Wasserstein of Wasserstein Loss for Learning Generative Models

Yonatan Dukler[*1] Wuchen Li[*1] Alex Tong Lin[*1] Guido Montúfar[*123]

Abstract

The Wasserstein distance serves as a loss function for unsupervised learning which depends on the choice of a ground metric on sample space. We propose to use a Wasserstein distance as the ground metric on the sample space of images. This ground metric is known as an effective distance for image retrieval, since it correlates with human perception. We derive the Wasserstein ground metric on image space and define a Riemannian Wasserstein gradient penalty to be used in the Wasserstein Generative Adversarial Network (WGAN) framework. The new gradient penalty is computed efficiently via convolutions on the L^2 (Euclidean) gradients with negligible additional computational cost. The new formulation is more robust to the natural variability of images and provides for a more continuous discriminator in sample space.

1. Introduction

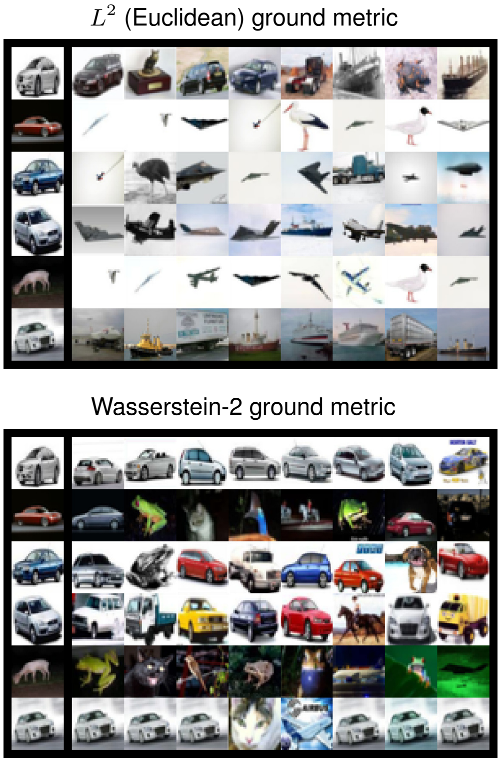
In recent years, optimal transport has become increasingly important in the formulation of training objectives for machine learning applications (Frogner et al., 2015; Montavon et al., 2016; Arjovsky et al., 2017). In contrast to traditional information divergences (arising in maximum likelihood estimation), the Wasserstein distance between probability distributions incorporates the distance between samples via a ground metric of choice. In this way, it provides a continuous loss function for learning probability models supported on possibly disjoint, lower dimensional subsets of the sample space. These properties are especially useful for training implicit generative models, with a prominent example being Generative Adversarial Networks (GANs). The application of the Wasserstein metric to define the objective function of GANs is known as Wasserstein GANs (WGANs) (Frogner et al., 2015; Arjovsky et al., 2017; Deshpande et al., 2018).

When training WGANs, one problem that remains is that of choosing a suitable ground metric for the sample space. The choice of the ground metric plays a crucial role in the training quality of WGANs. Usually the distance between two sample images is taken to be the mean square difference over the features, i.e., the L^2 (Euclidean) norm. This, however, does not incorporate additional knowledge that we have about the space of natural images. In order to improve training and direct focus to selected features, other Sobolev norms in image space have been studied (Adler & Lunz, 2018). Recent works are also investigating distances based on higher level representations of the samples, which can be obtained by means of techniques such as vector embeddings (Mroueh et al., 2017), auto-encoders, or other unsupervised and semi-supervised feature learning techniques (Nowak et al., 2006). Meanwhile, another distance that has been very successful in comparing images, has remained unnoticed in the context of WGANs, namely the Wasserstein distance on images (also named Earth Mover's distance or Monge-Kantorvich distance). In particular, the Wasserstein distance has been successful in image retrieval problems (Rubner et al., 2000; Zhang et al., 2007). It is known to correlate well with human perception for natural images, e.g., being robust to translations and rotations (Engquist & Yang, 2018; Puthawala et al., 2018). See Figure 1. In addition, this distance is very natural and does not require computing higher level representations of the images or any feature selection.

In this paper, we propose to apply the Wasserstein distance over the sample space of images with a ground metric over the discrete space of pixels for learning generative models. We call this ground metric the *Wasserstein ground metric*, and call the Wasserstein loss over the Wasserstein ground metric the Wasserstein of Wasserstein loss. At first sight, it may appear overly complicated to define a loss function of this form. Since computing the Wasserstein distance is already quite involved, a Wasserstein loss based on another Wasserstein ground metric may seem infeasible. Nonetheless, we will show that it is possible to derive an equivalent expression in the settings of gradient penalty of WGANs (Petzka et al., 2017). In details, the Wasserstein-2 ground

[*]Equal contribution [1]Department of Mathematics and [2]Department of Statistics, University of California, Los Angeles, CA 90095. [3]Max Planck Institute for Mathematics in the Sciences, 04103 Leipzig, Germany. Correspondence to: Guido Montúfar <montufar@math.ucla.edu>, Alex Tong Lin <atlin@math.ucla.edu>, Wuchen Li <wcli@math.ucla.edu>, Yonatan Dukler <ydukler@math.ucla.edu>.

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


L^2 (Euclidean) ground metric

Wasserstein-2 ground metric

Figure 1. Source image and 9 nearest neighbors from the CIFAR-10 dataset, with respect to the L^2 (top) and Wasserstein-2 (bottom) ground metrics. We note that the Wasserstein-2 distance is robust to translations and rotations, and gives neighbors that are perceptually similar. In contrast, the Euclidean distance is highly sensitive and oftentimes the nearest neighbors are predominantly white images.

metric exhibits a metric tensor structure (Otto, 2001; Villani, 2009). This introduces a Lipschitz condition based on the Wasserstein norm, rather than the L^2 norm of the standard WGAN setting.

In this work we focus on generative models for images and specifically the WGAN formulation, but the proposed Wasserstein of Wasserstein loss function can be applied to learning with other types of models or other types of data for which a natural distance between features can be introduced.

This paper is organized as follows. In Section 2, we introduce the Wasserstein loss function with Wasserstein ground metric. Based on duality and the metric tensor of the proposed problem, we derive an equivalent practical formulation. In Section 3 we discuss our application to Wasserstein of Wasserstein GANs (WWGANs). Numerical experiments illustrating the benefits of the new gradient norm penalty are provided in Section 4. Related works are reviewed in Section 5.

2. Wasserstein of Wasserstein Loss

In this section, we introduce the Wasserstein ground metric for the Wasserstein loss function. A motivating example is presented to demonstrate the utility of the proposed model.

2.1. Wasserstein loss

Consider a metric sample space $(\mathcal{X}, d_{\mathcal{X}})$. The Wasserstein-$p$ distance is defined as follows. Given a pair $\mathbb{P}_0, \mathbb{P}_1 \in \mathcal{P}_p(\mathcal{X})$ of probability densities with finite p-th moment, let

$$W_{p,d_{\mathcal{X}}}(\mathbb{P}_0, \mathbb{P}_1) = \inf_{\Pi} \left\{ \left(\mathbb{E}_{(X,Y)\sim\Pi} d_{\mathcal{X}}(X,Y)^p \right)^{\frac{1}{p}} \right\}, \quad (1)$$

where Π is a joint distribution of (X, Y) with marginals $X \sim \mathbb{P}_0, Y \sim \mathbb{P}_1$. We note that W_p depends on the choice of a distance function $d_{\mathcal{X}} \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ on sample space, which is usually called the ground metric.

In practice, the sample space $\mathcal{X}$ is typically very high dimensional, sometimes even being an (infinite dimensional) Banach space. We focus on the case where $\mathcal{X}$ is the space of images, which can be regarded as a density space over pixels, i.e., $\mathcal{X} = \mathcal{P}(\Omega)$, where $\Omega = [0, M] \times [0, M]$ is a discrete grid of pixels. With this in mind, we will define the distance function between pixels $d_\Omega \colon \Omega \times \Omega \to \mathbb{R}_+$.

2.2. Wasserstein loss function with Wasserstein ground metric

We now introduce the Wasserstein of Wasserstein loss. Here, the first 'Wasserstein' refers to the Wasserstein loss function over probability distributions on the space of images. The second 'Wasserstein' refers to the ground metric of this loss function. It is chosen as the Wasserstein distance over the space of images defined as histograms over pixels, having a ground metric over pixel locations.

That is, a raster image can be viewed as a 2D histogram with each pixel representing a bin for each channel. By defining a ground metric between pixels (e.g., the physical distance between pixels), we introduce the Wasserstein distance between images. This serves as the new ground metric for defining a Wasserstein distance between probability distributions over images. See Figure 2.

As mentioned in the introduction, the Wasserstein distance is also known as the Earth Mover's distance and is known as an effective metric in distinguishing images (Rubner et al., 2000). Motivated by this fact, we use the Earth Mover's distance (of images) as the ground metric,

$$\begin{aligned} d_{\mathcal{X}}(X,Y) :=& W_{q,d_\Omega}(X,Y) \\ =& \inf_{\pi} \left\{ \left(\mathbb{E}_{(x,y)\sim\pi} d_\Omega(x,y)^q \right)^{\frac{1}{q}} \right\}, \end{aligned} \quad (2)$$

where π is a joint distribution of (x, y) with marginals $x \sim X$, $y \sim Y$ both being images viewed as histograms over

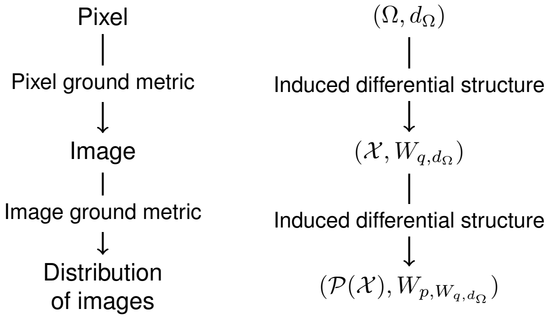

Pixel
Pixel ground metric
Image
Image ground metric
Distribution of images
(Ω, d_Ω)
Induced differential structure
$(\mathcal{X}, W_{q,d_\Omega})$
Induced differential structure
$(\mathcal{P}(\mathcal{X}), W_{p,W_{q,d_\Omega}})$

Figure 2. Illustration of Wasserstein-p loss function with Wasserstein-q ground metric.

pixels. Here $d_\mathcal{X} = W_{q,d_\Omega}(x,y)$ is named Wasserstein-q ground metric. It is defined with the pixel ground metric $d_\Omega \colon \Omega \times \Omega \to \mathbb{R}_+$ assigning distances to pairs of pixels.

In this work, combining the above approaches, we obtain a Wasserstein-p distance with Wasserstein-q ground metric as the loss function for training.

Definition 1. *Given a probability model $\{\mathbb{P}_G \colon G \in \Theta\} \subseteq \mathcal{P}_p(\mathcal{X})$ and a data distribution $\mathbb{P}_r \in \mathcal{P}_p(\mathcal{X})$, we propose the minimization problem*

$$\inf_G W_{p,W_{q,d_\Omega}}(\mathbb{P}_G, \mathbb{P}_r), \tag{3}$$

where $\mathcal{P}_p(\mathcal{X})$ is the set of densities with finite p-th moment, $W_{p,d_\mathcal{X}}$ is defined by (1) and W_{q,d_Ω} is given by (2).

The next example illustrates the difference between the proposed Wasserstein of Wasserstein loss and the Wasserstein loss with L^2 ground metric.

Motivating example. Consider the distribution $\mathbb{P}_r = \delta_X$ which assigns probability one to a single image X. Suppose the generative model attempts to estimate this via a distribution of the form $\mathbb{P}_G = \delta_Y$ which assigns probability one to a fake image Y. Now suppose that $X = \delta_x, Y = \delta_y$ are images with intensity 1 on pixel locations x, y, respectively, and intensity zero elsewhere. See Figure 3. In this case we have

$$W_{p,d_\mathcal{X}}(\mathbb{P}_r, \mathbb{P}_G) = d_\mathcal{X}(X, Y).$$

We check the following choices of the ground metric $d_\mathcal{X}$ between images X and Y.

1. Wasserstein-2 ground metric:

$$d_\mathcal{X}(X,Y) = W_{2,d_\Omega}(X,Y) = d_\Omega(x,y);$$

2. L^2 (Euclidean) ground metric:

$$d_\mathcal{X}(X,Y) = d_{L^2}(X,Y) = \begin{cases} 0 & \text{if } x = y \\ \text{constant} & \text{if } x \neq y \end{cases}.$$

We see that the Wasserstein distance with L^2 ground metric will assign two distant pixels the same cost as two adjacent pixels. This results in a highly discontinuous distance that is sensitive to single pixel translations! To make matters worse, in the case of continuous domain images, the L^2 distance will be infinite for all non-overlapping pixels. On the other hand, the Wasserstein of Wasserstein loss function is continuous with respect to continuous change of pixels in images. For learning image models with low dimensional support, the Wasserstein of Wasserstein loss function is still well defined, while the Wasserstein loss with L^2 ground metric function is ill-posed.

2.3. Duality formulation and properties

The computation required for the Wasserstein of Wasserstein loss function as stated in the previous section is unfeasible. To compute (3) one needs to handle a linear programming computation at both the level of probability distributions over images and individual images over pixels.

In this section, we present the Kantorovich duality formulation of Wasserstein of Wasserstein loss function with $p = 1$ and $q = 2$. As is done for Wasserstein GANs (Arjovsky et al., 2017), we consider an equivalent Lipschitz-1 condition, which can be practically applied in the framework of GANs.

Theorem 2 (Duality of Wasserstein of Wasserstein loss function)**.** *The Wasserstein-1 loss function over Wasserstein-2 ground metric has the following equivalent formulation:*

$$\begin{aligned} &W_{1,W_{2,d_\Omega}}(\mathbb{P}_G, \mathbb{P}_r) \\ =& \sup_{f \in C(\mathcal{X})} \Big\{ \mathbb{E}_{X \sim \mathbb{P}_G} f(X) - \mathbb{E}_{X \sim \mathbb{P}_r} f(X) \colon \\ &\int_\Omega \|\nabla_x \delta_X f(X)(x)\|^2_{d_\Omega} X(x) dx \leq 1 \Big\}, \end{aligned} \tag{4}$$

where ∇_x is the gradient operator in pixel space Ω and δ_X is the L^2 gradient in image space $\mathcal{X}$.

Proof. The result is from the duality of Wasserstein-1 metric, together with the Wasserstein-2 metric induced gradient operator. First, the Wasserstein-1 metric has a particular dual formulation, known as the Kantorovich duality:

$$W_{1,d_\mathcal{X}}(\mathbb{P}_0, \mathbb{P}_1) = \sup_f \mathbb{E}_{X \sim \mathbb{P}_0} f(X) - \mathbb{E}_{X \sim \mathbb{P}_1} f(X),$$

where the supremum is taken among all $f \colon \mathcal{X} \to \mathbb{R}$ satisfying a 1-Lipschitz condition with respect to the ground metric $d_\mathcal{X}$, i.e.,

$$\|\operatorname{grad} f(X)\|_{d_\mathcal{X}} \leq 1. \tag{5}$$

Second, consider the ground metric given by the Wasserstein-2 metric $d_\mathcal{X} = W_{2,d_\Omega}$ with ground metric d_Ω

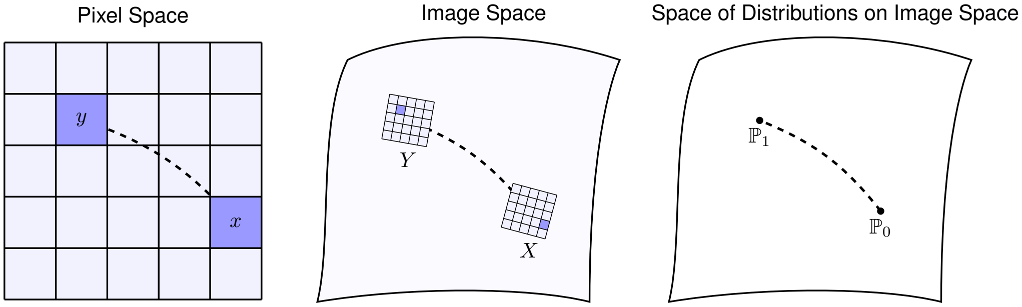


Figure 3. Depending on how we measure distances between pixel locations, the distance between images will be determined, and this in turn will determine how distances are measured between probability distributions.

of pixel space. Then the gradient operator in $(\mathcal{X}, d_{\mathcal{X}})$ is the Wasserstein-2 gradient, i.e.,

$$\mathrm{grad}\, f(X) = -\nabla_x \cdot (X(x)\nabla_x \delta_X f(X)(x)).$$

The 1-Lipschitz condition for $(\mathcal{X}, d_{\mathcal{X}})$ in equation 5 gives $\|\mathrm{grad}\, f(X)\|_{W_{2,d_\Omega}} \leq 1$, i.e.,

$$(\mathrm{grad}\, f(X), \mathrm{grad}\, f(X))_{W_{2,d_\Omega}} \leq 1.$$

It is rewritten as the following integral of the Lipschitz-1 condition w.r.t. the Wasserstein ground metric:

$$\int_\Omega \|\nabla_x \delta_X f(X)(x)\|^2_{d_\Omega} X(x) dx \leq 1.$$

Combining the above facts, we derive the formula for Wasserstein of Wasserstein loss function. □

Remark 3. *We note that the Kantorovich duality formula holds for any ground metric. The Wasserstein ground metric introduces differential structures and can be computed from the L^2 gradient. We review the Wasserstein gradient operators in Appendix A.*

The maximizer f in equation 4 corresponds to an Eikonal equation in image space $(\mathcal{X}, W_{2,d_\Omega})$. In other words, the Lipschitz-1 condition in Wasserstein norm has the form

$$\int_\Omega \|\nabla_x \delta_X f(X)(x)\|^2_{d_\Omega} X(x) dx = 1.$$

We call this equation the Wasserstein Eikonal equation.

Proposition 4 (Wasserstein Eikonal equation)**.** *The characteristic of characteristic for the Wasserstein Eikonal equation is the geodesic in pixel space.*

We defer the proof of the above proposition to Appendix A. Here the characteristic curve of our Eikonal equation is the geodesic curve in Wasserstein space $(\mathcal{X}, W_{2,d_\Omega})$. The characteristic curve of geodesics in Wasserstein space is again a geodesic in pixel space (Ω, d_Ω). We call this fact the *double characteristic property*. This is illustrated in Figure 3. In contrast, the characteristic of geodesics in L^2 space does not depend on pixel space. In the experiments section, we show that with the double characteristic property, the discriminator is continuous with respect to translations in pixel space, and is robust with respect to spatially independent noise added to the samples.

3. Wasserstein of Wasserstein GANs

In this section we apply the Wasserstein of Wasserstein loss function to implicit generative models.

3.1. Background

We start by reviewing generative adversarial networks (GAN). GANs are a deep learning approach to generative modelling that has demonstrated significant potential in the realm of image and text synthesis (Yu et al., 2017; Meng et al., 2018). The GAN model is composed of two competing agents: A discriminator and a generator. At each training step the generator produces synthesized images and the discriminator is given a batch of real and synthesized images to be classified as real or fake. The generator is trained to maximize the predictions of the discriminator while the discriminator is trained to classify generated images aside from real images. At the end of training the generator has learned how to trick the discriminator and ideally also estimate the underlying data distribution.

Mathematically if we define a trainable generative model $\mathbb{P}_G$ and discriminator D, the GAN objective formulation is as follows:

$$\min_{\mathbb{P}_G} \max_D \Big\{ \mathbb{E}_{x\sim\mathbb{P}_r} \log(D(X)) + \mathbb{E}_{x\sim\mathbb{P}_G} \log(1 - D(X)) \Big\}. \tag{6}$$

Here $\mathbb{P}_r$ is the true, or real, data distribution. The distribution $\mathbb{P}_G$ is defined in terms of a generator parameterized by $\theta \in \mathbb{R}^d$. Let the generator be given by $G_\theta \colon \mathbb{R}^m \to \mathcal{X};\ Z \mapsto X = G(\theta, Z)$. This takes a noise sample $Z \sim p(z) \in \mathcal{P}_2(\mathbb{R}^m)$ to an output sample with density given by $X = G(\theta, Z) \sim \rho(\theta, x) = \mathbb{P}_G$. Here $\mathbb{R}^d$ is the parameter space, $\mathbb{R}^m$ is the latent space, and $\mathcal{X}$ is the sample space.

The approach described above was found to suffer from difficulties at training, including lack of convergence and mode collapse, a phenomenon where the distribution $\mathbb{P}_G$ restricts to estimate a proper subset of $\mathbb{P}_r$. The above-mentioned challenges are often the result of the discontinuous nature of the loss in equation 6, and were also considered by Bernton et al. (2017). To resolve such problems, Arjovsky et al. (2017) proposed to use the Wasserstein metric with Euclidean ground metric as the objective, formulated as

$$\begin{aligned}&\min_{\mathbb{P}_G} W_{1,L^2}(\mathbb{P}_G, \mathbb{P}_r)\\ =&\min_{\mathbb{P}_G} \sup_{f \in C(\mathcal{X})} \Big\{\mathbb{E}_{X\sim\mathbb{P}_G} f(X) - \mathbb{E}_{X\sim\mathbb{P}_r} f(X) \colon \\ &\qquad \|\operatorname{grad} f(X)\|_2 \leq 1\Big\}.\end{aligned} \tag{7}$$

The Lipschitz condition in (7) was enforced via weight-clipping, ensuring $\|\operatorname{grad} f(X)\|_2 < C_0$, where C_0 is a constant. While now providing GAN with a continuous loss, WGAN with weight-clipping was noted to suffer from cyclic behavior and instability which was improved by Gulrajani et al. (2017) by changing the Lipschitz enforcing condition from hard weight-clipping to a soft gradient penalty term,

$$\begin{aligned}&\min_{\mathbb{P}_G} \sup_{f \in C(\mathcal{X})} \Big\{\mathbb{E}_{X\sim\mathbb{P}_G} f(X) - \mathbb{E}_{X\sim\mathbb{P}_r} f(X)\\ &+ \lambda \mathbb{E}_{X\sim\mathbb{P}_{\text{interp}}} (\nabla_X f(X) - 1)^2\Big\}.\end{aligned} \tag{8}$$

Here $\mathbb{P}_{\text{interp}}$ is an interpolation between $\mathbb{P}_r$ and $\mathbb{P}_G$, and λ is fixed. The gradient penalty term in equation 8 is not in full compliance with the Kantorovich duality of the problem as it also penalizes a discriminator of Lipschitz constants smaller than 1. To remedy this issue, Petzka et al. (2017) replace the gradient penalty term by

$$\lambda \mathbb{E}_{X\sim\mathbb{P}_{\text{interp}}} (\max(\nabla_X f(X) - 1, 0))^2.$$

We now derive our formulation that improves current methods which are based on the L^2 ground metric. Following Theorem 2, the Wasserstein of Wasserstein loss function can be rewritten to give the optimization problem

$$\begin{aligned}&\min_{\mathbb{P}_G} W_{1,W_{2,d_\Omega}}(\mathbb{P}_G, \mathbb{P}_r)\\ =&\min_{\mathbb{P}_G} \sup_{f \in C(\mathcal{X})} \Big\{\mathbb{E}_{X\sim\mathbb{P}_G} f(X) - \mathbb{E}_{X\sim\mathbb{P}_r} f(X) \colon \\ &\qquad \|\operatorname{grad} f(X)\|_{W_{2,d_\Omega}} \leq 1\Big\}.\end{aligned}$$

The above formulation is suitable for training GANs. Here we call the dual variable, f, the discriminator, while G is the generator. In the setting of GANs, neural networks are used to approximate the discriminator and generator, giving

$$\begin{aligned}&\min_\theta \sup_\phi \Big\{\mathbb{E}_{Z\sim p(z)} f_\phi(g(\theta, Z)) - \mathbb{E}_{X\sim\mathbb{P}_r} f_\phi(X) \colon \\ &\qquad \int_\Omega \|\nabla_x \delta_X f_\phi(X)(x)\|^2_{d_\Omega} X(x) dx \leq 1\Big\}.\end{aligned}$$

Here the generator G is expressed as a neural network with parameters $\theta \in \Theta$, and the discriminator is approximated by a neural network with parameters $\phi \in \Phi$. Our approach implements the 1-Lipschitz condition in terms of the Wasserstein gradient operator.

3.2. Discretization

We next present a discrete version of the Wasserstein-2 gradient. In practice, the image space $\mathcal{X}$ is not infinite dimensional, although in vision problems the dimension may be vast ($\mathcal{X} = \mathbb{R}^{28\times 28}$ or $\mathbb{R}^{32\times 32\times 3}$ for MNIST or CIFAR-10). To discretize, we first review the L^2-Wasserstein metric tensor (matrix) defined on a finite dimensional space. Consider a pixel space graph $\mathcal{G} = (V, E, \omega)$. Here $V = \{1, \dots, n\}$ is the vertex set (e.g., $n = 28 \times 28$), E is the edge set, and ω is a matrix of weights associated to the edges, with $\omega_{ij} = \omega_{ji}$, which defines a ground metric of pixels. We denote the neighborhood of node $i \in V$ by $N(i) = \{j \in V \colon (i,j) \in E\}$, and the degree of node i by $d_i = \frac{\sum_{j\in N(i)} \omega_{ij}}{\sum_{i=1}^n \sum_{i'\in N(i)} \omega_{ii'}}$. We can then define a Wasserstein-2 metric W on $\mathcal{X}$ (details in Appendix B), and further introduce the Wasserstein-2 gradient on discrete image space (cf. Solomon et al., 2014).

Proposition 5 (Wasserstein gradient on pixel space graph)**.** *Given a pixel space graph $\mathcal{G}$, the gradient of $f \in C^1(\mathcal{X})$ w.r.t. $(\mathcal{X}, W)$ satisfies*

$$\operatorname{grad} f(X) = L(X) \nabla_X f(X),$$

where ∇_X is the Euclidean gradient operator, and $L(X) \in \mathbb{R}^{n\times n}$ is the weighted Laplacian matrix defined as

$$L(X)_{ij} = \begin{cases} \frac{1}{2}\sum_{k\in N(i)} \omega_{ik}(\frac{X_i}{d_i} + \frac{X_k}{d_k}) & \textit{if } i = j;\\ -\frac{1}{2}\omega_{ij}(\frac{X_i}{d_i} + \frac{X_j}{d_j}) & \textit{if } j \in N(i);\\ 0 & \textit{otherwise}.\end{cases}$$

Moreover, the 1-Lipschitz condition w.r.t. $(\mathcal{X}, W)$, $\|\operatorname{grad} f(X)\|_W \leq 1$, is equivalent to

$$\nabla_X f(X)^\mathsf{T} L(X) \nabla_X f(X) \leq 1.$$

Remark 6. *We observe that the 1-Lipschitz condition is exactly the discrete analog of the one in equation (4),*

$$\begin{aligned}&\nabla_X f(X)^\mathsf{T} L(X) \nabla_X f(X)\\ =&\sum_{(i,j)\in E} \omega_{ij} (\nabla_{X_j} f(X) - \nabla_{X_i} f(X))^2 \frac{X_i/d_i + X_j/d_j}{2} \leq 1.\end{aligned}$$

We note that the Wasserstein gradient written in this form can be compared with the graph Laplacian on images (Bertozzi & Flenner, 2012; Zheng et al., 2011).

3.3. Computing the Wasserstein gradient via convolutions

We utilize the symmetry of the similarity graph of the image space to compute the Wasserstein gradient efficiently via convolutions as illustrated in Algorithm 2. We note the use of convolutions for the computation of the Wasserstein distance (Solomon et al., 2015; Bonneel et al., 2016) differs from ours as we merely compute the Wassserstein gradient. As the optimal transport plan can be defined for local distances and truncated at a given threshold, this leads to a sparse ω_{ij}, positive only for nearby pixels. We therefore can calculate all pairs $\nabla_{X_i} f(X) - \nabla_{X_j} f(X)$ with a given neighboring pattern by computing a set of kernels $K_{\mathcal{O}_1} \dots K_{\mathcal{O}_d}$ on the Euclidean gradient $\nabla_X f(X)$. The kernels $K_{\mathcal{O}_1} \dots K_{\mathcal{O}_d}$ are each defined as a convolution with fixed kernel of zeros with 1 and -1 in the corresponding neighbor pattern pixels. By creating a convolution filter for each neighbor pattern (e.g., right or up neighbor) we reach the desired output channels. In practice the different kernels $K_{\mathcal{O}_1} \dots K_{\mathcal{O}_k}$ are grouped to form a single 3D kernel. Likewise we apply the same kernel patterns, now with $\frac{1}{2}, \frac{1}{2}$ in the corresponding neighbor pattern pixels to obtain the terms $\frac{X_i/d_i + X_j/d_j}{2}$ for each i, j. This is done analogously, computing each kernel $M_{\mathcal{O}_k}$ over the images X/d. Applying entry-wise multiplication ($\odot$) and a summation collapsing all pixel locations and channels then yields an efficient and general method of calculating the Wasserstein gradient $\|\operatorname{grad} f\|_{W_{2,d(\Omega)}}$ for general local cost metrics on highly optimized convolution. The specific choice of the graph could serve to enhance different effects, which is a possibility that we leave for future study.

3.4. Wasserstein gradient regularization in GANs

We next adopt the gradient penalty into the loss function (cf. Petzka et al., 2017; Gulrajani et al., 2017) as

$$\min_\theta \sup_\phi \Big\{ \mathbb{E}_{z\sim p(z)} f_\phi(g(\theta, z)) - \mathbb{E}_{x\sim\mathbb{P}_r} f_\phi(x) + \lambda \mathbb{E}_{\hat{X}\sim\hat{\mathbb{P}}} \Big(\sqrt{\nabla_X f_\phi(\hat{X})^\intercal L(\hat{X}) \nabla_X f_\phi(\hat{X})} - 1 \Big)^2 \Big\},$$

where λ is chosen as a large constant and $\hat{\mathbb{P}}$ is the distribution of $\hat{X}$ taken to be the uniform on "Euclidean" lines connecting points drawn from $\mathbb{P}_G$ and $\mathbb{P}_r$. Our WWGAN training method is summarized in Algorithm 1.

Remark 7. *In practice, we may want to use images of unnormalized intensity, therefore the gradient penalty needs to account for change of total intensity. As proposed by Li (2018), we consider*

$$\tilde{L}(X) = \alpha \mathbf{1}\mathbf{1}^T + L(X). \tag{9}$$

Here $\mathbf{1} = (1, \dots, 1)^T \in \mathbb{R}^n$ is a constant vector. In Appendix C, we show how this adds one direction to the original tensor. Compared to $L(X)$ defined in the probability simplex, $\tilde{L}(X)$ is defined in the positive orthant. In the algorithm, we simply replace L by $\tilde{L}$ for unnormalized intensity.

Algorithm 1 WWGAN Gradient Penalty.

```
Require: Gradient penalty coefficient $\lambda$, discriminator iterations per generator iteration $n_{disc.}$, batch size $m$, ADAM hyperparameters $\alpha$, $\beta_1$, $\beta_2$, initial discriminator and generator parameters $\phi_0$ and $\theta_0$, $L$ matrix-function from graph structure for image space $\mathcal{G} = (V, E, \omega)$.
while $\theta$ has not converged do
    for $t = 1, \dots, n_{disc.}$ do
        for $i = 1, \dots, m$ do
            Sample real data $\boldsymbol{x} \sim \mathbb{P}_r$, latent variable $\boldsymbol{z} \sim p(\boldsymbol{z})$, a random number $\epsilon \sim U[0, 1]$.
            $\tilde{\boldsymbol{x}} \leftarrow G_\theta(\boldsymbol{z})$
            $\hat{\boldsymbol{x}} \leftarrow \epsilon \boldsymbol{x} + (1 - \epsilon)\tilde{\boldsymbol{x}}$
            $M^{(i)} \leftarrow D_\phi(\tilde{\boldsymbol{x}}) - D_\phi(\boldsymbol{x}) + \lambda(\sqrt{\nabla_{\hat{\boldsymbol{x}}} D_\phi(\hat{\boldsymbol{x}})^T L(\tilde{\boldsymbol{x}}) \nabla_{\hat{\boldsymbol{x}}} D_\phi(\hat{\boldsymbol{x}})} - 1)^2$
        end for
        $\phi \leftarrow$ Adam($\nabla_\phi \frac{1}{m} \sum_{i=1}^m M^{(i)}, \phi, \alpha, \beta_1, \beta_2$)
    end for
    Sample a batch of latent variables $\{\boldsymbol{z}^i\}_{i=1}^m \sim p(\boldsymbol{z})$
    $\theta \leftarrow$ Adam($\nabla_\theta \frac{1}{m} \sum_{i=1}^m -D_\phi(G_\theta(\boldsymbol{z}), \theta, \alpha, \beta_1, \beta_2)$)
end while
```

Algorithm 2 Wasserstein gradient norm $\|\operatorname{grad} f(X)\|_W$.

```
Require: The pixel graph: $\mathcal{G} = (V, E, \phi)$; local weights: $(w_{ij})$; neighbor relations arranged symmetrically: $\mathcal{O}_1 \dots \mathcal{O}_d$
Require: Euclidean gradient $\nabla_X f$
Wasserstein-grad $\leftarrow 0$
for neighbor relations $k = 1, \dots, d$ do
    Build kernel $K_{\mathcal{O}_k}$ to compute $\nabla_{X_i} f - \nabla_{X_{\mathcal{O}_k(i)}} f$
    Build corresponding kernel $M_{\mathcal{O}_k}$ to compute $\frac{X_i}{2d_i} + \frac{X_{\mathcal{O}_k}}{2d_{\mathcal{O}_k}}$
    $H \leftarrow K_{\mathcal{O}_k}(\nabla_X f)$
    $V \leftarrow M_{\mathcal{O}_k}(X)$
    $H \leftarrow H \odot H$    (entry-wise multiplication)
    $W \leftarrow H \odot V$
    Wasserstein-grad $\leftarrow$ Wasserstein-grad $+ \operatorname{sum}(W)$
end for
Return $\|\operatorname{grad} f(X)\|_W = \sqrt{\textit{Wasserstein-grad}}$
```

4. Experiments

In this section, we present experiments demonstrating the effects and utility of WWGAN. We use the CIFAR-10 and 64×64 cropped-CelebA image datasets. In both experiments the discriminator is a convolutional neural network with 3 hidden layers and leaky ReLU activations. For the generator we utilize a network with 3 hidden de-convolution layers and batch normalization (Ioffe & Szegedy, 2015). The dimensionality of the latent variable of the generator is set at 128. Batch normalization is not applied to the discriminator, in order to avoid dependencies when computing the gradient penalties. The model is then trained with the ADAM optimizer with fixed parameters $(\beta_1, \beta_2) = (0.9, 0)$. More implementation details are provided in Appendix C.

Figure 6 shows that in terms of computation time and quality of the generated images as measured by the Frechét Inception Distance (FID), WWGAN is comparable to state of the art WGAN-GP. Next, we take a look at the properties of the trained discriminators, which also serves to probe the shape of the probability densities over images defined by generators.

4.1. Perturbation stability

In this experiment we investigate how the discriminator trained with WWGAN on images benefits from the properties of the Wasserstein ground metric. Specifically, we test whether the discriminator trained with the new gradient penalty is more continuous with respect to natural variations of the images. Natural variabilities are continuous transformations of natural images that result in natural looking images, such as translations and rotations. If the transformations are applied gradually, one should expect to observe only gradual changes in the discriminator. The experiment is illustrated in Figure 4, where a randomly selected image from the CIFAR-10 dataset is gradually shifted vertically, shifting all pixels a single pixel downward at each step. In the figure, the sequence of shifted images is passed through the WWGAN and the WGAN-GP discriminators, which had been trained with their respective loss to reach an FID value of 40 for the generator. We observe with our WWGAN model, the discriminator values change continuously with the translation of the input image. In contrast, this type of continuity is not observed in models that are trained with the Euclidean Lipschitz condition. We note that WWGAN assigns a positive value to the image and gradually decreases to the end limit when the entire image is shifted away. Unlike WWGAN, WGAN-GP is highly sensitive to perturbations in image space and oscillates wildly, assigning highly positive (real label) and negative (fake labels) to images shifted less than 2 pixels away. We observed the same type of behavior across all images tested, as reported in Table 1.

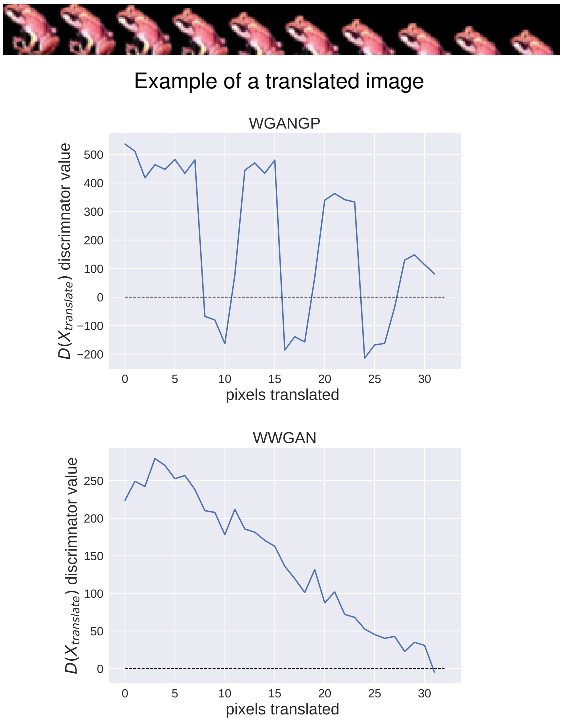


Figure 4. Discriminator for CIFAR-10 images translated by a vertical shift from 0 (no shift) to 32 pixels (complete image). The WWGAN discriminator is continuous to natural perturbations, e.g., vertical translation. WGAN-GP discriminator exhibits unpredictable behavior for small vertical perturbations, oscillating between real (positive values) and fake (negative values) labels. Both WWGAN, WGAN-GP discriminators tested were trained identically to reach an FID value of 40.

Method	Total variation (normalized)	zero-crossings
WGAN-GP	5.36	7.07
WWGAN	4.02	0.65

Table 1. For each image of the CIFAR-10 testing set we construct a vertical translation sequence and evaluate it on the discriminator of WWGAN and WGAN-GP. Normalized total variation and zero-crossing are computed for each curve and the average is reported. It is observed that WGAN-GP is more oscillatory than WWGAN.

4.2. Discriminator robustness to noise

In this experiment, we test the robustness of the discriminator to RGB salt and pepper noise, i.e., every pixel has a probability to be changed to either 0 or 1. In the plot 15% of the pixels are modified. We trained GANs with WGAN-GP and WWGAN until reaching an FID score of 40. We then measure the values of the trained discriminators on real images with RGB salt and pepper noise. In Figure 5, we see that WGAN-GP has separate clusters for noisy and clean images, while WWGAN is more robust to the noise and assigns more consistent values to all images.

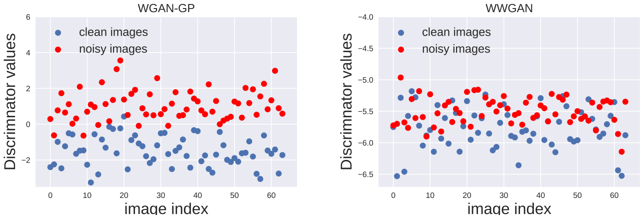


Figure 5. Robustness of the discriminator to noise on real CIFAR-10 images. The noise is RGB salt and pepper, where 15% of the pixels are modified. The WGAN-GP discriminator values cluster according to noise, giving different values to clean and noisy real images. The WWGAN discriminator is more robust to noise, and changes relatively little.

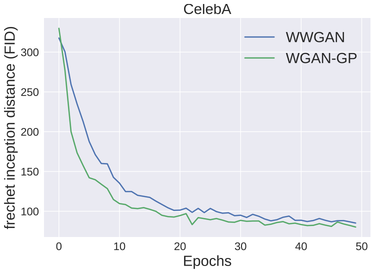


Figure 6. WWGAN gives comparable results with state of the art WGAN-GP training in terms of the FID of generated images. In terms of computation time, the overhead of WWGAN is negligible, with average epoch wall-clock times of 218.1 s and 236.9 s, respectively, in our experiments.

5. Related works

In this section, we review the connection between the proposed work and literature.

Ground Metric for function space. Banach GAN (Adler & Lunz, 2018) pointed out the importance of the ground metric in training with the Wasserstein loss. They apply Sobolev norms and their induced gradient operator. In contrast, we apply the optimal transport induced operator (Otto, 2001; Villani, 2009). The gradient operator depends on the new ground metric structure on sample space. We demonstrate that the optimal transport gradient provides for a practical 1-Lipschitz condition for training Wasserstein GANs.

Connection with Mean field games. Mean field games consider the optimal control problem in Wasserstein space (Cardaliaguet et al., 2015). In potential games, the Hamilton-Jacobi equation in Wasserstein space plays a vital role (Gangbo et al., 2008). In this paper, we present a new Hamilton-Jacobi equation in Wasserstein space. It is the Ekional equation in Wasserstein space as shown in Proposition 4. The new equation has naturally the double characteristics properties found in Mean field games. Here we demonstrate experimentally that the double characteristics property is very suitable for training GANs.

Geometric deep learning and Wasserstein metric on graphs. In geometric deep learning one considers mappings where the input space has a rich geometric structure (Bronstein et al., 2017). An example is the case where the input space consists of functions defined on a graph (e.g., raster images, where the graphs are grids). One can then define convolutions based on the group structures of these graphs. Here we propose to use a graph structure in the weighted Laplacian matrix. This matrix is connected to the *Wasserstein metric tensor on discrete space* (Chow et al., 2012; Maas, 2011; Mielke, 2011; Gu et al., 2015). A study is provided by Li (2018). The discrete Wasserstein metric tensor incorporates the graph structure of sample space into the training loss. The Wasserstein of Wasserstein loss function is an example in this direction.

Wasserstein natural gradients. Recent work also investigates natural gradients based on the Riemannian structures derived from optimal transport (Li & Montúfar, 2018). In this case, optimal transport serves to define an optimization method, rather than a loss function. This approach has also been applied to training of GANs, where it leads to an iterative regularizer for the generator (Lin et al., 2018).

6. Discussion

We proposed a Wasserstein loss function with Wasserstein ground metric for learning generative models. The Wasserstein ground metric introduces a graph / manifold structure into the sample space of the model and allows us to introduce meaningful priors to the learning model. Experiments demonstrate that this approach can contribute to making the generator and discriminator in GANs more stable with respect to noise and the natural variability of image data.

We consider the Wasserstein of Wasserstein loss an important advance at a conceptual level. It has a physical intuition. Consider a physical motion or translation in pixel space. It corresponds to a change in image space, and it changes the distribution over images accordingly. The double characteristic property of the Wasserstein Eikonal equation reflects this intuition analytically. We regard it as surprising that this high level approach can be translated to practical computational methods. Remarkably, our approach has a very small additional computational cost over the standard Wasserstein loss function with L^2 (Euclidean) ground metric.

In the future, we suggest to explore the consequences of our approach from the statistical and optimization point of view. Also, to continue exploring the role of the graph structure that is chosen to define the Wasserstein ground metric in relation to specific data types.

Acknowledgements

YD has been supported by the National Science Foundation under Graduate Research Fellowship Grant No. DGE-1650604. AL has been supported by AFOSR MURI FA9550-18-1-050. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n^{o} 757983).

References

Adler, J. and Lunz, S. Banach Wasserstein GAN. *arXiv:1806.06621 [cs, math]*, 2018.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein GAN. *arXiv:1701.07875 [cs, stat]*, 2017.

Bernton, E., Jacob, P. E., Gerber, M., and Robert, C. P. Inference in generative models using the wasserstein distance. *arXiv preprint arXiv:1701.05146*, 2017.

Bertozzi, A. L. and Flenner, A. Diffuse interface models on graphs for classification of high dimensional data. *Multiscale Modeling & Simulation*, 10(3):1090–1118, 2012.

Bonneel, N., Peyré, G., and Cuturi, M. Wasserstein barycentric coordinates: histogram regression using optimal transport. *ACM Transactions on Graphics (TOG)*, 35(4):71, 2016.

Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., and Vandergheynst, P. Geometric deep learning: Going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4): 18–42, 2017.

Cardaliaguet, P., Delarue, F., Lasry, J.-M., and Lions, P.-L. The master equation and the convergence problem in mean field games. *arXiv:1509.02505*, 2015.

Chow, S.-N., Huang, W., Li, Y., and Zhou, H. Fokker–Planck Equations for a Free Energy Functional or Markov Process on a Graph. *Archive for Rational Mechanics and Analysis*, 203(3):969–1008, 2012.

Deshpande, I., Zhang, Z., and Schwing, A. G. Generative modeling using the sliced wasserstein distance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3483–3491, 2018.

Engquist, B. and Yang, Y. Seismic imaging and optimal transport. *arXiv:1808.04801*, 2018.

Frogner, C., Zhang, C., Mobahi, H., Araya-Polo, M., and Poggio, T. Learning with a Wasserstein Loss. *arXiv:1506.05439 [cs, stat]*, 2015.

Gangbo, W., Nguyen, T., and Tudorascu, A. Hamilton-jacobi equations in the wasserstein space. *Methods Appl. Anal.*, 15(2):155–184, 06 2008.

Gu, J., Hua, B., and Liu, S. Spectral distances on graphs. *Discrete Applied Mathematics*, 190:56–74, 2015.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved training of wasserstein GANs. In *Advances in Neural Information Processing Systems 30*, pp. 5767–5777. Curran Associates, Inc., 2017.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

Li, W. Geometry of probability simplex via optimal transport. *arXiv:1803.06360 [math]*, 2018.

Li, W. and Montúfar, G. Natural gradient via optimal transport. *Information Geometry*, 1(2):181–214, Dec 2018.

Lin, A., Li, W., Osher, S., and Montúfar, G. Wasserstein proximal of GANs. *CAM report 18-53*, 2018.

Maas, J. Gradient Flows of the Entropy for Finite Markov Chains. *Journal of Functional Analysis*, 261(8):2250–2292, 2011.

Meng, R., Cui, Q., and Yuan, C. A survey of image information hiding algorithms based on deep learning. *Computer Modeling in Engineering & Sciences*, 117:425–454, 12 2018.

Mielke, A. A Gradient Structure for Reaction–diffusion Systems and for Energy-Drift-Diffusion Systems. *Nonlinearity*, 24(4):1329, 2011.

Montavon, G., Müller, K.-R., and Cuturi, M. Wasserstein Training of Restricted Boltzmann Machines. In *Advances in Neural Information Processing Systems 29*, pp. 3718–3726. Curran Associates, Inc., 2016.

Mroueh, Y., Sercu, T., and Goel, V. McGan: Mean and covariance feature matching GAN. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2527–2535. PMLR, 2017.

Nowak, E., Jurie, F., and Triggs, B. Sampling strategies for bag-of-features image classification. In *European Conference on Computer Vision*, pp. 490–503. Springer, 2006.

Otto, F. The Geometry of Dissipative Evolution Equations: The Porous Medium Equation. *Communications in Partial Differential Equations*, 26(1-2):101–174, 2001.

Petzka, H., Fischer, A., and Lukovnicov, D. On the regularization of Wasserstein GANs. *arXiv:1709.08894 [cs, stat]*, 2017.

Puthawala, M. A., Hauck, C. D., and Osher, S. J. Diagnosing forward operator error using optimal transport. *arXiv:1810.12993*, 2018.

Rubner, Y., Tomasi, C., and Guibas, L. J. The Earth Mover's Distance as a Metric for Image Retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

Solomon, J., Rustamov, R., Guibas, L., and Butscher, A. Earth mover's distances on discrete surfaces. *ACM Transactions on Graphics (TOG)*, 33(4):67, 2014.

Solomon, J., De Goes, F., Peyré, G., Cuturi, M., Butscher, A., Nguyen, A., Du, T., and Guibas, L. Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (TOG)*, 34(4):66, 2015.

Villani, C. *Optimal Transport: Old and New*. Number 338 in Grundlehren der mathematischen Wissenschaften. Springer, Berlin, 2009.

Yu, L., Zhang, W., Wang, J., and Yu, Y. Seqgan: Sequence generative adversarial nets with policy gradient. In *AAAI*, pp. 2852–2858, 2017.

Zhang, J., Marszałek, M., Lazebnik, S., and Schmid, C. Local features and kernels for classification of texture and object categories: A comprehensive study. *International Journal of Computer Vision*, 73(2):213–238, 2007.

Zheng, M., Bu, J., Chen, C., Wang, C., Zhang, L., Qiu, G., and Cai, D. Graph regularized sparse coding for image representation. *IEEE Transactions on Image Processing*, 20(5):1327–1336, 2011.