=1}^K$, $\{V_i\}_{i=1}^K$ equally as m, n, without loss of generality.*

It ought to be noticed that our Gaussian assumption above will not impose strong limitations on our following discussions, mainly because the Gaussian assumption remains local (cf. the vanilla GAN [2]) and the parameters of each Gaussian is not required to be observed (cf. Def. 4.1).

5.1 Benefits of Localization

In our previous interpretation of Blurry versus Sharp (§3.3), a claim remains unjustified that learning a set of local mappings is much easier compared with learning a globally compatible mapping. With the following observations: **a)** Lipschitz condition can be always satisfied with techniques (e.g., clipping [20] or gradient penalty [46]) during training phase. **b)** $\epsilon_{\text{classical}}, \eta, M_G$ remain constant for the same hypothesis space. **c)** The target-related term $D_{\text{LK}}(\nu, \hat{\nu})$ is identical in both the local and the global tasks when the pairing relation is unobserved, we reformulate Inequality (31) as

$$C + M_G D_{LK}(\mu, \hat{\mu}) < \epsilon_{\text{adv}} \tag{32}$$

where $C := \epsilon_{\text{classical}} + \eta - D_{\text{LK}}(\nu, \hat{\nu})$ a constant and the Lipschitz constant M_G is always non-negative.

By denoting probability measures underlying the global task as $\mu_X = \frac{1}{K}\sum_{i=1}^K \mu_i$ and $\nu_Y = \frac{1}{K}\sum_{i=1}^K \nu_i$ in the Euclidean sense, it is sufficient to compare the two terms below to justify our previous claim.

$$\epsilon^{local}_{adv} = \frac{1}{K}\sum_{i=1}^K D_{LK}(\mu_i, \hat{\mu}^m_i) \tag{33}$$

$$\epsilon_{adv}^{global} = D_{LK}(\frac{1}{K}\sum_{i=1}^{K}\mu_i, \hat{\mu}_X^{Km}) \tag{34}$$

Intuitively, the term ϵ_{adv}^{local} represents the average generalization errors for all the local tasks (i.e., $\mu_i \to \nu_i, \forall i \in [K]$), while ϵ_{adv}^{global} can be interpreted as the generalization error when the learning process is carried out globally (i.e., $\mu_X \to \nu_Y$). For convenience, we set the pairing relation $e \in \mathrm{Sym}(K)$ as $e(i) = i, \forall i \in [K]$ (the corresponding PTI-family denoted as F_e). With the following proposition, we eventually complete our unfinished interpretations for the empirical results.

Proposition 5.1. *In the settings above, we always have*

$$\epsilon_{adv}^{local} < \epsilon_{adv}^{global}$$

Proof. Please see Appendix A.4. □

5.2 Conditions of Generalization

We are now able to instantiate the generic inequality on generalization. It is worth to notice, in the next theorem, we characterize the classical generalization error term in the VC sense and study the condition for $\epsilon_{\text{adv}} = 0$, which, intuitively speaking, guarantees the generated distribution is even better than an estimated target distribution from m real samples.

Theorem 5.1. *Under the assumptions above, consider a generator $G \in F_e$ and a hypothesis space $\mathcal{H}$ with its VC-dimension bound by a constant Λ. Assume for each $i \in [K]$, the restriction of G to a pair of charts $f_i \doteq G_{\downarrow(U_i,V_i)} \in \mathcal{H}$ with $\psi_i \circ G \circ \varphi_i^{-1}$ satisfies Lipschitz condition with a constant M_G, then G generalizes globally with (Kn, Km) samples only if the following inequality is satisfied with a probability $1 - C(\epsilon,\Lambda)(nm\epsilon^2)^{\tau(\Lambda)}e^{-nm\alpha\epsilon^2}$,*

$$\epsilon + \frac{1}{nm}\max\{\sum_{i=1}^{n}\sum_{j=1}^{m} d(G(s_k^i), t_k^j)\}_{k=1}^{K} < \frac{1}{\sqrt{m}}\sqrt{tr(\Sigma_{\mathcal{N}})} + 2tr(\Sigma_{\mathcal{N}}) - M_G(\frac{1}{\sqrt{n}}\sqrt{tr(\Sigma_{\mathcal{M}})} + 2tr(\Sigma_{\mathcal{M}})) \tag{35}$$

where $C(\epsilon,\Lambda)$ and $\tau(\Lambda)$ are positive functions independent from n, m and $\alpha \in [1, 2]$ an absolute constant.

Proof. In order to apply Thm. 4.1, we first introduce the following lemma to show the equivalence between LK metric (Eq. (28)) with the assumed probability measures in Euclidean space and each local objective defined on the manifolds, which can be easily proved with the standard results on norm equivalence (e.g., [36]).

Lemma 5.1. *$\forall i \in [K]$, consider a measurable mapping $\tilde{f} : U_i \to V_i$ with $f \doteq \psi_i \circ \tilde{f} \circ \varphi_i^{-1}$ satisfying the Lipschitz condition, then $\int_{U_i}\int_{V_i} d(\tilde{f}(s), t)d\tilde{\mu}_i(s)d\tilde{\nu}_i(t) \simeq D_{LK}(f(\mu_i), \nu_i)$, i.e., there exist constants $0 < C_l < C_u < \infty$ such that*

$$C_l < \frac{\int_{U_i}\int_{V_i} d(\tilde{f}(s), t)d\tilde{\mu}_i(s)d\tilde{\nu}_i(t)}{D_{LK}(f(\mu_i), \nu_i)} < C_u \tag{36}$$

Therefore, for K independent local tasks, the global generalization condition in Thm. 4.1 can be written as

$$\max\{\epsilon_{\text{classical}}^i - \epsilon_{\text{adv}}^i + \eta^i\}_{i=1}^{K} < \min\{D_{\text{LK}}(\nu_i, \hat{\nu}_i^n) - M_G D_{\text{LK}}(\mu_i, \hat{\mu}_i^m)\}_{i=1}^{K} \tag{37}$$

which serves as a sufficient condition in the worst case.

We take the surrogate of the LHS by setting $\epsilon_{\text{adv}}^i = 0$, which corresponds to the situation when the observed samples on each pair of charts are identical. Therefore, we can reformulate the inequality as

$$\epsilon_{\text{classical}} + \max\{\eta^i\}_{i=1}^{K} < \min\{D_{\text{LK}}(\nu_i, \hat{\nu}_i^n) - M_G D_{\text{LK}}(\mu_i, \hat{\mu}_i^m)\}_{i=1}^{K} \tag{38}$$

Applying the result from [47], we could bound the left side by ϵ with a probability $1 - C(\epsilon,\Lambda)(nm\epsilon^2)^{\tau(\Lambda)}e^{-nm\alpha\epsilon^2}$, that is

$$\epsilon_{\text{classical}} + \max\{\eta^i\}_{i=1}^{K} < \epsilon + \frac{1}{nm}\max\{\sum_{i=1}^{n}\sum_{j=1}^{m} d(G(s_k^i), t_k^j)\}_{k=1}^{K} \tag{39}$$

The next step is to deal with the right side. With some algebras, we could deduce

$$\begin{aligned}&\min\{D_{\text{LK}}(\nu_i, \hat{\nu}_i^n) - M_G D_{\text{LK}}(\mu_i, \hat{\mu}_i^m)\}_{i=1}^{K} \\ &= \min\{\mathbb{E}\|\nu_i - \nu_i^n\| - M_G\mathbb{E}\|\mu_i - \mu_i^n\|\}_{i=1}^{K} \\ &+2\text{tr}(\Sigma_{\mathcal{N}}) - 2\text{tr}(\Sigma_{\mathcal{M}})\end{aligned} \tag{40}$$

In order to write the first minimization term in a closed form, we use the following theorem in information geometry.

Theorem 5.2. *[48, Theorem 4.4] The mean square error of a bias-corrected first-order efficient estimator is given asymptotically by the expansion (with N observed samples):*

$$\mathbb{E}[(\hat{u}^a - u^a)(\hat{u}^b - u^b)] = \frac{1}{N}g^{ab} + O(\frac{1}{N^2}) \tag{41}$$

where g^{ab} denotes the Fisher metric on the manifold constructed from a parametrized family of probability.

We thus apply this estimation on $\mathbb{E}\|\nu_i - \nu_i^n\|$ and $\mathbb{E}\|\mu_i - \mu_i^n\|$. As is well known, the matrix of the Fisher metric for a Gaussian $\mathcal{N}(x, \Sigma)$ is directly Σ, the covariance matrix itself.

By observing $\mathbb{E}\|\nu_i - \hat{\nu}_i^n\| = \sqrt{tr(\mathbb{E}[(\nu_i - \hat{\nu}_i^n)(\nu_i - \hat{\nu}_i^n)^T])}$ and $\mathbb{E}\|\mu_i - \hat{\mu}_i^m\| = \sqrt{tr(\mathbb{E}[(\mu_i - \hat{\mu}_i^n)(\mu_i - \hat{\mu}_i^n)^T])}$, we have

$$\begin{aligned}&\min\{D_{\text{LK}}(\nu_i, \hat{\nu}_i^n) - M_G D_{\text{LK}}(\mu_i, \hat{\mu}_i^m)\}_{i=1}^{K} \\ &= \frac{1}{\sqrt{m}}\sqrt{\text{tr}(\Sigma_{\mathcal{N}})} + 2\text{tr}(\Sigma_{\mathcal{N}}) - M_G(\frac{1}{\sqrt{n}}\sqrt{\text{tr}(\Sigma_{\mathcal{M}})} + 2\text{tr}(\Sigma_{\mathcal{M}}))\end{aligned} \tag{42}$$

which thus gives the condition of generalization above (with the $O(N^{-2})$ term omitted). □

Discussions & Guidance for Practitioners. A brief discussion on Theorem 5.1 and its practical guidance will conclude this section. As we can see, generalization happens with a higher probability when the RHS of Inequality (35) gets larger and the LHS gets smaller. On one hand, as the RHS is negatively related with the variance of the source distribution (as we can see from the terms $-\text{tr}(\Sigma_{\mathcal{M}})$ and $-\sqrt{\text{tr}(\Sigma_{\mathcal{M}})}$ in RHS), a larger RHS term can be brought by a smaller variance of each local source distribution, especially considering the multiplier effect of the Lipschitz constant

M_G on the $\text{tr}(\Sigma_{\mathcal{M}})$-related terms. On the other hand, to ensure a lower LHS term requires a *uniformly* lower empirical risk. As each local chart has an intuitive interpretation as a set of related images, we make the following suggestions on dataset construction and model design.

- As a smaller variance of each local source distribution means better generalization, images in one local chart of the source manifold should be as similar to each other as possible. For example, n copies of the same Bob's photo in one local chart gives a smaller variance than n different photos of Bob. When this local condition extends to the full training data, we suggest that the source set of images should better respect a mixture of single point distributions by, e.g., containing a collection of photos of n different persons, other than a collection of images where each one of them has some relations with others. In one word, the source set of images should better be of lower inner-similarity.
- As better generalization requires the empirical risks on all the tasks to be uniformly lower, a blind increase in the total number of images will hardly help generalization. It is the balance in the numbers of different objects that actually matters (empirically proved by [13]).
- Classical generalization capacity [44] and smoothness of learning model w.r.t. data manifolds [49] should be considered equivalently important in model design for such tasks (empirically proved by [7], [50], [51]).

6 Existence and Abundance of Perfect Generators for Style Transfer

To generalize our earlier results, we subsequently study the adversarial learning approach to style transfer in general. As a novel observation, we prove in this section the equivalence of the C^∞-diffeomorphism condition with the existence and abundance of perfect generators for style transfer. Noticeably, this remarkable condition is model-free and imposes no additional restrictions on corresponding distributions only if they are absolutely continuous over the manifold structure.

6.1 Existence of Global Optimum

Theorem 6.1 (Existence Theorem)**.** *There always exists a global optimum (G^*, γ^*) for Objective (6) such that*

$$\int_{\mathcal{M}\times\mathcal{N}} d(G^*(p), q) d\gamma^*(p, q) = 0 \tag{43}$$

if and only if $\mathcal{M} \equiv \mathcal{N}$.

Proof. **(Necessity)** As Objective (6) is minimized to 0 for some (G^*, γ^*), it means that there exists a sequence of Riemannian manifolds and couplings $\{(\mathcal{N}_t, \gamma_t)\}_{t=1}^{\infty}$, where $\mathcal{N}_t := G_t(\mathcal{M}), G_t \in C^\infty(\mathcal{M}, \mathbb{R}^n)$ and $\gamma_t \in \Pi(\mu, \nu)$, such that $d(G_t(p), q) < \epsilon_t$ a.e. γ_t, for negligible errors $\epsilon_t > 0$, $\lim_{t\to\infty} \epsilon_t = 0$.

Lemma 6.1. *Let G_t be a continuous mapping and the marginal distribution of $\gamma_t \in \Pi(\mu, \nu)$ such that $d(G_t(p), q) < \epsilon_t$ a.e. γ_t, for negligible errors $\epsilon_t > 0$, $\lim_{t\to\infty} \epsilon_t = 0$. We claim $\sup_{q' \in \mathcal{N}_t} \inf_{q\in\mathcal{N}} d(q', q) < O(\epsilon_t)$ and $\sup_{q\in\mathcal{N}} \inf_{q'\in\mathcal{N}_t} d(q, q') < O(\epsilon_t)$. In other words, $d_H(G_t(\mathcal{M}), \mathcal{N}) \le O(\epsilon_t)$.*

Proof. First, we prove the absolute continuity of γ_t. According to Assumption 2.2, μ, ν are absolutely continuous w.r.t. the respective volume form. Since γ_t is a coupling, which satisfies $\int_{\mathcal{M}} d\gamma_t(p, \cdot) = d\nu(\cdot)$, it implies that $d\gamma_t(p, \cdot)$ is absolutely continuous w.r.t. ν (a.t. Radon-Nikodym Theorem [36]). A similar proof goes with μ. Therefore, with the transitivity of absolute continuity, we have γ_t is absolutely continuous on $\mathcal{M} \times \mathcal{N}$.

Next, fix an arbitrary $p \in \mathcal{M}$ and its small open neighborhood U_p. With the absolute continuity of a coupling γ_t, there exist an open set $V_p \subset \mathcal{N}$ s.t.

$$\int_{U_p\times V_p} d(G_t(p'), q) d\gamma_t(p', q) \le O(\epsilon_t) \tag{44}$$

Therefore,

$$\int_{\mathcal{N}_t} \inf_{q\in V_p} d(q', q) dG_{t\#}\mu(p') \le O(\epsilon_t) \tag{45}$$

Since $G_{t\#}\mu$ is absolutely continuous (due to the continuity of G_t), so as the Euclidean metric d, the average bound $O(\epsilon_t)$ becomes a worst-case bound. Thus, with a slight abuse of notations, we have

$$\sup_{q'\in\mathcal{N}_t} \inf_{q\in\mathcal{N}} d(q', q) < O(\epsilon_t) \tag{46}$$

Similarly, we can prove

$$\sup_{q\in\mathcal{N}} \inf_{q'\in\mathcal{N}_t} d(q, q') < O(\epsilon_t) \tag{47}$$

which in turn, by definition of Hausdorff distance, means $d_H(G_t(\mathcal{M}), \mathcal{N}) \le O(\epsilon_t)$. □

By taking limit of t at both sides of $d_H(G_t(\mathcal{M}), \mathcal{N}) \le O(\epsilon_t)$, we have $\lim_{t\to\infty} d_H(G_t(\mathcal{M}), \mathcal{N}) = 0$. By applying Gromov's Theorem (Thm. 2.1), we therefore have for a sufficiently large T, $G_T(\mathcal{M}) \equiv \mathcal{N}$. Again, due to the continuity and openness of G_T, $G_T(\mathcal{M}) \equiv \mathcal{M}$ and transitivity of diffeomorphism, we finally reach the conclusion that $\mathcal{M} \equiv \mathcal{N}$.

(Sufficiency) By definition, the assumption $\mathcal{M} \equiv \mathcal{N}$ can be equivalently stated as, there exists a mapping $G \in C^\infty(\mathcal{M}; \mathcal{N})$ such that G is a diffeomorphism between $\mathcal{M}$ and $\mathcal{N}$. Therefore, we forward the proof by explicitly constructing a deterministic coupling between the pair of distributions $(G_{\#}\mu, \nu)$ on Riemannian manifold $\mathcal{N}$. Here, we use the standard notation $G_{\#}\mu$ to denote the induced probability measure of μ by mapping $G : \mathcal{M} \to \mathbb{R}^n$. Formally, $G_{\#}\mu(\cdot) := \mu(G^{-1}(\cdot))$.

In other words, we would like to prove the existence of $\gamma^* \in \Pi(G_{\#}\mu, \nu)$ such that

$$\int_{\mathcal{N}\times\mathcal{N}} d(q', q) d\gamma^*(q', q) = 0 \tag{48}$$

Lemma 6.2. *Due to the regularity conditions of $\mathcal{N}$, there exist constants C_l, C_u such that $C_l d_{\mathcal{N}}(q', q) \le d(q', q) \le C_u d_{\mathcal{N}}(q', q)$, where $d_{\mathcal{N}}$ is the geodesic distance metric on $\mathcal{N}$ [25].*

Proof. As the regularity conditions suggest, $\mathcal{N}$ is compact. Therefore, since $\mathcal{N}$ is of finite dimension, then with the

Heine-Borel property [36], $\sup_{q^{'},q} d_{\mathcal{N}}(q^{'},q) < \infty$, so as $d(q^{'},q)$. Furthermore, $d_{\mathcal{N}}(q^{'},q) = 0 \iff q^{'} = q$ a.e., so as $d(q^{'},q) = 0 \iff q^{'} = q$ a.e. Therefore, there exist positive constants C_l, C_u s.t. $C_l d_{\mathcal{N}}(q^{'},q) \leq d(q^{'},q) \leq C_u d_{\mathcal{N}}(q^{'},q)$. □

Therefore, we would turn to consider the existence of coupling γ^* such that

$$\int_{\mathcal{N}\times\mathcal{N}} d_{\mathcal{N}}(q^{'},q) d\gamma^*(q^{'},q) = 0 \tag{49}$$

which requires the aid of Moser's theorem from optimal transport theory.

Theorem 6.2 (Moser's Theorem [28])**.** *Let $\mathcal{N}$ be a n-dimensional Riemannian manifold, equipped with a normalized measure m induced from its volume form. Let $\mu_0 = \rho_0 m$, $\mu_1 = \rho_1 m$ be two probability measures on $\mathcal{N}$ where ρ_0, ρ_1 are bound below by a constant $K > 0$ and are locally Lipschitz. If the equation*

$$\Delta u = \rho_0 - \rho_1 \tag{50}$$

has solution $u \in C^{1,1}_{loc}(\mathcal{N})$ (i.e., ∇u is locally Lipschitz). Then there exists a flow $(T_t)_{0\leq t\leq 1}$ s.t. $\mu_t = (T_t)_{\#}\mu_0$.

In order to leverage Moser's theorem, the following lemmas need to be checked.

Lemma 6.3. *Probabilistic measures $f_{\#}\mu$ and ν are absolute continuous w.r.t. the normalized measure $m(dp) = vol_{\mathcal{N}}(dp)/\int_{\mathcal{N}} vol_{\mathcal{N}}(dp)$ on $\mathcal{N}$. As a result, we could write $f_{\#}\mu = \rho_0 m$ and $\nu = \rho_1 m$, where ρ_0, ρ_1 are locally Lipschitz.*

Proof. According to Assumption 2.2, the absolute continuity of ν is immediate. For $G_{\#}\mu$, consider an arbitrary $dS \subset \mathcal{N}$ s.t. $\text{vol}_N(dS) = 0$ then by definition, $G_{\#}\mu(dS) = \mu(G^{-1}(dS)) = 0$ where the last equality comes from the continuity of f, which follows the fact that G is a diffeomorphism.

Next, we claim ρ_0, ρ_1 are locally Lipschitz. Take $\nu = \rho_0 m$ as an example. First, the continuity of ρ_0 is by definition. Therefore, according to Miculescu et al. [52], there exists a sequence of locally Lipschitz functions to approximate ρ_0, therefore we simply replace the original ρ_0 with the limit with an infiniesimal error. □

Lemma 6.4. *The differential equation $\Delta u = \rho_0 - \rho_1$ has solutions in $C^{1,1}_{loc}(\mathcal{N})$.*

Proof. Please see Appendix A.5. □

With Lma. 6.3 & 6.4, Thm. 6.2 can be applied to construct a flow $\{T_t\}_{0\leq t\leq 1}$ on $\mathcal{N}$ s.t. $T_0 = \text{Id}$ while $\nu = T_{1\#}G_{\#}\mu$. (For definitions of flow, please see Def. 6.1.) Therefore, by constructing the coupling of μ, ν as $\gamma^*(p,q) = \delta_q^{T_1(G(p))} d\mu(p) d\nu(q)$ and $G^* = G$, we find (G^*, γ^*) as a global optimum for Objective (6), such that $\int_{\mathcal{M}\times\mathcal{N}} d(G^*(p), q) d\gamma^*(p,q) = 0$.

□

6.2 Abundance of Perfect Generators

From the statement and the proof of Thm. 6.1 above, we have already established the existence of a global optimum (G^*, γ^*) for arbitrary WGAN settings, if and only if $\mathcal{M}$ is C^∞-diffeomorphic to $\mathcal{N}$. In this part, we observe we could indeed construct infinitely many perfect generators from the one we have found. For intuition, please see Fig. 4.

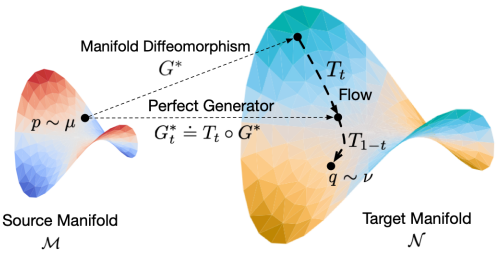

Fig. 4. An illustrative proof for Theorem 6.3 and part of Theorem 6.1.

In the last few lines of the proof of Thm. 6.1, we have once constructed a *flow* $\{T_t\}_{0\leq t\leq 1}$ on $\mathcal{N}$ as by-products. In case of readers' unfamiliarity of the concept of flow, we concisely provide its definition below.

Definition 6.1 (Flow on Riemannian manifold [25])**.** *We call $\{T_t\}_{0\leq t\leq 1}$ a flow on $\mathcal{N}$ if the following conditions are satisfied:*

1) *$\forall t \in [0,1]$, $T_t : \mathcal{N} \to \mathcal{N}$ is a continuous mapping from Riemannian manifold $\mathcal{N}$ to itself.*
2) *$T_0 = Id$, i.e., $\forall q \in \mathcal{N}, T_0(q) = q$.*
3) *$\forall t, s, t+s \in [0,1], T_s \circ T_t = T_{s+t}$.*

With an application of the properties of constructed flows on $\mathcal{N}$, we prove the following theorem, which, intuitively speaking, ensures the abundance of global optima for Objective (6).

Theorem 6.3 (Abundance Theorem)**.** *If $\mathcal{M} \equiv \mathcal{N}$, there exists an infinite number of global optima $\{(G_t^*, \gamma_t^*)\}_{0\leq t\leq 1}$ for Objective (6) such that*

$$\int_{\mathcal{M}\times\mathcal{N}} d(G_t^*(p), q) d\gamma_t^*(p,q) = 0 \tag{51}$$

Proof. Recall in the sufficiency part of the proof for Thm. 6.1, we have constructed a global optimum (G^*, γ^*) s.t.

$$\int_{\mathcal{M}\times\mathcal{N}} d(G^*(p), q) d\gamma^*(p,q) = 0 \tag{52}$$

and a flow $\{T_t\}_{0\leq t\leq 1}$ on $\mathcal{N}$ such that $T_0 = \text{Id}$, $T_{1\#}G_{\#}\mu = \nu$ a.e.

Then we claim, for each $0 \leq t \leq 1$, (G_t^*, γ_t^*) is also a global optimum of value 0, where $G_t^* \doteq T_t \circ G^*$ and $\gamma_t^* \doteq \delta_q^{T_{1-t}G_t^*(p)} d\mu(p) d\nu(q)$. First of all, it is easy to check the well-definedness that $G_t^* \in C^\infty(\mathcal{M}; \mathbb{R}^n)$ and $\gamma_t^* \in \Pi(\mu, \nu)$ mainly due to the continuity of T_t.

By replacing (G^*, γ^*) on the left side of Eq. (52) with the constructed (G_t^*, γ_t^*) for a arbitrary $t \in [0,1]$, we have

$$\begin{aligned}
&\int_{\mathcal{M}\times\mathcal{N}} d(G_t^*(p), q) d\gamma_t^*(p,q) \\
&= \int_{\mathcal{N}\times\mathcal{N}} d(q^{'},q) \delta_q^{q^{'}} dT_{1-t\#}G_{t\#}^*\mu(q^{'}) d\nu(q) \\
&= 0
\end{aligned} \tag{53}$$

where the last line comes from the fact that $T_t \circ T_{1-t} = T_1$ and $T_{1\#}G_{\#}\mu = \nu$ a.e. □

As a corollary, we show the implied mutual exclusivity between global optimum and mode collapse as follows.

Corollary 6.1 (Resolved mode collapse by $\mathcal{M} \equiv \mathcal{N}$)**.** *Any generator $G_t{}^*$ constructed in Thm. 6.3 is perfect. In other words, there will be no mode collapse for any one of them.*

Proof. For a random variable $p \sim \mu$ and an arbitrary $t \in [0, 1]$, we have $T_{1-t}G_t(p) \sim \nu$ from the proof above. Furthermore, as the diffeomorphism G is invertible (Def. 2.1), there always exists a dual flow $\{R_t\}_{0\leq t\leq 1}$ on $\mathcal{M}$ s.t. $T_{1-t}G_t^*(p) = G_t^*(R_t p)$. Therefore, by reparametrizing the source distribution μ as $\mu \circ R_t$, it is easy to see G_t^* is a perfect generator, for which no mode collapse will occur. □

Future Directions. With the proved abundance of perfect generators under the C^∞-diffeomorphism between the source and the target manifolds, it would be promising for future works to apply this result to improve the image generation with GANs. For example, as the GAN generation process can be viewed as a special case of "style transfer" from a noise distribution to a given image distribution, the diffeomorphism between the noise manifold and the real image manifold should also guarantee the existence of infinitely many generators. It may mitigate mode collapse to some degree and hence increase the quality of image generation. In practice, to ensure the diffeomorphism with tolerable errors, it would be an interesting direction for future works to construct the noise distribution via sampling noises from the persistent Cěch complex [53] of the real data distribution [54].

7 Conclusions

From a novel geometric viewpoint, this paper presents the first theoretical framework for understanding the empirical phenomena and the underlying mechanism of image style transfer with adversarial learning. By providing comprehensive interpretations on previously unclear experimental results in style transfer, clarifying the generalization capacity of the state-of-the-art pix2pix model and establishing the equivalence between C^∞-diffeomorphism and the existence of infinitely many perfect generators, we strongly demonstrate the utility and fruitfulness of our framework for analysis. For future directions, it will be interesting to further leverage this geometric tool to study the implication of diffeomorphism on the possible convergence to Nash equilibrium between the generator and discriminator. In practice, future works may explore better architectures for generative learning under the guidance of our theory. As our theoretical results can be easily decoupled with the image settings, we also suggest to study the transfer problem between a source manifold structure to a target one via adversarial learning in other application domains.

Acknowledgments

We would like to thank the anonymous reviewers and the editors for their constructive comments and input to improve our manuscript. This work was supported in part by National Natural Science Foundation of China (61972099, U1636204, U1836213, U1836210, U1736208), and Natural Science Foundation of Shanghai (19ZR1404800). Min Yang is a faculty of Shanghai Institute of Intelligent Electronics & Systems, Shanghai Institute for Advanced Communication and Data Science, and Engineering Research Center of CyberSecurity Auditing and Monitoring, Ministry of Education, China.

References

[1] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 5967–5976, 2017.

[2] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014.

[3] A. Siarohin, E. Sangineto, S. Lathuilière, and N. Sebe, "Deformable gans for pose-based human image generation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3408–3416.

[4] L. Yu, W. Zhang, J. Wang, and Y. Yu, "Seqgan: Sequence generative adversarial nets with policy gradient," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 2852–2858.

[5] Y. Saito, S. Takamichi, and H. Saruwatari, "Statistical parametric speech synthesis incorporating generative adversarial networks," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, pp. 84–96, 2018.

[6] J. Yoon, J. Jordon, and M. van der Schaar, "GAIN: Missing data imputation using generative adversarial nets," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 5689–5698.

[7] J. Cao, Y. Guo, Q. Wu, C. Shen, J. Huang, and M. Tan, "Adversarial learning with local coordinate coding," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 707–715.

[8] J. Harer, O. Ozdemir, T. Lazovich, C. Reale, R. Russell, L. Kim, and p. chin, "Learning to repair software vulnerabilities with generative adversarial networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 7933–7943.

[9] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *Comput. Research Reposit.*, vol. abs/1411.1784, 2014.

[10] R. J. Aumann, *Game theory*. Springer, 1989.

[11] S. Arora, R. Ge, Y. Liang, T. Ma, and Y. Zhang, "Generalization and equilibrium in generative adversarial nets (gans)," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 224–232.

[12] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 2223–2232, 2017.

[13] 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," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 8789–8797, 2017.

[14] M. Sela, E. Richardson, and R. Kimmel, "Unrestricted facial geometry reconstruction using image-to-image translation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1576–1585.

[15] C. Yang, T. Kim, R. Wang, H. Peng, and C.-C. J. Kuo, "Show, attend and translate: Unsupervised image translation with self-regularization and attention," *IEEE T Image Process.*, vol. 28, pp. 4845–4856, 2019.

[16] E. Reinhard, M. Ashikhmin, B. Gooch, and P. Shirley, "Color transfer between images," *IEEE Comput. Graph. App.*, vol. 21, pp. 34–41, 2001.

[17] M. D. Zeiler, G. W. Taylor, L. Sigal, I. A. Matthews, and R. Fergus, "Facial expression transfer with input-output temporal restricted boltzmann machines," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2011, pp. 1629–1637.

[18] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 2414–2423, 2016.

[19] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M.-H. Yang, "Universal style transfer via feature transforms," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 386–396.

[20] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 214–223.

[21] M. Arjovsky and L. Bottou, "Towards principled methods for training generative adversarial networks," *Comput. Research Reposit.*, vol. abs/1701.04862, 2017.

[22] L. Mescheder, A. Geiger, and S. Nowozin, "Which training methods for gans do actually converge?" in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 3478–3487.

[23] D. Balduzzi, S. Racaniere, J. Martens, J. Foerster, K. Tuyls, and T. Graepel, "The mechanics of n-player differentiable games," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 354–363.

[24] H. Lu, Y. Fainman, and R. Hecht-Nielsen, "Image manifolds," *Appl. Artif. Neural Netw. Image Process.*, pp. 52–63, 1998.

[25] J. Jost, *Riemannian Geometry and Geometric Analysis*. Springer Sci. & Busin. Media, 2008.

[26] Z. Lin, A. Khetan, G. Fanti, and S. Oh, "Pacgan: The power of two samples in generative adversarial networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 1498–1507.

[27] X. Pan, M. Zhang, and D. Ding, "Theoretical analysis of image-to-image translation with adversarial learning," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 4006–4015.

[28] C. Villani, *Optimal Transport: Old and New*. Springer Sci. Busin. Media, 2008, vol. 338.

[29] J.-Y. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros, "Generative visual manipulation on the natural image manifold," in *Eur. Conf. Comput. Vis.*, 2016, pp. 597–613.

[30] N. Lei, K. Su, L. Cui, S.-T. Yau, and X. Gu, "A geometric view of optimal transportation and generative model," *Comput. Aided Geom. Design*, pp. 1–21, 2019.

[31] J. Lee, *Introduction to Topological Manifolds*. Springer Sci. & Busin. Media, 2010, vol. 940.

[32] K. Fukaya, "Hausdorff convergence of riemannian manifolds and its applications," in *Recent Topics in Different. and Analyt. Geometry*. Elsevier, 1990, pp. 143–238.

[33] K. L. Chung, *A Course in Probability Theory*. Academic press, 2001.

[34] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *Comput. Research Reposit.*, vol. abs/1706.06083, 2017.

[35] M. Nakahara, *Geometry, Topology and Physics*. CRC Press, 2003.

[36] W. Rudin, *Real and Complex Analysis*. McGraw-hill Education, 2010.

[37] W. Dai, Y. Chen, G.-R. Xue, Q. Yang, and Y. Yu, "Translated learning: Transfer learning across different feature spaces," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2009, pp. 353–360.

[38] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[39] E. D. Sontag, "Vc dimension of neural networks," *NATO ASI Series F Comput. Sys. Sci.*, vol. 168, pp. 69–96, 1998.

[40] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 2319–2323, 2000.

[41] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, pp. 2323–2326, 2000.

[42] G. E. Hinton and S. T. Roweis, "Stochastic neighbor embedding," in *Proc. Int. Conf. Mach. Learn.*, 2003, pp. 857–864.

[43] P. J. Cameron, *Permutation Groups*. Cambridge University Press, 1999, vol. 45.

[44] V. N. Vapnik, *Statistical Learning Theory*. Wiley New York, 1998.

[45] S. Łukaszyk, "A new concept of probability metric and its applications in approximation of scattered data sets," *Comput. Mech.*, vol. 33, pp. 299–304, 2004.

[46] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 5767–5777.

[47] N. Vayatis and R. Azencott, "Distribution-dependent vapnik-chervonenkis bounds," in *Eur. Comput. Learn. Theory*, 1999, pp. 230–240.

[48] S.-i. Amari and H. Nagaoka, *Methods of Information Geometry*. Am. Math. Soc., 2007.

[49] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *J Mach. Learn. Research*, vol. 7, pp. 2399–2434, 2006.

[50] Z. Huang, J. Wu, and L. Van Gool, "Manifold-valued image generation with wasserstein adversarial networks," *Proc. AAAI Conf. Artif. Intell.*, pp. 3886–3893, 2019.

[51] G.-J. Qi, L. Zhang, and H. Hu, "Global versus localized generative adversarial nets," *Comput. Research Reposit.*, vol. abs/1711.06020, 2017.

[52] R. Miculescu *et al.*, "Approximation of continuous functions by lipschitz functions," *Real Analys. Exchange*, vol. 26, pp. 449–452, 2000.

[53] H. Edelsbrunner and J. Harer, "Persistent homology: A survey," *Contemp. Math.*, vol. 453, pp. 257–282, 2008.

[54] V. Khrulkov and I. Oseledets, "Geometry score: A method for comparing generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2621–2629.

Xudong Pan received the B.Sc. degree from Fudan University, China in 2018. He is currently a Ph.D. candidate in the School of Computer Science, Fudan University. His current research interests include theoretical and applied machine learning.

Mi Zhang received the B.Sc. degree from National University of Defense Technology, China in 2001, the M.Sc. degree from Fudan University, China in 2004 and the Ph.D. degree in computer science from University College Dublin in 2010. She is currently an associate professor in the School of Computer Science, Fudan University. Her research interests include theoretical and applied machine learning.

Daizong Ding received the B.Sc. degree from Fudan University, China in 2017. He is currently working towards the Ph.D. degree in the School of Computer Science, Fudan University. His current research interests cover machine learning, social network analysis and time series analysis.

Min Yang received the B.Sc. and the Ph.D. degrees in computer science from Fudan University in 2001 and 2006, respectively, where he is currently a professor in the School of Computer Science. His research interests include system security and AI security.