Theoretical Analysis of Image-to-Image Translation with Adversarial Learning

Xudong Pan [1] **Mi Zhang** [1] **Daizong Ding** [1]

Abstract

Recently, a unified model for image-to-image translation tasks within adversarial learning framework (Isola et al., 2017) has aroused widespread research interests in computer vision practitioners. Their reported empirical success however lacks solid theoretical interpretations for its inherent mechanism. In this paper, we reformulate their model from a brand-new geometrical perspective and have eventually reached a full interpretation on some interesting but unclear empirical phenomenons from their experiments. Furthermore, by extending the definition of generalization for generative adversarial nets (Arora et al., 2017) to a broader sense, we have derived a condition to control the generalization capability of their model. According to our derived condition, several practical suggestions have also been proposed on model design and dataset construction as a guidance for further empirical researches.

1. Introduction

Generative adversarial nets (GAN) (Goodfellow et al., 2014) have been a trending topic in machine learning community recent years, leading to a number of derived models (Mirza & Osindero, 2014; Arjovsky et al., 2017) and related theoretical works (Arjovsky & Bottou, 2017; Lei et al., 2017). With wide and fruitful applications in various scenarios such as speech synthesis (Saito et al., 2018), text generation (Zhang et al., 2017) and a considerable amount of visual tasks (Denton et al., 2015; Wu et al., 2016; Kataoka et al., 2017), the idea behind GAN and its derivations is relatively simple and intuitive. It aims at learning a mapping from an artificial distribution, usually priorly known gaussian for original GAN and an unknown distribution of *labels* for conditional GAN (Mirza & Osindero, 2014), to a real one. Via attaining an equilibrium of the minimax game (Aumann, 1989) between a *generator* (i.e. an adaptive model that maps a gaussian noise to a fake sample) and a *discriminator* (i.e. an adaptive model to distinguish a fake sample from a real distribution), the adversarial learning models will finally learn a realistic distribution for further generative tasks (Arora et al., 2017).

Noticeably, last year has also witnessed an empirical success on traditional image-to-image translation tasks with the aid of a model under conditional GAN paradigm (Isola et al., 2017), arousing widespread research interest in computer vision practitioners (Zheng et al., 2017; Choi et al., 2017). Image-to-image translation, as a generic name for various specific tasks in image processing, includes tasks such as facial expression transfer (e.g. poker face $\rightarrow$ smile face), artistic style transfer (e.g. Van Gogh's $\rightarrow$ Monet's). Generally speaking, the goal of image-to-image translation is to process an image from a source collection to make it indistinguishable among a target collection of images. Although related models and methods abound in literature (Reinhard et al., 2001; Gatys et al., 2016; Zeiler et al., 2011), the first attempt to tackle image-to-image translation as a whole instead of focusing on one of its specific tasks exclusively, ought to be attributed to the pioneering work of Isola et al. (2017), where the powerfulness of adversarial learning with conditional GAN has been once again exhibited sufficiently. We briefly review Isola's original model as the optimization problem below,

$$\min_G \max_D \mathcal{L}_{\text{cGAN}}(G, D) + \lambda \mathcal{L}_{L_1}(G) \tag{1}$$

with $\mathcal{L}_{\text{cGAN}}$ the *conditional GAN-loss* (or generally, *adversarial loss*) defined as

$$\mathcal{L}_{\text{cGAN}}(G, D) = \mathbb{E}_{x,y\sim p_r(x,y)}[\log D(x, y)] + \mathbb{E}_{x\sim p_r(x)}[1 - \log D(x, G(x))] \tag{2}$$

and $\mathcal{L}_{L_1}$ the *L1 loss* (or *identity loss*) as

$$\mathcal{L}_{L_1}(G) = \mathbb{E}_{x,y\sim p_r(x,y)}[\|y - G(x)\|_1] \tag{3}$$

where $p_r(x, y)$ denotes the distribution of *paired images* (e.g. in facial expression transfer, Bob's poker face and his ground-truth smile face) and $p_r(x)$ the distribution of images over the source collection, with G, D respectively the generator and the discriminator, λ the regularization factor.

[1]Shanghai Key Laboratory of Intelligent Information Processing, School of Computer Science, Fudan University, China. Correspondence to: Mi Zhang <mi_zhang@fudan.edu.cn>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018. Copyright 2018 by the author(s).

As reported in their experiments (Isola et al., 2017), several interesting but theoretically unclear results have attracted our attentions,

- Omitting the adversarial loss, i.e. solving $\mathcal{L}_{L_1}(G)$ alone, will "lead to reasonable but blurry results" (i.e. generated related target images, however with details hard to recognize), which we refer to as *Blurry versus Sharp*.

- Omitting the identity loss, i.e. setting the regularization factor λ to 0, "gives much sharper results, but results in some artifacts" (i.e. generated realistic images however unrelated to the given source image), which we refer to as *Source of Artifacts*.

Despite a number of studies devoted to analyzing and improving the training dynamics and generalization capability of GAN (Arjovsky & Bottou, 2017; Arora et al., 2017), there is rarely applicable theoretical results for analyzing conditional GAN, thus Isola's original model 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 (Lu et al., 1998), 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 have eventually reached some inspiring theoretical results.

In this paper, for the convenience of mathematical manipulation, we will study a slightly different objective from Isola's original model (Eq. 1), by substituting the ordinary conditional GAN loss (Eq. 2) with the Wasserstein GAN (WGAN) loss below as Eq. 4. As is well known, such a replacement is usually adopted by experimenters to stabilize the model's training dynamics (Arjovsky & Bottou, 2017).

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

where p_g denotes the distribution of images over the target collection, with $\Pi(p_r, p_g)$ the set of joint distributions for pairs of images (x, y) s.t. the marginal distributions are equal to p_r, p_g. Note the explicit term of discriminator in GAN (Eq. 2) is actually replaced by the inner optimal transport task (Villani, 2008) implicitly in WGAN loss (Eq. 4).

Therefore, the corresponding objective can be formulated as

$$\overbrace{\min_G \underbrace{\mathcal{L}_{\text{adv}}(G)}_{\text{adversarial loss}} + \lambda \underbrace{\mathcal{L}_{L_1}(G)}_{\text{identity loss}}}^{\text{target model}} \tag{5}$$

Aiming at exploring the intrinsic mechanism of our target model, we first formulate the image-to-image translation task with adversarial learning from a geometrical viewpoint (Section 2). With some basic results from topology and analysis, we have proved that the adversarial loss has an equivalent form (Eq. 16), degenerated as a set of individual learning tasks between paired *charts* (i.e. local neighborhoods homeomorphism to Euclidean space). We call such a result as *natural localization of adversarial loss* (Theorem 3.1). As a direct application of our theorem, theoretical interpretations has been presented fully for Source of Artifacts and partially for Blurry versus Sharp (Section 3.3).

In order to explore the properties of our target model more quantitatively, we have extended the definition of generalization for GAN proposed by Arora et al. (2017) to a broader case for analysis of conditional GAN (Definition 4.2). We have then pointed out the relation between generalizations in different senses with a generic inequality for the first time as far as we know (Theorem 4.1) and have finally obtained the full picture of the mechanism behind Blurry versus Sharp (Section 5.1).

As a step further, we have derived a condition on controlling generalization for our target model with additional statistical settings (Theorem 5.1). As we will see, the derived inequality (Eq. 27) imposes concrete constraints on both the sample complexity and model complexity, which provides practical guidance on model design and dataset construction for further applications (Section 5.2). Conclusions and future directions are provided in Section 6.

Generally, our contributions are outlined as follows,

1. A proposed geometrical formulation of image-to-image translation task with adversarial learning (Section 2)

2. Reduction of the adversarial loss to a set of independent learning tasks between paired charts (Theorem 3.1)

3. An extended definition of generalization for conditional GAN (Definition 4.2) and a derived condition on generalization (Theorem 5.1) for our target model

4. Theoretical interpretations for several unclear empirical phenomenons reported in previous works (Section 3.3 & 5.1), together with a practical guidance on model design and dataset construction for practitioners (Section 5.2)

2. Preliminaries

In Section 2.1 & 2.2, we equip a set of images with additional geometrical structures. In Section 2.3, we correspondingly reformulate image-to-image translation with adversarial learning by extending the concept of generator and discriminator. A reformulation of our target model will thus be given in Eq. 7 & 8 as a basis for analysis in the remainder of this paper.

2.1. Set of Images as Smooth Manifold

Without loss of generality, we mainly focus on the image-to-image translation task from a source set of RGB images $\mathcal{I}_S$ to a target set $\mathcal{I}_T$, with images of the same resolution $w \times h$. As is well-known, an image can always be considered as an element in an ambient Euclidean space (here, specifically $\mathbb{R}^{3\times w\times h}$). In fact, there also exists an intrinsic structure over the image set alongside with the ambient space, as is validated by various empirical works previously (Lu et al., 1998; Zhu et al., 2016). Such an intrinsic structure is usually formulated as a smooth manifold mathematically (Arjovsky & Bottou, 2017; Lei et al., 2017). For the basics of intrinsic geometry, see standard texts such as Lee's (2010).

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

Assumption 2.1. *There exist smooth, locally compact d-dimensional manifolds $\mathcal{M}, \mathcal{N}$ embedded in $\mathbb{R}^{w\times h}$, with constructed atlas as $\{(U_i, \varphi_i)\}_{i=1}^K$, $\{(V_i, \psi_i)\}_{i=1}^K$, respecting 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)*

As a comment, 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 Measure on Image Manifold

With Assumption 2.1, we are able to divide the image sets $I_S \subset \mathcal{M}$, $I_T \subset \mathcal{N}$ into finer subsets, formally, that is $\mathcal{I}_S = \cup_{k=1}^K \mathcal{I}_S^k$, $\mathcal{I}_T = \cup_{k=1}^K \mathcal{I}_T^k$, 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.

Assumption 2.2. *For each $k \in [K]$, there exists probability measures μ_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 $\mathbb{R}^d$.*

With probability measures defined on each chart (precisely, its homeomorphism as $\mathbb{R}^d$), we would like to "glue" them together to induce a unified probability measure (denoted respectively as $\mu_{\mathcal{M}}, \nu_{\mathcal{N}}$) globally over the underlying manifold structure, with the aid of the following lemma.

Lemma 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 $\mu_{\mathcal{M}} : \mathcal{B}(\mathcal{M}) \to [0, 1]$ is defined by*

$$d\mu_{\mathcal{M}}(s) = \frac{1}{K}\sum_{i=1}^{K} \mathbf{1}_{s\in U_i} d\mu_i \circ \varphi_i(s) \tag{6}$$

Then $\mu_{\mathcal{M}}$ is a probability measure defined on $\mathcal{M}$.

Proof. See Appendix A. Although Definition 6 seemingly contains some notation abusing (consider if $s \notin U_i$, $\varphi_i(s)$ is not defined), we can actually avoid this awkwardness according to the pairwise disjointness assumption, 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 Image-to-Image Translation with Adversarial Learning

After assuming additional geometrical structure on image set, the definition of generator and discriminator in our target model requires slight modifications correspondingly.

On Generator In our context, the generator should be redefined as a mapping between manifolds instead of flat Euclidean spaces. Formally, we denote the generator as $G : \mathcal{M} \to \mathcal{N}$, a measurable mapping w.r.t. $\mathcal{M}, \mathcal{N}$.

On Discriminator As we have pointed out, within the WGAN setting, the role of discriminator is correspondingly abdicated to the set of joint distributions $\Pi(p_r, p_g)$ and the norm $\| \bullet \|$ contained in Eq. 4. However, the latter is usually not well-defined in manifold settings. As a natural way to make Eq. 4 & 5 proper, we further equip the manifold structure $\mathcal{N}$ underlying the target set with a Riemmanian metric τ (Jost, 2008), with a little more technical conditions for regularity. Eventually it comes to our formulation of the adversarial loss and the corresponding identity loss, with relatively minor modifications compared with Eq. 3 & 5.

$$\mathcal{L}_{adv}'(G) = \inf_{\gamma\in\Pi(\mu_{\mathcal{M}},\nu_{\mathcal{N}})} \mathbb{E}_{(s,t)\sim\gamma} d_{\mathcal{N}}(G(s), t) \tag{7}$$

$$\mathcal{L}_{L_1}'(G) = \mathbb{E}_{s,t\sim p_r(s,t)} d_{\mathcal{N}}(G(s), t) \tag{8}$$

where $d_{\mathcal{N}}(\bullet, \bullet)$ denotes the τ-induced geodesic distance on $\mathcal{N}$ (Jost, 2008). For compatibility with inner-relatedness, we further assume $\forall i \neq j \in [K]$, $\forall x, y \in V_i, z \in V_j$, $d_{\mathcal{N}}(x, y) \leq d_{\mathcal{N}}(x, z)$.

3. Natural Localization of Adversarial Loss

A widely recognized difficulty for obtaining analytic solutions for adversarial loss lies in the nested optimization problem (Goodfellow et al., 2014) (here, specifically $\min_G \mathcal{L}_{\text{adv}}$).

In order to avoid such an obstacle, we will prove in this section that, within our proposed framework above, the inner infimum term in Eq. 7 could be solved in closed form with non-trivial constraints on candidate set of generator G (Theorem 3.1). Furthermore, we have observed that the closed form has a decomposition 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)), with the relations uniquely determined by the candidate sets (Eq. 16). This result directly leads to theoretical interpretations fully for Source of Artifacts and partially for Blurry versus Sharp (Section 3.3).

3.1. An Equivalent Form of $\mathcal{L}_{adv}^{'}$

We start our derivation by giving the explicit form of the probability measures $\mu_{\mathcal{M}}$, $\nu_{\mathcal{N}}$ on manifolds, constructed with the aid of Lemma 2.1.

$$d\mu_{\mathcal{M}}(s) \doteq \frac{1}{K}\sum_{i=1}^{K}\mathbf{1}_{s\in U_i} d\mu_i \circ \varphi_i(s) \tag{9}$$

$$d\nu_{\mathcal{N}}(t) \doteq \frac{1}{K}\sum_{i=1}^{K}\mathbf{1}_{t\in V_i} d\nu_i \circ \psi_i(t) \tag{10}$$

For simplicity, we denote $d\tilde{\mu}_i = d\mu_i \circ \varphi_i$ and $d\tilde{\nu}_i = d\nu_i \circ \psi_i$, $\forall i \in [K]$.

We then expand $\mathcal{L}_{adv}^{'}$ with assumed pairwise disjointness property and obtain

$$\inf_{\gamma\in\Pi(\mu_{\mathcal{M}},\nu_{\mathcal{N}})} \mathbb{E}_{(s,t)\sim\gamma}\sum_{i=1}^{K}\sum_{j=1}^{K}\mathbf{1}_{s\in U_i}\mathbf{1}_{t\in V_j} d_{\mathcal{N}}(G(s),t) \tag{11}$$

By exchanging the expectation operator with summations according to Fubini's theorem (Rudin, 2010) and writing the expectation directly in integral form, we have

$$\inf_{\gamma\in\Pi(\mu_{\mathcal{M}},\nu_{\mathcal{N}})} \sum_{i=1}^{K}\sum_{j=1}^{K}\int_{U_i}\int_{V_j} d_{\mathcal{N}}(G(s),t)d\gamma(s,t) \tag{12}$$

With a similar technique adopted in Dai et al. (2008), for every $\gamma \in \Pi(\mu_{\mathcal{M}}, \nu_{\mathcal{N}})$, there exist functions $\Delta : \mathcal{N}\times\mathcal{N} \to \mathbb{R}^+ \cup \{0\}$ and $f_\gamma : \mathcal{M} \to \mathcal{N}$, satisfying

$$d\gamma(s,t) = d\gamma(t|s)d\mu_{\mathcal{M}}(s) = \Delta(f_\gamma(s),t)d\mu_{\mathcal{M}}(s)d\nu_{\mathcal{N}}(t) \tag{13}$$

where Δ has an intuitive interpretation as a metric of similarity between elements on manifold $\mathcal{N}$, usually independent of the choice of path and compatible with inner-relatedness. Recall $\mathcal{N}$ as a Riemannian manifold is naturally equipped with a 'divergence' metric τ. We claim it is proper to absorb the term $\Delta(f_\gamma(s),t)$ into $d_{\mathcal{N}}(G(s),t)$ with the following observations.

a) Equivalence of optimization problems (without boundary condition) (Boyd & Vandenberghe, 2004)
   - $\min_G \min_{f_\gamma} \Delta(f_\gamma(s),t)d_{\mathcal{N}}(G(s),t)$
   - $\min_G \Delta(G(s),t)d_{\mathcal{N}}(G(s),t)$

   considering the relatively large learning capacity of G, usually implemented as a neural network (Sontag, 1998).

b) It is possible to alter the choice of the original metric τ to be the metric $\tau^{'}$ induced by a modified distance function $d_{\mathcal{N}}^{'}(\bullet,\bullet) = \Delta(\bullet,\bullet)d_{\mathcal{N}}(\bullet,\bullet)$, which is asserted by the following lemma.

Lemma 3.1. *Consider Riemannian manifold $(\mathcal{N},\tau)$ with curvature locally bounded above and below, $\tau \in C^\infty$ and its induced distance function denoted as $d_{\mathcal{N}}$, then for any path-independent function $f : \mathcal{N}\times\mathcal{N} \to \mathbb{R}^+ \cup \{0\}$, there exists a Riemannian metric $\tau^{'}$ on $\mathcal{N}$, induced by the distance function*

$$d_{\mathcal{N}}^{'}(x,y) = f(x,y)d_{\mathcal{N}}(x,y) \quad \forall x,y \in \mathcal{N}$$

Proof. See Appendix A. □

After we rearrange $d_{\mathcal{N}}(G(s),t)d\gamma$ as $d_{\mathcal{N}}^{'}(G(s),t)d\gamma'$, 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 $\min_G \mathcal{L}_{adv}^{'}(G)$ can be reformulated as (following Eq. 9, 10, 12, with details in Appendix A)

$$\min_G \min_{A\in\mathbf{H}(K)} \sum_{i=1}^{K}\sum_{j=1}^{K}\int_{U_i}\int_{V_j} A^{ij} d_{\mathcal{N}}^{'}(G(s),t)d\tilde{\mu}_i(s)d\tilde{\nu}_j(t) \tag{14}$$

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

As is discussed above, the form of Eq. 14 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. $A \in \mathbf{H}(K)$. Although it is almost infeasible to obtain a closed form solution for arbitrary mapping G, we are curious of the possibility by imposing non-trivial constraints on the candidate set of G. In our approach, we first propose the following definition.

Definition 3.1 (pairwise topological immersion family (PTI–family)). *Given topological manifolds $\mathcal{M} = \{(U_i,\varphi_i)\}_{i=1}^K$ and $\mathcal{N} = \{(V_i,\psi_i)\}_{i=1}^K$, 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 $[K]$ (Cameron, 1999), is called pairwise topological immersion mappings indexed by p, w.r.t $\mathcal{M}$, $\mathcal{N}$.*

Postponing remarks on this definition (Section 3.3.1), a main result of this paper will be provided subsequently, which shows that, we can indeed obtain a meaningful closed form of solution for the inner minimization problem, by constraining the candidate set of G as an arbitrary PTI-family (Def. 3.1).

Theorem 3.1. *[Natural Localization of Adversarial Loss] For any $p \in Sym(K)$, the optimization problem below (compared with Eq. 14)*

$$\min_{G \in F_p} \min_{A \in \mathbf{H}(K)} \sum_{i=1}^{K} \sum_{j=1}^{K} \int_{U_i} \int_{V_j} A^{ij} d'_{\mathcal{N}}(G(s), t) d\tilde{\mu}_i(s) d\tilde{\nu}_j(t) \tag{15}$$

is equivalent to

$$\min_{G \in F_p} \sum_{i=1}^{K} \int_{U_i} \int_{V_{p(i)}} d'_{\mathcal{N}}(G(s), t) d\tilde{\mu}_i(s) d\tilde{\nu}_{p(i)}(t) \tag{16}$$

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.

Sketch of Proof. Fix $i, j \in [K]$, s.t. $j \neq p(i)$ and arbitrary $G \in F_p$. The key step lies in comparing the terms (remember the positiveness of A^{ij})

$$T_{\text{non-paired}} = \int_{U_i} \int_{V_j} d'_{\mathcal{N}}(G(s), t) d\tilde{\mu}_i(s) d\tilde{\nu}_j(t)$$

and

$$T_{\text{paired}} = \int_{U_i} \int_{V_{p(i)}} d'_{\mathcal{N}}(G(s), t) d\tilde{\mu}_i(s) d\tilde{\nu}_{p(i)}(t)$$

For any $s \in U_i$, $G(s) \in V_{p(i)} \cap V_j = \varnothing$, which leads to $\forall t \in V_{p(i)}, t' \in V_j$, $d_{\mathcal{N}}(G(s), t) \leq d_{\mathcal{N}}(G(s), t')$. And thus $T_{\text{non-paired}} \geq T_{\text{paired}}$. A rigorous proof can be found in Appendix A. □

3.3. Discussions & Interpretations

The final part of this section is devoted to a discussion on what Definition 3.1 and Theorem 3.1 actually mean, together with their significant roles in interpretations for the empirical phenomenons.

3.3.1. Discussion 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 $\{\mu_i\}_{i=1}^K$, $\{\nu_i\}_{i=1}^K$. For example, in facial expression translation tasks (Choi et al., 2017), U_i contains a set of Bob's poker face, while V_j, V_k are respectively sets of Alice's and Bob's face with smile. A PTI-family F_p exactly characterizes the generating tendency of a given generator G. Let us come back to the example. Fix $i \neq j \neq k$. Assume $p, q \in Sym(K)$ with $p(i) = j$ while $q(i) = k$. Thus with the input as an image of Bob's poker face, generators from F_p tends to generate a sample of Alice's smiling face, while those from F_q prefer a sample of smiling Bob. Note that, 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 Theorem 3.1, which is not just an intention to give a closed form for further analysis. More essentially, such a theorem points out the role of $\{F_p\}_{p \in \text{Sym(K)}}$ as *attractors* (for attractors in a general sense, see Luenberger's (1979)) 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. 15) will immediately degenerate to learning tasks on paired charts $\{(U_i, V_{p(i)})\}_{i=1}^K$ (Eq. 16). The generator will thus be trapped in the subset F_p until the end of the training. This theorem can be considered as a support to a recent result called *imaginary adversary*, which points out that in WGAN setting, the minimax game between generator and discriminator can be resolved under some technical conditions (Lei et al., 2017).

3.3.2. Partial Interpretations for Empirical Results

Source of Artifacts Although it brings sharper results with the adversarial loss, a non-negligible proportion of artifacts is observed in experiments (Isola et al., 2017; Choi et al., 2017). 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 to mistake. 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 clever approach is by imposing oracle as a regularization term with L1-loss (Eq. 8), which plays the role as a *rectifier* for choice of p.

Blurry versus Sharp In previous empirical studies, after learning with identity loss (Eq. 8) alone, the final generator usually produces more blurry images compared with the generator after learning with the adversarial loss (Eq. 7). 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 Theorem 3.1, learning with adversarial loss theoretically only requires to learn 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* (Rudin, 2010). Intuitively, learning local mappings independently requires much smaller capacity of G, compared with learning a globally compatible one (a theoretical justification, see Proposition 5.1). Recently, a model with a similar consideration by *artificially* localizing the adversarial loss to improve the generating quality was proposed (Qi et al., 2017). However, their work mainly targets on image generation (i.e. only the target manifold structure is considered) and stays on an empirical level, with little theoretical analysis for the inherent mechanism.

As a complement and a step further, we will provide a formal analysis on the benefit of localization detailedly (Section 5.1) to complete our interpretations. Due to the indispensable role of the concept of generalization in analyzing model's learning capability (Vapnik & Vapnik, 1998), we will first present an extended definition of generalization for conditional GAN in the next section.

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, there have been previous efforts on proposing specific definitions for GAN considering its difference from conventions. One of these definitions is provided as follows, with our notations.

Definition 4.1. *(Arora et al., 2017) [Generalization w.r.t Divergence] A divergence $D(\bullet,\bullet)$ is said to generalize with m training samples and error ϵ if for the learned distribution $\nu_{\mathcal{N}}$, the following inequality holds with high probability,*

$$|D(\hat{\nu}_{real}, \hat{\nu}_G) - D(\nu_{real}, \nu_G)| < \epsilon \tag{17}$$

where $\hat{\nu}_{real}, \hat{\nu}_G$ are empirical versions of real and generated distributions with ν_{real} the real distribution as ground-truth.

Although their work marks the first attempt to study the generalization capability of GAN, such a definition has several potential shortcomings: **a)** generalization is defined w.r.t specific divergence, instead of the generator itself. From our perspective, it is still the generator that holds the fundamental position in generative tasks. **b)** lack of the extensibility to conditional GAN, which however plays an increasingly significant role in empirical research and applications. Such a deficiency directly makes it improper to be applied to analyze our target model.

In order to alleviate these possible downsides, we propose an extended version of generalization for both GAN and its deviations with respect to a learned generator.

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

$$D(G(\hat{\mu}^m_{\mathcal{M}}), \nu_{\mathcal{N}}) - D(\hat{\nu}^n_{\mathcal{N}}, \nu_{\mathcal{N}}) < \epsilon \tag{18}$$

where $\hat{\mu}^m_{\mathcal{M}}, \hat{\nu}^n_{\mathcal{N}}$ are estimators of source and target distributions, with $\mu_{\mathcal{M}}, \nu_{\mathcal{N}}$ the corresponding ground-truth distributions and $G(\hat{\mu}^m_{\mathcal{M}}) \doteq \hat{\mu}^m_{\mathcal{M}} \circ G^{-1}$, the induced distribution on $\mathcal{N}$ (Chung, 2001).

Compared with Definition 4.1, our extension explicitly contains the generator as an essential factor for generalization. Furthermore, instead of assuming the source distribution as a gaussian priorly known, we depict it with an empirical estimator from m observed samples. Notice our definition is actually an extension of Definition 4.1, since, by limiting m to infinity and assuming G of sufficient learning capability (in a classical sense), Inequality 18 will directly degenerate to Inequality 17 in the previous definition.

4.2. Relations of Generalization in Different Senses

As an auxiliary theorem for further analysis of our target model in the next section, we will derive the relation of generalization in different senses as well.

We first specify the divergence $D(\bullet,\bullet)$ in Definition 4.2 as Lukaszyk-Karmowski metric (Łukaszyk, 2004)

$$D_{\text{LK}}(\nu, \nu^{'}) = \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \|x - x^{'}\| d\nu(x) d\nu^{'}(x^{'}) \tag{19}$$

where $\nu, \nu^{'}$ are arbitrary probability measures supported on $\mathbb{R}^d$ (compared with Eq. 16). Note the Euclidean form above brings convenience for analysis and it actually only requires several technical steps to extend the following result to an intrinsic form (Lemma 5.1).

Theorem 4.1. *Consider generator $G : \mathbb{R}^d \to \mathbb{R}^d$ satisfying Lipschitz condition with constant M_G and μ_X, ν_Y are probability measures on $\mathbb{R}^d$ 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 probability $1-\delta$

$$\mathbb{E}_{x\sim\mu_X, y\sim\nu_Y}\|G(x) - y\| - \sum_{i=1}^{n_X}\sum_{j=1}^{n_Y}\frac{\|G(x_i) - y_j\|}{n_X n_Y} < \epsilon_{classical} \tag{20}$$

where $\epsilon_{classical} \doteq \epsilon(n_X, n_Y, \mu_X, \nu_Y, \delta)$ the upper bound and ERM-principle (Vapnik & Vapnik, 1998) 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 error ϵ_{adv} with probability $1-\delta$, i.e.

$$D_{LK}(G(\hat{\mu}^{n_X}_X), \nu_Y) - D_{LK}(\hat{\nu}^{n_Y}_Y, \nu_Y) < \epsilon_{adv} \tag{21}$$

if the following condition is satisfied

$$\epsilon_{classical} - \epsilon_{adv} + \eta < D_{LK}(\nu_Y, \hat{\nu}_Y^{n_Y}) - M_G D_{LK}(\mu_X, \hat{\mu}_X^{n_X}) \tag{22}$$

Proof. See Appendix A. ☐

As Theorem 4.1 indicates, unlike the classical generalization bound (especially in VC sense (Vapnik & Vapnik, 1998)), the generalization error in adversarial learning is also affected by the variation of distributions in source and target distributions.

5. Benefits of Localization and Conditions of Generalization

By auxiliary of the extended definition of generalization above, we are now able to complete our interpretations for Blurry versus Sharp (Section 5.1). As a step further, we will derive a concrete condition (Theorem 5.1) to control the generalization capability of our target model, which will directly provide practical guidance on model design and dataset construction for practitioners.

We start by specifying some additional statistical settings, only for the sake of concreteness. 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 gaussian settings *locally*.

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 $i \in [K]$, $\mu_i = \mathcal{N}(\bullet; x_i, \Sigma_\mathcal{M})$, $\nu_i = \mathcal{N}(\bullet; y_i, \Sigma_\mathcal{N})$, where $\mathcal{N}(\bullet; x, \Sigma)$ 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 much limitation on our discussion, mainly because its influence remains local (compared with original GAN (Goodfellow et al., 2014)) and each gaussian is partially unknown (compared with Arora et al. (2017)).

5.1. Benefits of Localization

In our previous interpretation for Blurry versus Sharp (Section 3.3), a claim has remained unjustified that learning a set of local mappings is much easier compared with learning a globally compatible one. With the following observations: **a)** Lipschitz condition can be always satisfied with techniques during training phase (Arjovsky et al., 2017). **b)** $\epsilon_{\text{classical}}, \eta, M_G$ remain constant for the same hypothesis space. **c)** The target-related term $D_{LK}(\nu, \hat{\nu})$ is identical in local and global task when the pairing relation is unobserved, we reformulate Inequality 22 as

$$C + \lambda D_{LK}(\mu, \hat{\mu}) < \epsilon_{\text{adv}} \tag{23}$$

where $C \doteq \epsilon_{\text{classical}} + \eta - D_{LK}(\nu, \hat{\nu})$ a constant and $\lambda \doteq M_G > 0$.

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 Euclidean sense, it is sufficient to compare the two terms below to justify our previous claim.

$$\epsilon_{adv}^{local} = \frac{1}{K}\sum_{i=1}^K D_{LK}(\mu_i, \hat{\mu}_i^m) \tag{24}$$

$$\epsilon_{adv}^{global} = D_{LK}(\frac{1}{K}\sum_{i=1}^K \mu_i, \hat{\mu}_X^{Km}) \tag{25}$$

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

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

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

Proof. See Appendix A. Intuitively, let us consider an extremal situation when $\mu_i = \delta_{x_i}, \nu_i = \delta_{y_i}$ and $m \to \infty$ (δ_x is the Dirac function as a distribution) for each $i \in [K]$. Thus $\epsilon_{adv}^{local} = 0$, while $\epsilon_{adv}^{global} \geq \sup_{i,j\in[K]} \|x_i - x_j\| > 0$. ☐

5.2. Conditions of Generalization

As a step further, we would like to derive some technical conditions under which the target model will generalize well, in the sense of adversarial learning (Definition 4.2).

In order to apply Theorem 4.1, we introduce the following lemma which points out the equivalence between L-K metric (Eq. 19) with assumed probability measures in Euclidean space and each local objective defined on manifolds.

Lemma 5.1. $\forall i \in [K]$, *consider a measureable mapping $\tilde{f} : U_i \to V_i$ with $f \doteq \psi_i \circ \tilde{f} \circ \varphi_i^{-1}$ satisfies Lipschitz condition, then $\int_{U_i}\int_{V_i} d_\mathcal{N}'(\tilde{f}(s), t) d\tilde{\mu}_i(s) d\tilde{\nu}_i(t) \simeq D_{LK}(f(\mu_i), \nu_i)$, i.e. there exists constants $0 < C_l < C_u < \infty$ such that*

$$C_l < \frac{\int_{U_i}\int_{V_i} d_\mathcal{N}'(\tilde{f}(s), t) d\tilde{\mu}_i(s) d\tilde{\nu}_i(t)}{D_{LK}(f(\mu_i), \nu_i)} < C_u \tag{26}$$

Sketch of Proof. The key observation lies in, with the measureability of $\tilde{f}$ and smoothness of φ_i, ψ_i, the induced mapping $\tilde{\nu}^{'} = \frac{1}{\tilde{E}}\tilde{f}(\mu_i)$ and $\nu^{'} = \frac{1}{E}(\psi_i \circ \tilde{f})(\mu_i)$ are also probability measures respectively on $\tilde{f}(U_i) \subset V_i$ and $(\psi_i \circ \tilde{f})(U_i) \subset \psi_i(V_i)$ (with $\tilde{E}, E$ some normalizing factor), with bounded range a.e., which basically comes from the Lipschitz condition. We also use the assumption $\tilde{f}(U_i) \subset V_i$ and $tr(\Sigma_\mathcal{N}) < \infty$. For the rest of the proof, see Appendix A. □

We are now able to instantiate the generic inequality on generalization in the form of the following theorem. Note we depict the classical generalization error term in VC sense and study the condition for $\epsilon_{adv} = 0$, which means 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 VC-dimension bound by 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 constant M_G, then G generalizes globally with (Kn, Km) samples only if the following inequality is satisfied with probability $1 - C(\epsilon, \Lambda)(nm\epsilon^2)^{\tau(\Lambda)}e^{-nm\alpha\epsilon^2}$,*

$$\epsilon + \frac{1}{nm}\max\{\sum_{i=1}^{n}\sum_{j=1}^{m} d_\mathcal{N}(G(s_k^i), t_k^j)\}_{k=1}^{K} <$$
$$\frac{1}{\sqrt{m}}\sqrt{tr(\Sigma_\mathcal{N})} + 2tr(\Sigma_\mathcal{N}) - M_G(\frac{1}{\sqrt{n}}\sqrt{tr(\Sigma_\mathcal{M})} + 2tr(\Sigma_\mathcal{M})) \tag{27}$$

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

Proof. See Appendix A. Besides Theorem 4.1 & Lemma 5.1, we have applied a general form of Vapnik-Chervonenkis theorem (Vayatis & Azencott, 1999) for worst case analysis and a non-asymptotic theorem from information geometry as follows,

Theorem 5.2. *(Amari & Nagaoka, 2007) The mean square error of a biased-corrected first-order efficient estimator $\hat{u}$ to μ is given by the expansion (with N observed samples):*

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

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

□

Discussions & Guidance for Practitioners A brief discussion on Theorem 5.1 and its possible guidance on practice will serve as the last topic. As we can see, generalization happens with a higher probability when the right side of Inequality 27 yields larger and the left side becomes smaller. The former situation corresponds to a smaller variance of each local source distribution, especially when M_G the Lipschitz constant lets the tr$(\Sigma_\mathcal{M})$ term dominate. The latter situation corresponds to a *uniformly* lower empirical risk. As each local chart has an intuitive interpretation as a set of related images, it is reasonable to make the following suggestions on dataset construction and model design.

- The source set of images should be of lower inner-similarity, i.e. a set of N different individuals' poker face will give a better generator rather than a set of N different photos of the same person's poker face.
- A blind increase in total number of images will hardly help generalization, while the balancedness in numbers of different objects is what actually matters.
- Classical generalization capacity (Vapnik & Vapnik, 1998) and smoothness of learning model w.r.t. data manifolds (Belkin et al., 2006) should be considered equivalently important in model design for such tasks.

6. Conclusion and Further Directions

In this paper, we have focused on providing a solid theoretical interpretations for some critical but unclear empirical phenomenons reported in Isola et. al (2017). Via reformulating Isola's model within a brand-new geometrical framework (Section 2), we have proved that the target model has a natural localized form as independent learning tasks on paired charts (Theorem 3.1), which directly provides a candidate interpretation for their experimental results (Section 3.3). Furthermore, with our extension of the generalization concept for GAN to conditional GAN case (Definition 4.2), we have successfully described the inherent mechanism of the target model in a full picture (Section 5.1). Our derived generalization condition (Theorem 5.1) also provides constructive guidance for further empirical studies (Section 5.2).

Actually, our theoretical results can be easily decoupled from the image-to-image translation setting to a much general case, that is, learning translation from a source manifold structure to a target one via adversarial learning. Further directions in applications, such as applying our theoretical results for improving the current models or devising new architectures for better generating and translation performance, are potentially fruitful. For theorists, our framework for analysis can be considered as an attempt to understand the far-more complicated mechanism behind adversarial learning models in a specific context. More exciting theoretical results based on our theoretical framework awaits further dedications.

Acknowledgements

The authors would like to thank the anonymous referees for their valuable comments and helpful suggestions. This work is funded by the National Program on Key Basic Research (NO. 2015CB358800).

References

Amari, S.-i. and Nagaoka, H. *Methods of information geometry*, volume 191. American Mathematical Soc., 2007.

Arjovsky, M. and Bottou, L. Towards principled methods for training generative adversarial networks. *CoRR*, abs/1701.04862, 2017.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *ICML*, 2017.

Arora, S., Ge, R., Liang, Y., Ma, T., and Zhang, Y. Generalization and equilibrium in generative adversarial nets (gans). In *ICML*, 2017.

Aumann, R. J. Game theory. In *Game Theory*, pp. 1–53. Springer, 1989.

Belkin, M., Niyogi, P., and Sindhwani, V. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7:2399–2434, 2006.

Boyd, S. and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2004.

Cameron, P. J. *Permutation groups*, volume 45. Cambridge University Press, 1999.

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.

Chung, K. L. *A course in probability theory*. Academic press, 2001.

Dai, W., Chen, Y., Xue, G.-R., Yang, Q., and Yu, Y. Translated learning: Transfer learning across different feature spaces. In *NIPS*, 2008.

Denton, E. L., Chintala, S., Szlam, A., and Fergus, R. Deep generative image models using a laplacian pyramid of adversarial networks. In *NIPS*, 2015.

Gatys, L. A., Ecker, A. S., and Bethge, M. Image style transfer using convolutional neural networks. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2414–2423, 2016.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., and Bengio, Y. Generative adversarial nets. In *NIPS*, 2014.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. Image-to-image translation with conditional adversarial networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5967–5976, 2017.

Jost, J. *Riemannian geometry and geometric analysis*. Springer Science & Business Media, 2008.

Kataoka, Y., Matsubara, T., and Uehara, K. Automatic manga colorization with color style by generative adversarial nets. *2017 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, pp. 495–499, 2017.

Lee, J. *Introduction to topological manifolds*, volume 940. Springer Science & Business Media, 2010.

Lei, N., Su, K., Cui, L., Yau, S.-T., and Gu, X. A geometric view of optimal transportation and generative model. *CoRR*, abs/1710.05488, 2017.

Lu, H., Fainman, Y., and Hecht-Nielsen, R. Image manifolds. 1998.

Luenberger, D. G. *Introduction to dynamic systems: theory, models, and applications*, volume 1. Wiley New York, 1979.

Łukaszyk, S. A new concept of probability metric and its applications in approximation of scattered data sets. *Computational Mechanics*, 33(4):299–304, 2004.

Mirza, M. and Osindero, S. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014.

Qi, G.-J., Zhang, L., and Hu, H. Global versus localized generative adversarial nets. *CoRR*, abs/1711.06020, 2017.

Reinhard, E., Ashikhmin, M., Gooch, B., and Shirley, P. Color transfer between images. *IEEE Computer Graphics and Applications*, 21:34–41, 2001.

Rudin, W. Real and complex analysis real and complex analysis third edition. 2010.

Saito, Y., Takamichi, S., and Saruwatari, H. Statistical parametric speech synthesis incorporating generative adversarial networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26:84–96, 2018.

Sontag, E. D. Vc dimension of neural networks. *NATO ASI Series F Computer and Systems Sciences*, 168:69–96, 1998.

Vapnik, V. N. and Vapnik, V. *Statistical learning theory*, volume 1. Wiley New York, 1998.

Vayatis, N. and Azencott, R. Distribution-dependent vapnik-chervonenkis bounds. In *EuroCOLT*, 1999.

Villani, C. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

Wu, J., Zhang, C., Xue, T., Freeman, B., and Tenenbaum, J. B. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *NIPS*, 2016.

Zeiler, M. D., Taylor, G. W., Sigal, L., Matthews, I. A., and Fergus, R. Facial expression transfer with input-output temporal restricted boltzmann machines. In *NIPS*, 2011.

Zhang, Y., Gan, Z., Fan, K., Chen, Z., Henao, R., Shen, D., and Carin, L. Adversarial feature matching for text generation. In *ICML*, 2017.

Zheng, Z., Zheng, H., Yu, Z., Gu, Z., and Zheng, B. Photo-to-caricature translation on faces in the wild. *CoRR*, abs/1711.10735, 2017.

Zhu, J.-Y., Krähenbühl, P., Shechtman, E., and Efros, A. A. Generative visual manipulation on the natural image manifold. In *ECCV*, 2016.