Learning Generative Models across Incomparable Spaces

Charlotte Bunne [1] **David Alvarez-Melis** [2] **Andreas Krause** [1] **Stefanie Jegelka** [2]

Abstract

Generative Adversarial Networks have shown remarkable success in learning a distribution that faithfully recovers a reference distribution *in its entirety*. However, in some cases, we may want to only learn some aspects (e.g., cluster or manifold structure), while modifying others (e.g., style, orientation or dimension). In this work, we propose an approach to learn generative models across such *incomparable* spaces, and demonstrate how to steer the learned distribution towards target properties. A key component of our model is the *Gromov-Wasserstein* distance, a notion of discrepancy that compares distributions *relationally* rather than absolutely. While this framework subsumes current generative models in identically reproducing distributions, its inherent flexibility allows application to tasks in manifold learning, relational learning and cross-domain learning.

1. Introduction

Generative Adversarial Networks (GANs, Goodfellow et al. (2014)) and its variations (Radford et al., 2016; Arjovsky et al., 2017; Li et al., 2017) are powerful models for learning complex distributions. Broadly, these methods rely on an *adversary* that compares samples from the true and learned distributions, giving rise to a notion of divergence between them. The divergences implied by current methods require the two distributions to be supported in sets that are *identical* or at the very least *comparable*; examples include Optimal Transport (OT) distances (Salimans et al., 2018; Genevay et al., 2018) or Integral Probability Metrics (IPM) (Müller, 1997; Sriperumbudur et al., 2012; Mroueh et al., 2017). In all of these cases, the spaces over which the distributions are defined must have the same dimensionality (e.g., the space of 28×28-pixel vectors for MNIST), and the generated distribution that minimizes the objective has the same support as the reference one. This is of course desirable when the goal is to generate samples that are *indistinguishable* from those of the reference distribution.

Many other applications, however, require modeling only topological or relational aspects of the reference distribution. In such cases, the absolute location of the data manifold is irrelevant (e.g., distributions over learned representations, such as word embeddings, are defined only up to rotations), or it is not available (e.g., if the data is accessible only as a weighted graph indicating similarities among sample points). Another reason for modeling only topological aspects is the desire to, e.g., change the appearance or style of the samples, or down-scale images. Divergences that directly compare samples from the two distributions, and hence most current generative models, do not apply to those settings.

In this work, we develop a novel class of generative models that can learn across *incomparable* spaces, e.g., spaces of different dimensionality or data type. Here, the relational information between samples, i.e., the topology of the reference data manifold, is preserved, but other characteristics, such as the ambient dimension, can vary. A key component of our approach is the *Gromov-Wasserstein* (GW) distance (Mémoli, 2011), a generalization of classic Optimal Transport distances to incomparable ground spaces. Instead of directly comparing points in the two spaces, the GW distance computes pairwise intra-space distances, and compares those distances across spaces, greatly increasing the modeling scope. Figure 1 illustrates the new model.

To realize this model, we address several challenges. First, we enable the use of the Gromov-Wasserstein distance in various learning settings by improving its robustness and ensuring unbiased learning. Similar to existing OT-based generative models (Salimans et al., 2018; Genevay et al., 2018), we leverage the differentiability of this distance to provide gradients for the generator. Second, for efficiency, we further parametrize it via a learnable adversary. The added flexibility of the GW distance necessitates to constrain the adversary. To this end, we propose a novel orthogonality regularization, which might be of independent interest.

A final challenge —which doubles as one of the main ad-

[1]Department of Computer Science, Eidgenössische Technische Hochschule (ETH), Zürich, Switzerland [2]Computer Science and Artificial Intelligence Laboratory (CSAIL), Massachusetts Institute of Technology (MIT), Cambridge, USA. Correspondence to: Charlotte Bunne <bunnec@ethz.ch>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

Figure 1. The Gromov-Wasserstein generative adversarial network (GW GAN) learns across incomparable spaces, such as different dimensions or data type (from graphs to Euclidean space). The key idea is that its learning objective is purely based on intra-space distances (e.g., pairwise distances d or shortest paths sp) in the generator and data space, respectively.

vantages of this approach— arises from the added flexibility of the generator: it allows to freely alter superficial characteristics of the generated distribution while still learning the basic structure of the reference distribution. We show examples how to steer these additional degrees of freedom via regularization or adversaries in the model. The resulting model subsumes the traditional (i.e., same-space) adversarial models as a special case, but can do much more. For example, it learns cluster structure across spaces of different dimensionality and across different data types, e.g., from graphs to Euclidean space. Thus, our GW GAN can also be viewed as performing dimensionality reduction or manifold learning, but, departing from classical approaches to these problems, it recovers, in addition to the manifold structure of the data, the probability distribution defined over it. Moreover, we propose a general framework for stylistic modifications by integrating a style adversary; we demonstrate its use by changing the thickness of learned MNIST digits. In summary, this work provides a framework that substantially expands the potential applications of generative adversarial learning.

Contributions. We make the following contributions:

i. We introduce a new class of generative models that can learn distributions across different dimensionalities or data types.
ii. We demonstrate the model's range of applications by deploying it to manifold learning, relational learning and cross-domain learning tasks.
iii. More generally, our modifications of the Gromov-Wasserstein discrepancy enable its use as a loss function in various machine learning applications.
iv. Our new approach to approximately enforce orthogonality in neural networks based on the orthogonal Procrustes problem also applies beyond our model.

2. Model

Given a dataset of n observations $\{x_1, ..., x_n\}, x_i \in \mathcal{X}$ drawn from a reference distribution $p \in \mathcal{P}(\mathcal{X})$, we aim to learn a generative model g_θ parametrized by θ purely based on relational and intra-structural characteristics of the dataset. The generative model $g_\theta : \mathcal{Z} \to \mathcal{Y}$, typically a neural network, maps random noise $z \in \mathcal{Z}$ to a generator space $\mathcal{Y}$ that is independent of data space $\mathcal{X}$.

2.1. Gromov-Wasserstein Discrepancy

Learning generative models typically relies on a statistical divergence between the target distribution and the model's current estimate. Classical statistical divergences only apply when comparing distributions whose supports lie in the same metric space, or when at least a meaningful distance between points in the two supports can be computed. When the data space $\mathcal{X}$ and generator space $\mathcal{Y}$ are different, these divergences no longer apply.

Hence, instead, we will use a more suitable divergence measure. Rather than relying on a metric *across* the spaces, the *Gromov-Wasserstein (GW)* distance (Mémoli, 2011) compares distributions by computing a discrepancy between the metrics defined *within* each of the spaces. As a consequence, it is oblivious to specific characteristics or the dimensionality of the spaces.

Given n samples of the compared distributions $p \in \mathcal{P}(\mathcal{X})$ and $q \in \mathcal{P}(\mathcal{Y})$, the discrete formulation of the GW distance needs a similarity (or distance) matrix between the samples and a probability vector for each space, say $(D, \mathbf{p})$ and $(\bar{D}, \mathbf{q})$, with $(D, \mathbf{p}) \in \mathbb{R}^{n\times n} \times \Sigma_n$, where $\Sigma_n := \{\mathbf{p} \in \mathbb{R}^+_n; \sum_i \mathbf{p}_i = 1\}$ is the n-dimensional probability simplex. Then, the GW discrepancy is

$$\begin{aligned} \mathrm{GW}(D, \bar{D}, \mathbf{p}, \mathbf{q}) &:= \min_{T \in \mathcal{U}_{\mathbf{p},\mathbf{q}}} \mathcal{E}_{D,\bar{D}}(T) \\ &:= \min_{T \in \mathcal{U}_{\mathbf{p},\mathbf{q}}} \sum_{ijkl} L(D_{ik}, \bar{D}_{jl}) T_{ij} T_{kl}, \end{aligned} \tag{1}$$

where $\mathcal{U}_{\mathbf{p},\mathbf{q}} := \{T \in (\mathbb{R}_+)^{n\times n}; T\mathbb{1}_n = \mathbf{p}, T^\top \mathbb{1}_n = \mathbf{q}\}$ is the set of all couplings T between $\mathbf{p}$ and $\mathbf{q}$. The loss function L in our case is $L(a, b) = L_2(a, b) := \frac{1}{2}|a - b|^2$. If $L = L_2$, then $\mathrm{GW}^{1/2}$ defines a (true) distance (Mémoli, 2011).

2.2. Gromov-Wasserstein Generative Model

To learn across incomparable spaces, one key idea of our model is to use the Gromov-Wasserstein distance as a loss

Algorithm 1 Training Algorithm of the Gromov-Wasserstein Generative Model.

Require: α: learning rate, n_g: the number of iterations of the generator per adversary iteration, m: mini-batch size, N: number of training iterations, θ_0: initial parameters of generator g_θ, $\omega_0 = (\check{\omega}_0, \hat{\omega}_0)$: initial parameters of adversary f_ω

for $t = 0$ to N **do**
  sample $X = (x_i)_{i=1}^m$ from dataset
  sample $Z = (z_j)_{j=1}^m \sim \mathcal{N}(0,1)$, $Y = (y_j)_{j=1}^m = g_{\theta_t}((z_j)_{j=1}^m)$
  $\forall (i,j), D^{\check{\omega}}_{ij} := \|f_{\check{\omega}}(x_i) - f_{\check{\omega}}(x_j)\|_2$ and $D^{\hat{\omega}}_{ij} := \|f_{\hat{\omega}}(y_i) - f_{\hat{\omega}}(y_j)\|_2$
  $\mathcal{L} = \overline{\mathrm{GW}}_\epsilon(D^{\check{\omega}}, D^{\hat{\omega}}, \mathbf{p}, \mathbf{q})$, where $\mathbf{p}, \mathbf{q}$ are uniform distributions ▷ $\overline{\mathrm{GW}}_\epsilon$ is defined in Eq. (7)
  if $t \mod n_g + 1 = 0$ **then**
    $\mathcal{L}_{\mathrm{reg}} \leftarrow \mathcal{L} - R_\beta(f_{\check{\omega}}(X), X) - R_\beta(f_{\hat{\omega}}(Y), Y)$ ▷ R_β is defined in Eq. (5)
    $\omega_{t+1} \leftarrow \omega_t + \alpha \times \nabla_{\omega_t} \mathcal{L}_{\mathrm{reg}}$
  else
    $\theta_{t+1} \leftarrow \theta_t - \alpha \times \nabla_{\theta_t} \mathcal{L}$
  end if
end for

function to compare the generated and true distribution. As in traditional adversarial approaches, we parametrize the generator $g_\theta : \mathcal{Z} \to \mathcal{Y}$ as a neural network that maps noise samples z to features y. We train g_θ by using GW as a loss, i.e., for mini-batches X and Y of reference and generated samples, respectively, we compute pairwise distance matrices D and $\bar{D}$ and solve the GW problem, taking $\mathbf{p}$ and $\mathbf{q}$ as uniform distributions.

While this procedure alone is often sufficient for simple problems, in high dimensions, the statistical efficiency of classical divergence measures can be poor and a large number of input samples is needed to achieve good discrimination between generated and data distribution (Salimans et al., 2018). To improve discriminability, we learn the intra-space metrics adversarially. An adversary f_ω parametrized by ω maps data and generator samples into feature spaces in which we compute Euclidean intra-space distances:

$$D^\omega_{ij} := \|f_\omega(x_i) - f_\omega(x_j)\|_2, \text{ where } f_\omega \colon \mathcal{X} \to \mathbb{R}^s \quad (2)$$

with f_ω modeled by a neural network. The feature mapping may, for instance, reduce the dimensionality of $\mathcal{X}$ and extract important features. The original loss minimization problem of the generator thus becomes a minimax problem

$$\min_\theta \max_{\omega=(\check{\omega},\hat{\omega})} \quad \mathrm{GW}(D^{\check{\omega}}, D^{\hat{\omega}}, \mathbf{p}, \mathbf{q}), \quad (3)$$

where $D^{\check{\omega}}$ and $D^{\hat{\omega}}$ denote pairwise distance matrices of samples originating from the generator and reference domain, respectively, mapped into the feature space via f_ω (Eq. (2)). We refer to our model as GW GAN.

3. Training

We optimize the adversary f_ω and generator g_θ in an alternating scheme, where we train the generator more frequently than the adversary to avoid the adversarially-learned distance function to become degenerate (Salimans et al., 2018). Algorithm 1 shows the GW GAN training algorithm.

While training of standard GANs suffers from undamped oscillations and mode collapse (Metz et al., 2017; Salimans et al., 2016), following an argument of Salimans et al. (2018), the GW objective is well defined and statistically consistent if the adversarially learned intra-space distances D^ω are non-degenerate. Trained with the GW loss, the generator thus does not diverge even when adversary f_ω is kept fixed. Empirical validation (see Appendix A) confirms this: we stopped updating the adversary f_ω while continuing to update the generator. Even with fixed adversary, the generator further improved its learned distribution and did not diverge.

Note that Problem (3) makes very few assumptions on the spaces $\mathcal{X}$ and $\mathcal{Y}$, requiring only that a metric be defined on them. This remarkable flexibility can be exploited to enforce various characteristics on the generated distribution. We discuss examples in Section 3.1. However, this same flexibility combined with the added degrees of freedom due to the learned metric, demands to regularize the adversary to ensure stable training and prevent it from overpowering the generator. We propose an effective method to do so in Section 3.2.

Moreover, using the Gromov-Wasserstein distance as a differentiable loss function for training a generative model requires modifying its original formulation to ensure robust and fast computation, unbiased gradients, and numerical stability, as described in detail in Section 3.3.

3.1. Constraining the Generator

The GW loss encourages the generator to recover the relational and geometric properties of the reference dataset, but leaves other global aspects undetermined. We can thus

shape the generated distribution by enforcing desired properties through constraints. For example, while any translation of a distribution would achieve the same GW loss, we can enforce centering around the origin by penalizing the norm of the generated samples. Figure 2a illustrates an example.

For computer vision tasks, we need to ensure that the generated samples still look like natural images. We found that a total variation regularization (Rudin et al., 1992) induces the right bias here and hence greatly improves the results (see Figures 2b, c, and d).

Moreover, the invariances of the GW loss allow for shaping stylistic characteristics of the generated samples by integrating design constraints into the learning process. In contrast, current generative models (Arjovsky et al., 2017; Salimans et al., 2018; Genevay et al., 2018; Li et al., 2017) cannot perform style transfer as modifications in surface-level features of the generated samples conflict with their adversarially computed loss used for training the generator. We propose a modular framework, which enables style transfer to the generated samples when given an additional style reference besides the provided data samples. We incorporate design constraints into the generator's objective via a *style adversary* c, i.e., any function that quantifies a certain style and thereby, as a penalty, enforces this style on the generated samples. The resulting objective is

$$\min_{\theta} \max_{\omega=(\tilde{\omega},\hat{\omega})} \quad \mathrm{GW}(D^{\tilde{\omega}}, D^{\hat{\omega}}, \mathbf{p}, \mathbf{q}) - \lambda \times c(g_\theta(z)). \quad (4)$$

As a result, the generator learns structural content of the target distribution via the adversarially learned GW loss, and stylistic characteristics via the style adversary. We demonstrate this framework via the example of stylistic changes to learned MNIST digits in Section 4.4 and in Appendix G.

3.2. Regularizing the Adversary

During training, the adversary maximizes the objective function (3). However, the GW distance is easily maximized by stretching the space and thus distorting the intra-space distances used for its computation. To avoid such arbitrary distortion of the space, we propose to regularize the adversary f_ω by (approximately) enforcing it to define a unitary transformation, thus restricting the magnitude of stretching it can do. Note that *directly* parametrizing f_ω as an orthogonal matrix would defeat its purpose, as the Frobenius norm is unitarily invariant. Instead, we allow f_ω to take a more general form, but limit its expansivity and contractivity through approximate orthogonality.

Previous work has explored various orthogonality-based regularization methods to stabilize neural networks training(Vorontsov et al., 2017). Saxe et al. (2014) introduced a new class of random orthogonal initial conditions on the weights of neural networks stabilizing the initial training phase. By enforcing the weight matrices to be Parseval tight frames, layerwise orthogonality constraints are introduced in Cisse et al. (2017); Brock et al. (2017; 2019); they penalize deviations of the weights from orthogonality via $R_\beta(W_k) := \beta\|W_k^\top W_k - I\|_F^2$, where W_k are weights of layer k and $\|\cdot\|_F$ is the Frobenius norm.

However, these approaches enforce orthogonality on the weights of each layer rather than constraining the network f_ω in its entirety to function as an approximately orthogonal operator. An empirical comparison to these layerwise approaches (shown in Appendix D) reveals that, for GW GAN, regularizing the full network is desirable. To enforce the approximation of f_ω as an orthogonal operator, we introduce a new orthogonal regularization approach, which ensures orthogonality of a network by minimizing the distance to the closest orthogonal matrix P^*. The regularization term is defined as

$$R_\beta(f_\omega(X), X) := \beta\|f_\omega(X) - XP^{*\top}\|_F^2, \quad (5)$$

where P^* is an orthogonal matrix that most closely maps X to $f_\omega(X)$, and β is a hyperparameter. The matrix $P^* = \arg\min_{P\in O(s)} \|f_\omega(X) - XP^\top\|_F$, where $O(s) = \left\{P \in \mathbb{R}^{s\times s} \mid P^\top P = I\right\}$ and s is the dimensionality of the feature space, can be obtained by solving an orthogonal Procrustes problem. If the dimensionality s of the feature space equals the input dimension, then P^* has a closed-form solution $P^* = UV^\top$, where U and V are the left and right singular vectors of $f_\omega(X)^\top X$, i.e. $U\Sigma V^\top = \mathrm{SVD}(f_\omega(X)^\top X)$ (Schönemann, 1966). Otherwise, we need to solve for P^* with an iterative optimization method.

This novel Procrustes-based regularization principle for neural networks is remarkably flexible since it constrains global input-output behavior without making assumptions about specific layers or activations. It preserves the expressibility of the network while efficiently enforcing orthogonality. We use this orthogonal regularization principle for training the adversary of the GW generative model across different applications and network architectures.

3.3. Gromov-Wasserstein as a Loss Function

To serve as a robust training objective for general machine learning settings we modify the naïve formulation of the Gromov-Wasserstein discrepancy in various ways.

Regularization of Gromov-Wasserstein Optimal transport metrics and extensions such as the Gromov-Wasserstein distance are particularly appealing because they take into account the underlying geometry of the data when comparing distributions. However, their computational cost is prohibitive for large-scale machine learning problems. More precisely, Problem (1) is a quadratic programming problem,

and solving it directly is intractable for large n. Regularizing this objective with an entropy term results in significantly more efficient optimization (Peyré et al., 2016). The resulting smoothed problem can be solved through projected gradient descent methods, where the projection steps rely on the Sinkhorn-Knopp scaling algorithm (Cuturi, 2013). Concretely, the entropy-regularized version of the Gromov-Wasserstein discrepancy proposed by Peyré et al. (2016) has the form

$$\mathrm{GW}_\epsilon(D, \bar{D}, \mathbf{p}, \mathbf{q}) = \min_{T \in \mathcal{U}_{\mathbf{p},\mathbf{q}}} \mathcal{E}_{D,\bar{D}}(T) - \epsilon H(T), \qquad (6)$$

where $\mathcal{E}_{D,\bar{D}}(T)$ is defined in Equation (1), $H(T) := -\sum_{ij} T_{ij}(\log(T_{ij}) - 1)$ is the entropy of coupling T, and ϵ a parameter controlling the strength of regularization. Besides leading to significant speedups, entropy smoothing of optimal transport discrepancies results in distances that are *differentiable* with respect to their inputs, making them a more convenient choice as loss functions for machine learning algorithms. Since the Gromov-Wasserstein distance as loss function in generative models compares noisy features of the generator and the data by computing correspondences between intra-space distances, a soft rather than a hard alignment T might be a desirable property. The entropy smoothing yields couplings that are sparser than their non-regularized counterparts, ideal for applications where soft alignments are desired (Cuturi & Peyré, 2016).

The effectiveness of entropy smoothing of the Gromov-Wasserstein discrepancy has been shown in other downstream application such as shape correspondences (Solomon et al., 2016) or the alignment of word embedding spaces (Alvarez-Melis & Jaakkola, 2018).

Motivated by Salimans et al. (2018) and justified by the envelope theorem (Carter, 2001), we do not backpropagate the gradient through the iterative computation of the GW_ϵ coupling T (Problem (6)).

Normalization of Gromov-Wasserstein With entropy regularization, GW_ϵ is not a distance any more, as the discrepancy of identical metric measure spaces is then no longer zero. Similar to the Wasserstein metric (Bellemare et al., 2017), the estimation of GW_ϵ from samples yields biased gradients. Inspired by Bellemare et al. (2017), we use a normalized entropy-regularized Gromov-Wasserstein discrepancy defined as

$$\begin{aligned}\overline{\mathrm{GW}}_\epsilon(D, \bar{D}, \mathbf{p}, \mathbf{q}) := \; & 2 \times \mathrm{GW}_\epsilon(D, \bar{D}, \mathbf{p}, \mathbf{q}) \\ & - \mathrm{GW}_\epsilon(D, D, \mathbf{p}, \mathbf{p}) - \mathrm{GW}_\epsilon(\bar{D}, \bar{D}, \mathbf{q}, \mathbf{q}).\end{aligned} \qquad (7)$$

Numerical Stability of Gromov-Wasserstein Computing the entropy-regularized Gromov-Wasserstein formulation relies on a projected gradient algorithm (Peyré et al., 2016), in which each iteration involves a projection into the transportation polytope, efficiently computed with the Sinkhorn-Knopp algorithm (Cuturi, 2013), a matrix-scaling procedure that alternatingly updates marginal scaling variables. In the limit of vanishing regularization ($\epsilon \to 0$) these scaling factors diverge, resulting in numerical instabilities.

To improve the numerical stability, we compute GW_ϵ using a stabilized version of the Sinkhorn algorithm (Schmitzer, 2016). This significantly increases the robustness of the Gromov-Wasserstein computation. Performing Sinkhorn updates in the log-domain further increases the stability of the algorithm, by avoiding numerical overflow while preserving its efficient matrix multiplication structure.

Normalizing the intra-space distances of the generated and the data samples, respectively, further improves the numerical stability of the Gromov-Wasserstein computation. However, to preserve information on the scale of the samples, we use normalized distances for the Sinkhorn iterates, while the final loss is calculated using the original distances.

4. Empirical Results

In this section, we empirically demonstrate the effectiveness of the GW GAN formulation and regularization, and illustrate its versatility by tackling various novel settings for generative modeling, including learning distributions across different dimensionalities, data types and styles.

4.1. Learning across Identical Spaces

As a sanity check, we first consider the special case where the two distributions are defined on identical spaces (i.e., the usual GAN setting). Specifically, we test the model's ability to recover 2D mixtures of Gaussians, a common proof of concept task for mode recovery (Che et al., 2017; Metz et al., 2017; Li et al., 2018). For the experiments on synthetic datasets, generator and adversary architectures are multilayer perceptrons (MLPs) with ReLU activation functions. Figure 2a shows that the GW GAN reproduces a mixture of Gaussians with learned adversary f_ω that stabilizes the learning. We observe that ℓ_1-regularization indeed helps position the learned distributions around the origin. Comparative results with and without ℓ_1-regularization are shown in Appendix B. As opposed to the OT GAN proposed by Salimans et al. (2016), our model robustly learns Gaussian mixtures with differing number of modes and arrangements (see Appendix E). The Appendix shows several training runs. While the generated distributions vary in orientation in the Euclidean plane, the cluster structure is clearly preserved.

To illustrate the ability of the GW GAN to generate images, we train the model on MNIST (LeCun et al., 1998), fashion-MNIST (Xiao et al., 2017) and gray-scale CIFAR-

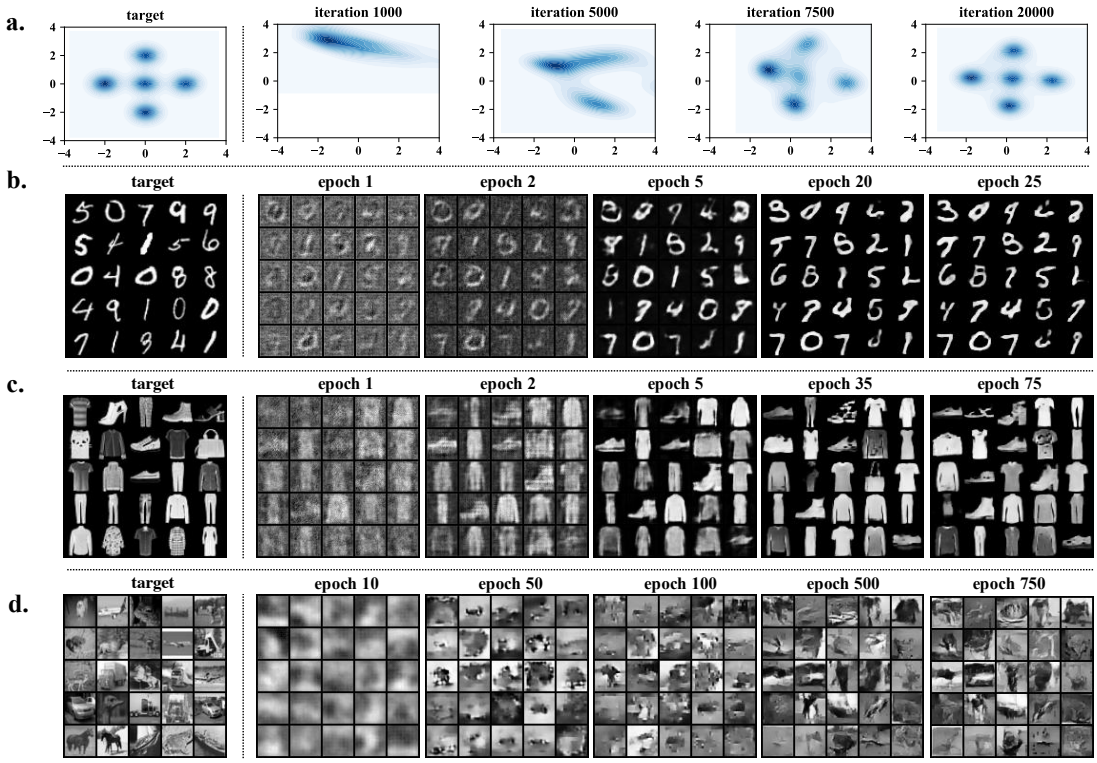

Figure 2. The Gromov-Wasserstein GAN can learn distributions of different dimensionality. **a.** Results of learning a mixture of Gaussian distributions with adversary f_ω ($\beta = 1$). ℓ_1-regularization allows centering the distribution across the origin (ℓ_1-penalty: $\lambda = 0.001$). Each plot shows 1000 generated samples. Learning to generate **b.** MNIST digits ($\beta = 32$), **c.** fashion-MNIST ($\beta = 35$) and **d.** gray-scale CIFAR10 ($\beta = 40$). Trained with total variation denoising ($\lambda = 0.5$).

10 (Krizhevsky et al., 2014). Both generator and adversary follow the deep convolutional architecture introduced by Chen et al. (2016), whereby the adversary f_ω maps into $\mathbb{R}^s$ rather than applying a final `tanh`. To stabilize the initial training phase, the weights of the adversary network were initialized with random orthogonal matrices as proposed by Saxe et al. (2014). We train the model using Adam with a learning rate of 2×10^{-4}, $\beta_1 = 0.5$, $\beta_2 = 0.99$ (Kingma & Ba, 2015). Figure 2b, c and d display generated images throughout the training process. The adversary was constrained to approximate an orthogonal operator. The results highlight the effectiveness of the orthogonal Procrustes regularization, which allows successful learning of complex distributions using different network architectures. Additional experiments on the influence of adversary f_ω are provided in Appendix C.

Having validated the overall soundness of the GW GAN on traditional settings, we now demonstrate its usefulness in tasks that go beyond the scope of traditional generative adversarial models, namely, learning across spaces that are not directly comparable.

4.2. Learning across Dimensionalities

Arguably, the simplest instance of *incomparable* spaces are Euclidean spaces of different dimensionality. In this section, we investigate whether the Gromov-Wasserstein GAN can learn to generate a distribution defined on a space of different dimensionality than that of the reference. We consider both directions: learning to a smaller and higher dimensional space. In this experimental setup, we compute intra-space distances using the Euclidean distance without a parametrized adversary. The generator network follows an MLP architecture with ReLU activation functions. The training task consists of translating between a mixture of Gaussian distributions in two and three dimensions. The results, shown in Figure 3, demonstrate that our model successfully recovers the global structure and relative distances of the modes of the reference distribution, despite the different dimensionality.

4.3. Learning across Data Modalities and Manifolds

Next, we consider distributions with more complex structure, and test whether our model is able to *recover* manifold structure on the generated distribution. Using the popular

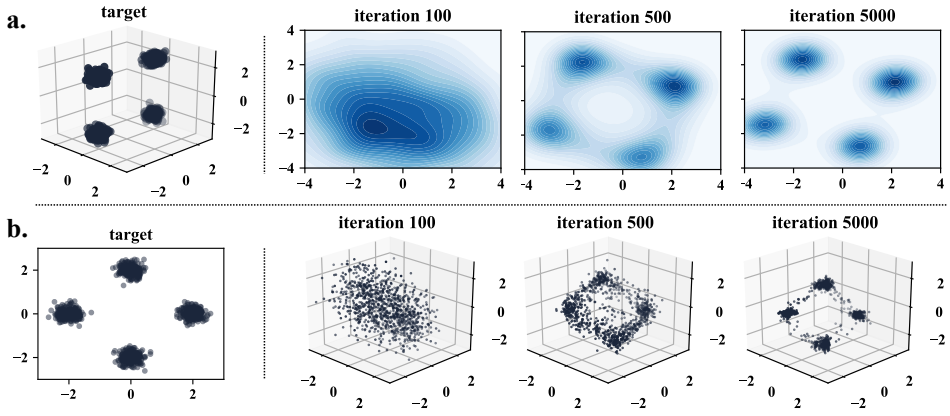

Figure 3. The GW GAN can be applied to generate samples of **a.** reduced and **b.** increased dimensionality compared to the target distribution. All plots show 1000 samples.

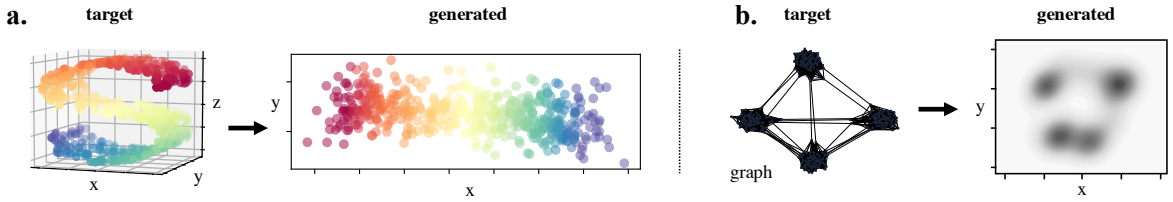

Figure 4. By learning from intra-space distances, the GW GAN learns the manifold structure of the data. **a.** The model can be applied to dimensionality reduction tasks and reproduce a three-dimensional S-curve in two-dimensions. Intra-space distances of the data samples are Floyd-Warshall shortest paths of the corresponding k-nearest neighbor graph. **b**. Similarly, it can map a graph into $\mathbb{R}^2$. The plots display 500 samples.

three-dimensional S-shaped dataset as example, we define distances between the samples via shortest paths on their k-nearest neighbor graph, computed using the Floyd-Warshall algorithm (Floyd, 1962). For the generated distribution we use a space of the same intrinsic dimensionality (two) as the reference manifold. The results in Figure 4a show that the generated distribution learnt with GW GAN successfully recovers the manifold structure of the data.

Taking the notion of incomparability further, we next consider a setting when the reference distribution is accessible only through relational information, i.e., a weighted graph without absolute representations of the samples. While conceptually very different from previous scenarios, applying our model to this setting is just as simple as previous scenarios. Once a notion of distance is defined over the reference graph, our model learns the distribution based on pairwise relations as before. Given merely a graph, we use pairwise shortest paths as the intra-space distance metric, and use the 2D Euclidean space for the generated distribution. Figure 4b shows that GW GAN is able to successfully learn a distribution that approximately recovers the neighborhood structure of the reference graph.

4.4. Shaping Learned Distributions

The Gromov-Wasserstein GAN enjoys remarkable flexibility, allowing us to actively influence stylistic characteristics of the generated distribution.

While structure and content of the distribution are learned via the adversary f_ω, stylistic features can be introduced via a style adversary as outlined in Section 3.1. As a proof of concept of this modular framework, we learn MNIST digits and enforce their font style to be bold via additional design constraints. The style adversary is parametrized by a binary classifier trained on handwritten letters of the EMNIST dataset (Cohen et al., 2017) which were assigned thin and bold class labels $l \in \{0, 1\}$. The training objective of the generator g_θ is augmented with the classification result of the trained binary classifier (Eq. (4)). Further details are provided in Appendix G. After the generator has satisfactorily learnt the data distribution based on training with loss $\overline{\text{GW}}_\epsilon$, the style adversary c is activated. Figure 5 shows that the style adversary affects the generator to increase the thickness of the MNIST digits, while the structural content learned in the first stage is retained.

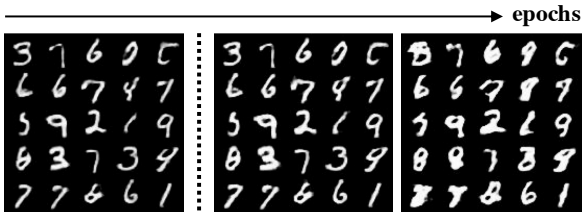

Figure 5. **Cross-Style Generation**. By decoupling topological information from superficial characteristics, our method allows for stylistic aspects to be enforced (boldness in this case, enforced with a style adversary) upon the generated distribution *while* preserving the principal characteristics of the reference distribution (MNIST).

5. Related Work

Generative adversarial models have been extensively studied and applied in various fields including image synthesis (Brock et al., 2019), semantic image editing (Wang et al., 2018), style transfer (Zhu et al., 2017), and semi-supervised learning (Kingma et al., 2014). As the literature is extensive, we provide a brief overview on GANs and focus on selected approaches targeting tasks in cross-domain learning.

Generative Adversarial Networks Goodfellow et al. (2014) proposed generative adversarial networks (GANs) as a zero-sum game between a generator and a discriminator, which learns to distinguish between generated and data samples. Despite their success and improvements in optimization, the training of GANs is difficult and unstable (Salimans et al., 2016; Arjovsky & Bottou, 2017). To remedy these issues, various extensions of this framework have been proposed, most of which seek to replace the game objective with more stable or general losses. These include using Maximum Mean Discrepancy (MMD) (Dziugaite et al., 2015; Li et al., 2017; Bińkowski et al., 2018), other IPMs (Mroueh et al., 2017; 2018), or Optimal Transport distances (Arjovsky et al., 2017; Salimans et al., 2018; Genevay et al., 2018). Due to their relevance, we discuss the latter in detail below. A crucial characteristic that distinguishes our approach from other generative models is its ability to learn across different domains and modalities.

GANs and Optimal Transport (OT) To compare probability distributions supported on low dimensional manifolds in high dimensional spaces, recent GAN variants integrate OT metrics in their training objective (Arjovsky et al., 2017; Salimans et al., 2018; Genevay et al., 2018; Gulrajani et al., 2017). Since OT metrics are computationally expensive, Arjovsky et al. (2017) use the dual formulation of the 1-Wasserstein distance. Other approaches approximate the primal via entropically smoothed generalizations of the Wasserstein distance (Salimans et al., 2018; Genevay et al., 2018). Our work departs from these methods in that it relies on a much more general instance of Optimal Transport (the Gromov-Wasserstein distance) as a loss function, which allows us to compare distributions even if cross-domain pairwise distances are not available.

GANs for Cross-Domain Learning GANs have been successfully applied to style transfer between images (Isola et al., 2017; Karacan et al., 2016; Zhu et al., 2017), text-to-image synthesis (Reed et al., 2016; Zhang et al., 2017), visual manipulation (Zhu et al., 2016; Engel et al., 2018) or font style transfer (Azadi et al., 2018). However, to achieve this, these methods depend on conditional variables, training sets of aligned data pairs or cycle consistency constraints. Kim et al. (2017) utilize two different, coupled GANs to discover cross-domain relations given unpaired data. However, the method's applicability is limited as all images in one domain need to be representable by images in the other domain.

Gromov-Wasserstein Learning Since its introduction by Mémoli (2011), the Gromov-Wasserstein discrepancy has found applications in many learning problems that rely on a coupling between different metric spaces. Being an effective method to solve matching problems, it has been used in shape and object matching (Mémoli, 2009; 2011; Solomon et al., 2016; Ezuz et al., 2017), for aligning word embedding spaces (Alvarez-Melis & Jaakkola, 2018) and for matching weighted directed networks (Chowdhury & Mémoli, 2018). Other recent applications of the GW distance include the computation of barycenters of a set of distance or kernel matrices (Peyré et al., 2016) and heterogeneous domain adaptation where source and target samples are represented in different feature spaces (Yan et al., 2018). While relying on a shared tool —the GW discrepancy— this paper leverages it in a very different framework, generative modeling, where questions of efficiency, degrees of freedom, minimax objectives and end-to-end learning pose various challenges that need to be addressed to successfully use this tool.

6. Conclusion

In this paper, we presented a new generative model that can learn a distribution in a space that is different from, and even *incomparable* to, that of the reference distribution. Our model accomplishes this by relying on relational —rather than absolute— comparisons of samples via the Gromov-Wasserstein distance. Such disentanglement of data and generator spaces opens up a wide array of novel possibilities for generative modeling, as portrayed by our experiments on learning across different dimensional representations and learning across modalities (weighted graph to Euclidean representations). Validated here through simple experiments on digit thickness control, the use of crafted regularization losses on the generator to impose certain stylistic characteristics makes for an exciting avenue of future work.

Acknowledgements

This research was supported in part by NSF CAREER Award 1553284 and The Defense Advanced Research Projects Agency (grant number YFA17 N66001-17-1-4039). The views, opinions, and/or findings contained in this article are those of the author and should not be interpreted as representing the official views or policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the Department of Defense. Charlotte Bunne was supported by the Zeno Karl Schindler Foundation. We thank Suvrit Sra for a question that initiated this research, and MIT Supercloud and the Lincoln Laboratory Supercomputing Center for providing computational resources.

References

Alvarez-Melis, D. and Jaakkola, T. Gromov-Wasserstein Alignment of Word Embedding Spaces. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2018.

Arjovsky, M. and Bottou, L. Towards Principled Methods for Training Generative Adversarial Networks. In *International Conference on Learning Representations (ICLR)*, 2017.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein Generative Adversarial Networks. In *International Conference on Machine Learning (ICML)*, volume 70. PMLR, 2017.

Azadi, S., Fisher, M., Kim, V. G., Wang, Z., Shechtman, E., and Darrell, T. Multi-Content GAN for Few-Shot Font Style Transfer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Bellemare, M. G., Danihelka, I., Dabney, W., Mohamed, S., Lakshminarayanan, B., Hoyer, S., and Munos, R. The Cramer Distance as a Solution to Biased Wasserstein Gradients. *arXiv preprint arXiv:1705.10743*, 2017.

Bińkowski, M., Sutherland, D. J., Arbel, M., and Gretton, A. Demystifying MMD GANs. In *International Conference on Learning Representations (ICLR)*, 2018.

Brock, A., Lim, T., Ritchie, J. M., and Weston, N. Neural Photo Editing with Introspective Adversarial Networks. In *International Conference on Learning Representations (ICLR)*, 2017.

Brock, A., Donahue, J., and Simonyan, K. Large Scale GAN Training for High Fidelity Natural Image Synthesis. *International Conference on Learning Representations (ICLR)*, 2019.

Carter, M. *Foundations of Mathematical Economics*. MIT Press, 2001.

Che, T., Li, Y., Jacob, A. P., Bengio, Y., and Li, W. Mode Regularized Generative Adversarial Networks. 2017.

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., and Abbeel, P. InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Chowdhury, S. and Mémoli, F. The Gromov-Wasserstein distance between networks and stable network invariants. *arXiv preprint arXiv:1808.04337*, 2018.

Cisse, M., Bojanowski, P., Grave, E., Dauphin, Y., and Usunier, N. Parseval Networks: Improving Robustness to Adversarial Examples. In *International Conference on Machine Learning (ICML)*, volume 70. PMLR, 2017.

Cohen, G., Afshar, S., Tapson, J., and van Schaik, A. EMNIST: an extension of MNIST to handwritten letters. *arXiv preprint arXiv:1702.05373*, 2017.

Cuturi, M. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.

Cuturi, M. and Peyré, G. A Smoothed Dual Approach for Variational Wasserstein Problems. *SIAM Journal on Imaging Sciences*, 9, 2016.

Dziugaite, G. K., Roy, D. M., and Ghahramani, Z. Training generative neural networks via Maximum Mean Discrepancy optimization. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2015.

Engel, J., Hoffman, M., and Roberts, A. Latent Constraints: Learning to Generate Conditionally from Unconditional Generative Models. In *International Conference on Learning Representations (ICLR)*, 2018.

Ezuz, D., Solomon, J., Kim, V. G., and Ben-Chen, M. GWCNN: A Metric Alignment Layer for Deep Shape Analysis. In *Computer Graphics Forum*, volume 36. Wiley Online Library, 2017.

Floyd, R. W. Algorithm 97: Shortest path. *Communications of the ACM*, 5(6):345, 1962.

Genevay, A., Peyré, G., and Cuturi, M. Learning Generative Models with Sinkhorn Divergences. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 84. PLMR, 2018.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved Training of Wasserstein GANs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. Image-to-Image Translation with Conditional Adversarial Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.

Karacan, L., Akata, Z., Erdem, A., and Erdem, E. Learning to Generate Images of Outdoor Scenes from Attributes and Semantic Layouts. *arXiv preprint arXiv:1612.00215*, 2016.

Kim, T., Cha, M., Kim, H., Lee, J. K., and Kim, J. Learning to Discover Cross-Domain Relations with Generative Adversarial Networks. In *International Conference on Machine Learning (ICML)*, volume 70, 2017.

Kingma, D. and Ba, J. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*, volume 5, 2015.

Kingma, D. P., Mohamed, S., Rezende, D. J., and Welling, M. Semi-Supervised Learning with Deep Generative Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

Krizhevsky, A., Nair, V., and Hinton, G. The CIFAR-10 Dataset, 2014.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 1998.

Li, C., Alvarez-Melis, D., Xu, K., Jegelka, S., and Sra, S. Distributional Adversarial Networks. *International Conference on Learning Representations (ICLR), Workshop Track*, 2018.

Li, C.-L., Chang, W.-C., Cheng, Y., Yang, Y., and Póczos, B. MMD GAN: Towards Deeper Understanding of Moment Matching Network. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Mémoli, F. Spectral Gromov-Wasserstein Distances for Shape Matching. In *Computer Vision Workshops (ICCV Workshops)*. IEEE, 2009.

Mémoli, F. Gromov-Wasserstein Distances and the Metric Approach to Object Matching. *Foundations of Computational Mathematics*, 11(4), 2011.

Metz, L., Poole, B., Pfau, D., and Sohl-Dickstein, J. Unrolled generative adversarial networks. *International Conference on Learning Representations (ICLR)*, 2017.

Mroueh, Y., Sercu, T., and Goel, V. McGAN: Mean and Covariance Feature Matching GAN. In *International Conference on Machine Learning (ICML)*, volume 70. PMLR, 2017.

Mroueh, Y., Li, C.-L., Sercu, T., Raj, A., and Cheng, Y. Sobolev GAN. In *International Conference on Learning Representations (ICLR)*, 2018.

Müller, A. Integral Probability Metrics and Their Generating Classes of Functions. *Advances in Applied Probability*, 29(2), 1997.

Peyré, G., Cuturi, M., and Solomon, J. Gromov-Wasserstein Averaging of Kernel and Distance Matrices. In *International Conference on Machine Learning (ICML)*, volume 48, 2016.

Radford, A., Metz, L., and Chintala, S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In *International Conference on Learning Representations (ICLR)*, 2016.

Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., and Lee, H. Generative Adversarial Text to Image Synthesis. In *International Conference on Machine Learning (ICML)*, volume 48, 2016.

Rudin, L. I., Osher, S., and Fatemi, E. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4), 1992.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., and Chen, X. Improved Techniques for Training GANs. In *Advances in Neural Information Processing Systems (NeurIPS)*. 2016.

Salimans, T., Zhang, H., Radford, A., and Metaxas, D. Improving GANs Using Optimal Transport. In *International Conference on Learning Representations (ICLR)*, 2018.

Saxe, A. M., McClelland, J. L., and Ganguli, S. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.

Schmitzer, B. Stabilized Sparse Scaling Algorithms for Entropy Regularized Transport Problems. *arXiv preprint arXiv:1610.06519*, 2016.

Schönemann, P. H. A generalized solution of the Orthogonal Procrustes problem. *Psychometrika*, 31(1), 1966.

Solomon, J., Peyré, G., Kim, V. G., and Sra, S. Entropic Metric Alignment for Correspondence Problems. *ACM Transactions on Graphics (TOG)*, 35(4), 2016.

Sriperumbudur, B., Fukumizu, K., Gretton, A., Schölkopf, B., and Lanckriet, G. On the Empirical Estimation of Integral Probability Metrics. *Electronic Journal of Statistics*, 6, 2012.

Vorontsov, E., Trabelsi, C., Kadoury, S., and Pal, C. On orthogonality and learning recurrent networks with long term dependencies. In *International Conference on Machine Learning (ICML)*, volume 70. PMLR, 2017.

Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., and Catanzaro, B. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Xiao, H., Rasul, K., and Roland, V. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Yan, Y., Li, W., Wu, H., Min, H., Tan, M., and Wu, Q. Semi-Supervised Optimal Transport for Heterogeneous Domain Adaptation. *International Joint Conference on Artificial Intelligence (IJCAI)*, 2018.

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., and Metaxas, D. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *International Conference on Computer Vision (ICCV)*, 2017.

Zhu, J.-Y., Krähenbühl, P., Shechtman, E., and Efros, A. A. Generative Visual Manipulation on the Natural Image Manifold. In *European Conference on Computer Vision (ECCV)*, 2016.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *International Conference on Computer Vision (ICCV)*, 2017.

Appendix

A. Training of the GW GAN with Fixed Adversary

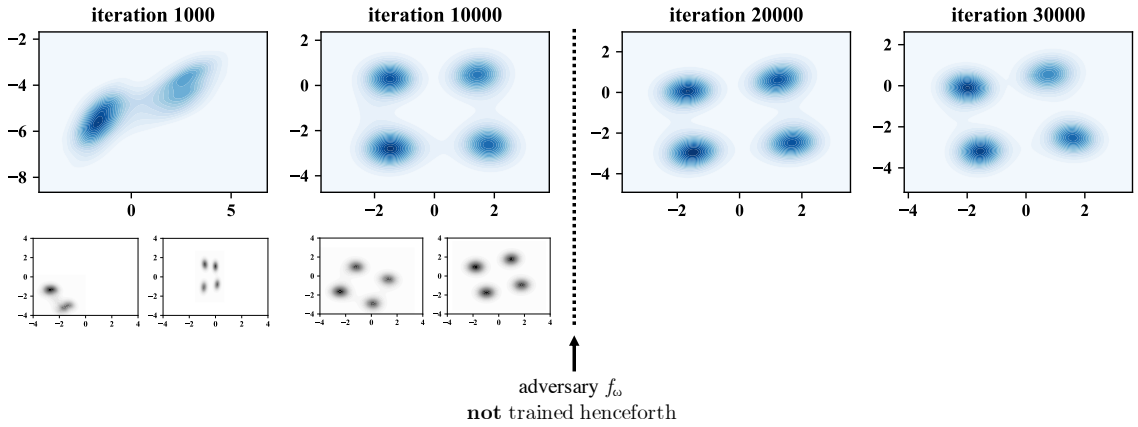

Figure 6. Results demonstrating the consistency of the GW training objective when fixing adversary f_ω. Even after holding the adversary f_ω fixed and stopping its training after 10000 iterations, the generator does not diverge and remains consistent. All plots display 1000 samples.

B. Influence of Generator Constraints on the GW GAN

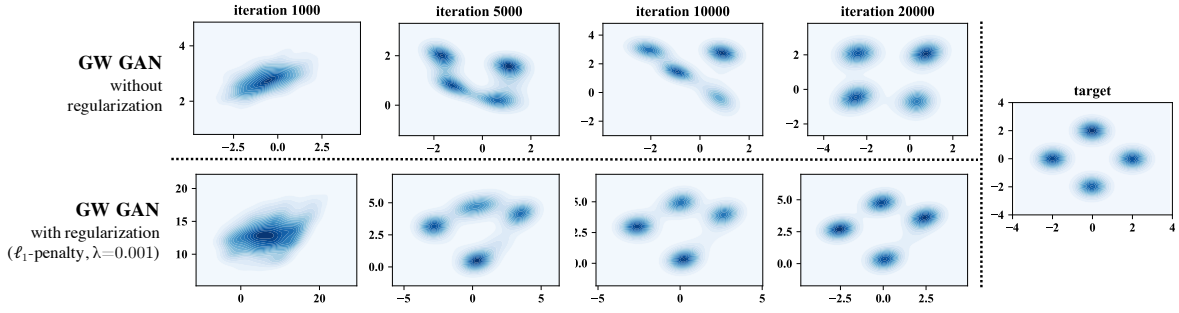

Figure 7. The GW GAN recovers relational and geometric properties of the reference distribution. Global aspects can be determined via constraints of the generator g_θ. When learning a mixture of four Gaussians we can enforce centering around the origin by penalizing the ℓ_1-norm of the generated samples. The results show training the GW GAN with and without a ℓ_1-penalty. While the GW GAN recovers all modes in both cases, learning with ℓ_1-regularization centers the resulting distribution around the origin and determines its orientation in the Euclidean plane. All plots display 1000 samples.

C. Influence of the Adversary

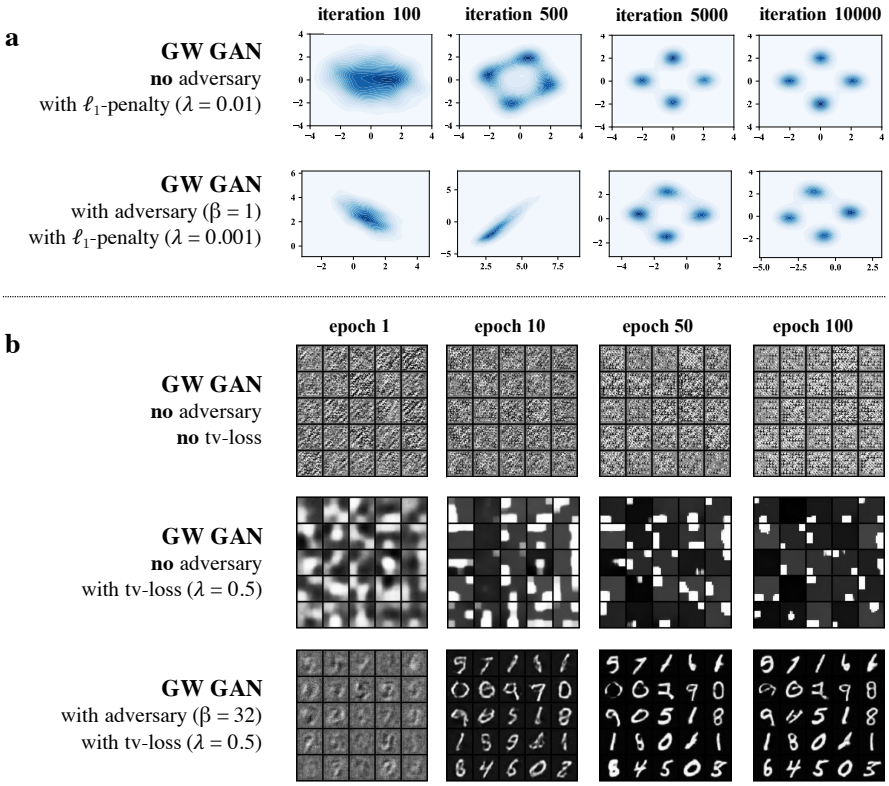

Figure 8. Learning the intra-space metrics adversarially is crucial for high dimensional applications. **a.** For simple applications such as generating 2D Gaussian distributions, the GW GAN performs well irrespective of the use of an adversary. **b.** However, for higher dimensional inputs, such as generating MNIST digits, the use of the adversary is crucial. Without the adversary, the GW GAN is not able to recover the underlying target distributions, while the total variation regularization (tv-loss) effects a clustering of local intensities.

D. Comparison of the Effectiveness of Orthogonal Regularization Approaches

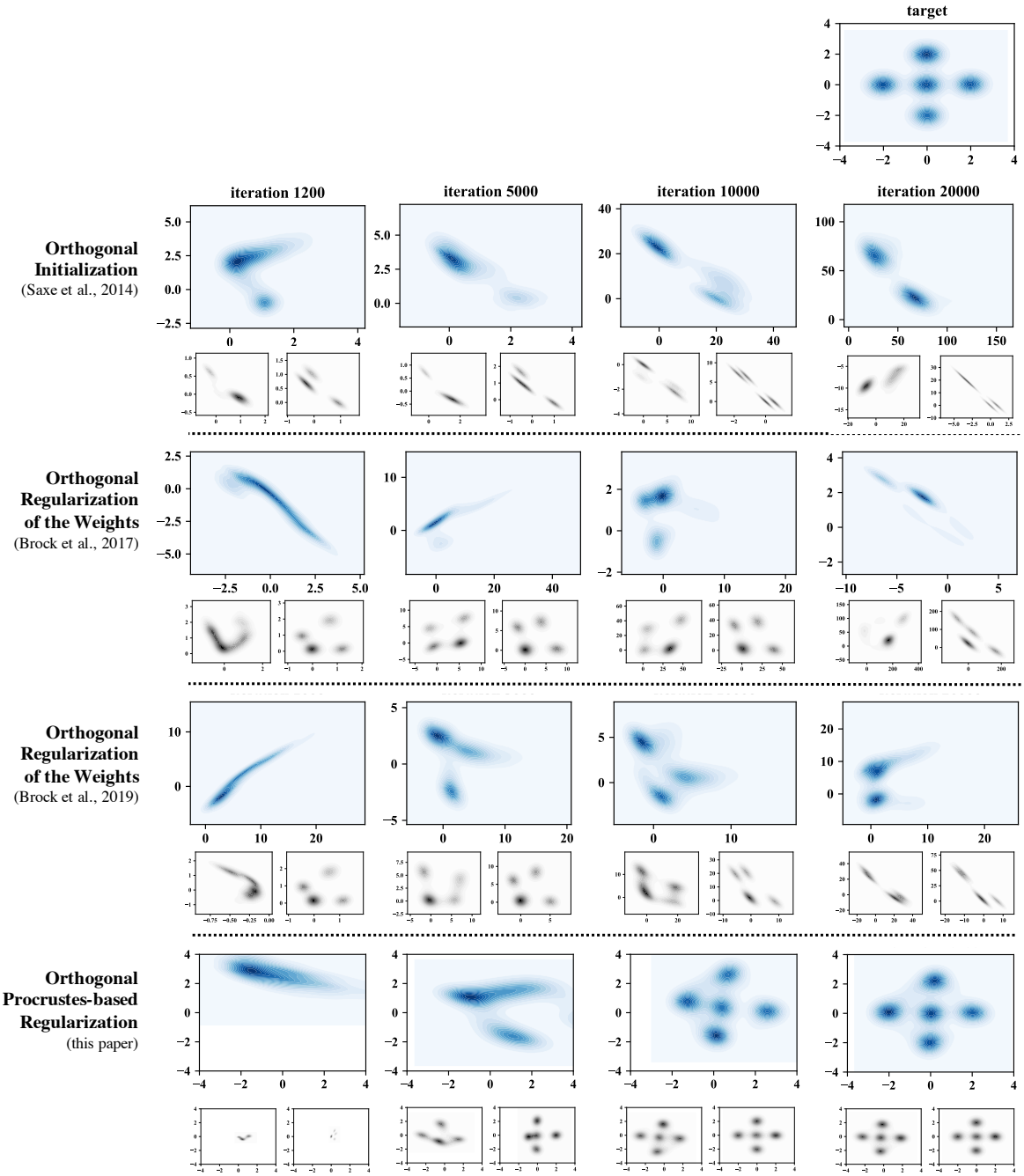

Figure 9. To avoid arbitrary distortion of the space by the adversary, we propose to regularize f_ω by (approximately) enforcing it to define a unitary transformation. We compare different methods of orthogonal regularization of neural networks; including random initialization of the network's weights at the beginning of the training (Saxe et al., 2014), and layerwise orthogonality constraints which penalize deviations of the weights from orthogonality ($R_\beta(W_k) := \beta\|W_k^\top W_k - I\|_{\mathrm{F}}^2$) (Brock et al., 2017). Brock et al. (2019) remove the diagonal terms from the regularization, which enforces the weights to be orthogonal but does not constrain their norms ($R_\beta(W_k) := \beta\|W_k^\top W_k \odot (1 - I)\|_{\mathrm{F}}^2$). The approaches of Saxe et al. (2014); Brock et al. (2017; 2019) are not able to tightly constrain f_ω. As a result, the adversary is able to stretch the space and thus maximize the Gromov-Wasserstein distance. Only the orthogonal Procrustes-based orthogonality approach introduced in this paper is able to effectively regularize adversary f_ω preventing arbitrary distortions of the intra-space distances in the feature space. All plots display 1000 samples.

E. Comparison of the GW GAN with Salimans et al. (2018)

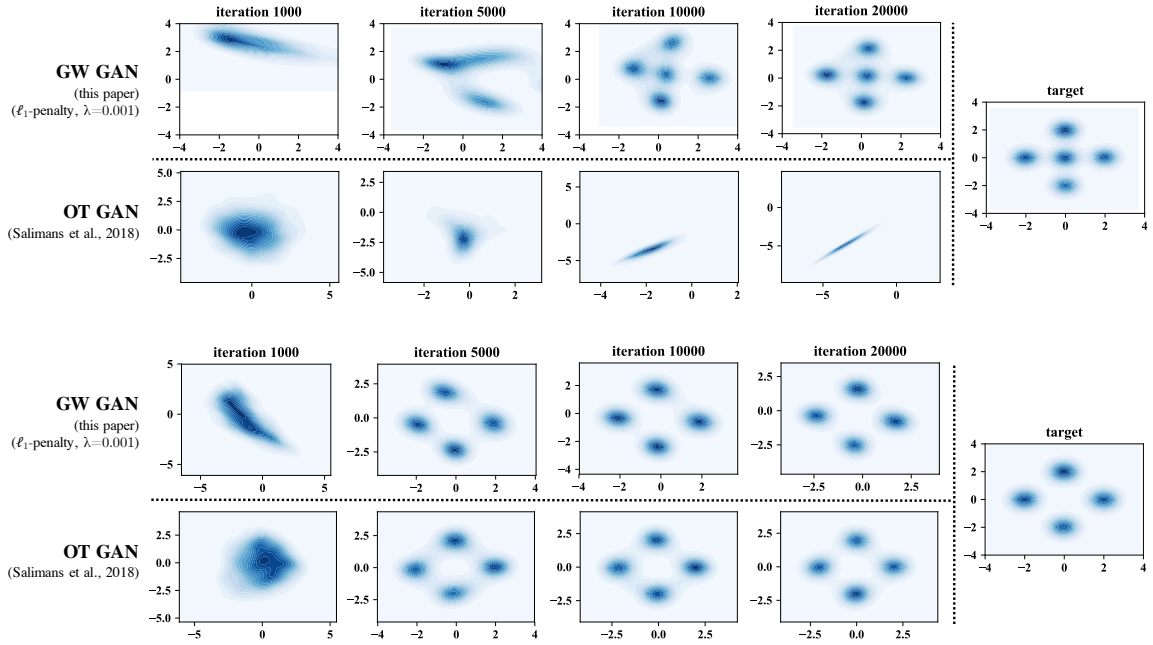

Figure 10. Comparison of the performance of the GW GAN and an OT based generative model proposed by Salimans et al. (2018) (OT GAN). The GW GAN learns mixture of Gaussians with differing number of modes and arrangements. The approach of Salimans et al. (2018) is not able to recover a mixture of five Gaussians with a centering distribution. In the case of a mixture of four Gaussian distributions, both models are able to recover the reference distribution in a similar number of iterations. All plots display 1000 samples.

F. Comparison of Training Times

Model	**Average Training Time** (Seconds per Epoch)
WASSERSTEIN GAN with Gradient Penalty (Gulrajani et al., 2017)	17.57 ± 2.07
SINKHORN GAN (Genevay et al., 2018) (default configuration, $\epsilon = 0.1$)	145.52 ± 1.90
SINKHORN GAN (Genevay et al., 2018) (default configuration, $\epsilon = 0.005$)	153.86 ± 1.64
GW GAN (this paper, $\epsilon = 0.005$)	156.62 ± 1.06

Table 1. Training time comparisons of PyTorch implementations of different GAN architectures. The generative models were trained on generating MNIST digits and their average training time per epoch was recorded. All experiments were performed on a single GPU for consistency.

G. Training Details of the Style Adversary

We introduce a novel framework which allows a modular application of style transfer tasks by integrating a *style adversary* into the architecture of the Gromov-Wasserstein GAN. In order to demonstrate the practicability of this modular framework, we learn MNIST digits and enforce their font style to be bold via additional design constraints. The style adversary is parametrized by a binary classifier trained on handwritten letters of the EMNIST dataset (Cohen et al., 2017) which were assigned bold and thin class labels based on the letterwise ℓ_1-norm of each image. As the style adversary is trained based on a different dataset, it is independent of the original learning task. The binary classifier is parametrized by a convolutional neural network and trained by computing a binary cross-entropy loss. The dataset, classification results of bold and thin letters as well as the loss curve of training the binary classifier are shown in figure 11.

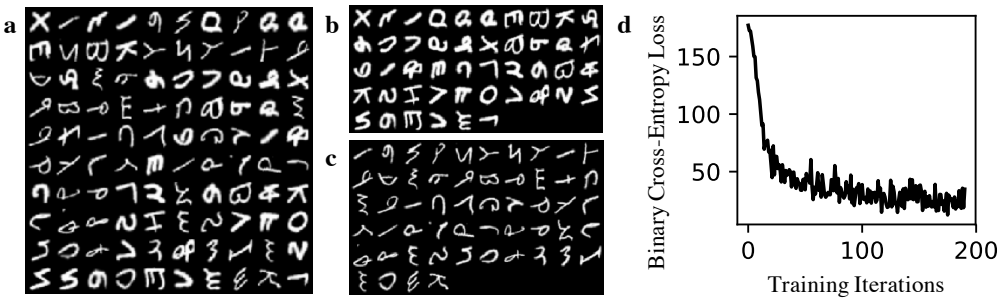

Figure 11. Training of a binary classifier to discriminate between bold and thin letters. **a.** Training set of the EMNIST dataset including bold and thin letters. Output of the trained network of letters labelled as **b.** bold and **c.** thin. **d.** Loss curve corresponding to the training of the binary classifier.