Sinkhorn Adversarial Attack and Defense

A V Subramanyam, *Member, IEEE*

***Abstract*—Adversarial attacks have been extensively investigated in the recent past. Quite interestingly, a majority of these attacks primarily work in the l_p space. In this work, we propose a novel approach for generating adversarial samples using Wasserstein distance. Unlike previous approaches, we use an unbalanced optimal transport formulation which is naturally suited for images. We first compute an adversarial sample using a gradient step and then project the resultant image into Wasserstein ball with respect to original sample. The attack introduces perturbation in the form of pixel mass distribution which is guided by a cost metric. Elaborate experiments on MNIST, Fashion-MNIST, CIFAR-10 and Tiny ImageNet demonstrate a sharp decrease in the performance of state-of-art classifiers. We also perform experiments with adversarially trained classifiers and show that our system achieves superior performance in terms of adversarial defense against several state-of-art attacks. Our code and pre-trained models are available at https://bit.ly/2SQBR4E.**

***Index Terms*—Sinkhorn, Dual Wasserstein, Adversarial attack and defense.**

I. Introduction

THE radical success of deep learning has led to wide scale deployment of deep learning systems for multiple tasks in real-world. While these systems have shown tremendous progress over the years, they are known to be vulnerable to imperceptible perturbations [1], [2]. In particular, when deep learning systems are deployed in sensitive applications such as autonomous vehicles, face recognition or malware detection, it is necessary to elaborately evaluate the system' robustness against various adversarial attacks.

The goal of adversarial attack in a classification setting is defined as follows. Let x be an input sample, y be the ground truth label, $\hat{y}$ be a label other than y, and F is a trained classification model. Then, in order to generate an adversarial sample, the adversary adds an imperceptible perturbation Δx to x such that,

$$F(x + \Delta x) \neq y \text{ or } F(x + \Delta x) = \hat{y}. \tag{1}$$

Szegedy *et al.* [1] proposed the first work on generation of adversarial sample for deep neural networks. The authors formulated it as an optimization problem and obtained the adversarial sample using an additive perturbation by solving the objective using L-BFGS. Since then, a lot of works have been proposed in this direction. Noting that L-BFGS is computationally expensive, Goodfellow *et al.* [2] proposed a fast gradient sign method (FGSM) to efficiently compute the adversarial sample. Kurakin *et al.* [3] extended FGSM for targeted attacks where $F(x + \Delta x)$ can be guided to output a target label other than the ground truth label. These attacks can be categorised as l_p norm based attacks. In particular, a vast majority of attacks are proposed using l_p norm only. However, some notable works perform the attacks using Wasserstein distance [4], [5], [6], [7], the focus of our work.

In this work, we propose a Wasserstein distance based attack. Prominent works in this domain, Wong *et al.* [4] and Hu *et al.* [5], derive the objective formulation from balanced optimal transport (OT) problem. Balanced OT is defined under the assumption that the total probability masses between the distributions over which the Wasserstein distance is computed are the same. However, this definition is restrictive when the distance is to be computed over signals such as images. In order to address this issue, some works relax the equality constraints [8], [9] and solve a partial transport problem. In our work, we cast the formulation as an unbalanced optimal transport problem where the requirement of same probability masses is relaxed [10]. Unbalanced OT has also shown a promising performance in other domains such as computational imaging [11], domain adaptation [12], multi-label classification [13], document retrieval [14], crowd counting [15], natural language generation [16].

Another advantage of Wasserstein distance, in contrast to pixel-wise l_p distance, is that it accounts for the geometry of the signals by incorporating a cost matrix for transporting mass from source to target. Wasserstein distance has been used in tasks such as image synthesis [17], [18] and has demonstrated to be better metric compared to pixel-wise losses like MSE. We show an illustration using Figure 1. We can see that both images are very similar to human vision. However, the l_∞ distance is 1 between them. On the other hand Wasserstein distance is only 0.0015. This indicates that though the images can be perceptually very similar, l_∞ distance can be very large compared to Wasserstein distance. In other words, even a minor adversarial attack may overshoot the available budget under l_∞ norm. Whereas, Wasserstein distance may lead to large distance only when the attack is moderate to severe and under such attacks it shall significantly deteriorate the performance of the undefended model. Thus, Wasserstein distance may be better suited while adversarially modifying the images.

In the proposed work, we first obtain an l_p norm perturbed sample using FGSM which is then projected into the vicinity of original sample measured in terms of Wasserstein distance. While projecting the sample into Wasserstein ball, the objective also minimizes l_2 distance between output adversarial sample and l_p norm perturbed sample. We solve the problem in dual domain using modified Sinkhorn iterations [19]. Further, we also demonstrate that when the adversarial samples are used for training the model, the trained model exhibits a strong defense against several state-of-art attacks.

A V Subramanyam is with the Department of Electronics and Communication Engineering, Indraprastha Institute of Information Technology, New Delhi, India, 110020 India. E-mail: subramanyam@iiitd.ac.in.

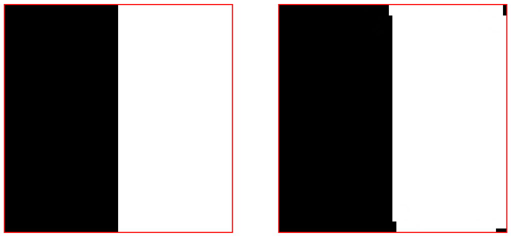

Fig. 1: Original (left) and rotated (right) image. Red box indicates the boundary of image and is not a part of the image itself.

II. Related Works

Adversarial Attacks FGSM' simplicity and efficiency triggered a huge interest in gradient based attacks which are predominantly in l_p space. Kurakin *et al.* [3] proposed a basic iterative FGSM which generates adversarial examples using small step size. The attack is more severe than one-step FGSM and is also scalable to large scale datasets like ImageNet. Madry *et al.* [20] introduced projected gradient descent as a universal first-order adversary and demonstrated it as the strongest first order attack. This work also emphasizes that more complex models can perform better against one-step perturbations. However, this also decreases the transferability.

Carlini and Wagner [21] introduced l_2, l_0 and l_∞ attacks to demonstrate the vulnerability of neural networks against defensive distillation [22]. They evaluate a combination of seven different objective functions with 3 different box constraints and highlight that cross-entropy based objective function is the worst performing among all objective functions. Further, they also investigate transferability of attacks. Here, an unsecured standard model is used to determine strongly misclassified adversarial example which can also successfully attack the distilled models.

Papernot *et al.* [23] proposed a Jacobian based method to construct an adversarial saliency map. The saliency map is constructed using forward derivative of the network with respect to the input features. The derivatives which take higher positive values lead to high saliency. These high saliency values then identify whether the corresponding features will increase the likelihood of target class or decrease the likelihood of other classes. Further, the features are perturbed to obtain the adversarial samples. This method has a significant benefit as it perturbs only a small fraction of the input features. Saliency maps have also been used in other works such as [24].

Athalye *et al.* [25] synthesized 2D and 3D adversarial objects using an expectation over transformation. Since many attacks do not survive the real world scenarios like viewpoint variations, authors propose to use an expectation over affine transformations or rendering of texture in case of 3D. In [26], Croce and Hein proposed AutoAttack which addresses the fixed step size, budget and optimization issues of projected gradient descent based attacks.

In [27], Dong *et al.* identified that iterative FGSM is less transferable as it tends to overfit the model. To address this issue, the authors proposed a momentum based iterative FGSM which is both stronger and transferable compared to basic iterative methods. The authors point out that iterative methods easily get trapped into local maxima which results in poor transferabilty as the decision boundaries of different models are not the same. On the other hand, incorporating momentum stabilizes the update direction and allows escaping from local maxima. Su *et al.* [28] proposed an extreme attack by modifying the RGB values of a single pixel using differential evolution [29]. This method does not use any gradient information and has better transferability as very less target model information is needed.

In [30], Zhang *et al.* theoretically analyzed the regularization terms for adversarial defense. The authors show that the presence of additional regularization term that minimizes the difference between predictions for clean and attacked samples lead to an upper bound on the error between accuracies for robust models and standard models. Wong *et al.* point out that adversarial training using FGSM where the adversarial perturbation from previous iteration is used is not robust. They propose to use random initialization and demonstrate that adversarial robustness can be achieved via standard adversarial training [31]. In [32], authors proposed to find a direction normal to decision classifier and iteratively perturb the sample till an adversarial example is generated. Xie *et al.* proposed Diverse Inputs Iterative Fast Gradient Sign Method (DI^2-FGSM) to improve the transferability of attacks. Before feeding the image to an iterative FGSM, the image is processed using different transformations. The diverse patterns that are generated due to transformations lead to better transferability [33].

Adversarial Defense Countering adversarial attacks, the goal of adversarial defense is to achieve the accuracy comparable to that of untargeted model. The defense methods either use adversarial examples during training or modify the network itself. Adversarial training is often considered as a first line of defense [1], [2], [32] and also demonstrates the strongest defense. Among other class of defenses which modify the network are defensive distillation [22], gradient regularization [34], biologically inspired models [35], [36], convex ReLU relaxation [37], image enhancement [38], image restoration [39].

Entropic Regularized Optimal Transport

$$f(\Pi) = \min_{\Pi \in \mathcal{U}(x,z)} \langle C, \Pi \rangle + \gamma \langle \Pi, \ln \Pi \rangle \tag{2}$$

$$\mathcal{U}(x,z) = \Pi \in R_{+}^{n \times n} : \Pi \mathbf{1} = x, \Pi^{\top} \mathbf{1} = z,$$

where Π is the transportation plan, $\ln \Pi$ operates element-wise, $C_{+}^{n \times n}$ is the cost matrix, and, $\mathbf{1}$ is an $n-$dimensional vector of all ones. $\langle ., . \rangle$ denotes Frobenius product of matrices. Since the term $\gamma \langle \Pi, \ln \Pi \rangle$ is strongly convex, the objective in Equation 2 is strongly convex and admits an optimal solution. In addition, the higher computational complexity for computation of exact OT ($\mathcal{O}(n^3 logn)$) owing to interior-point methods [40], is also addressed by this entropic regularized version and has been demonstrated to achieve an $\mathcal{O}(n^2)$ in the celebrated work by Cuturi *et al.* [41].

Regularized Balanced Optimal Transport In this paper, we focus on Wasserstein space attacks. In contrast to pixel based distance measures, Wasserstein distances incorporate the geometry of the pixels. Quite recently, Wong *et al.* [4] proposed a projected Sinkhorn attack characterized by projected gradient descent followed by projection onto Wasserstein ball. More formally, let $l(x, y)$ be a cross-entropy loss, α be step-size and ∇_x denotes the gradient of the function with respect to x. Then,

$$w = x + \alpha \nabla_x l(x, y) \tag{3}$$

Now, w can be projected either into an l_p ball or Wasserstein ball. Here, we consider projection into Wasserstein ball only. Then, to drop w into Wasserstein ball of ϵ radius, we solve for,

$$\min_{z,\Pi} \frac{\lambda}{2}\|w - z\|_2^2 + \sum_{ij} \Pi_{ij} \ln \Pi_{ij} \tag{4}$$
$$\text{subject to } \Pi\mathbf{1} = x, \Pi^\top \mathbf{1} = z, \langle \Pi, C\rangle < \epsilon.$$

Upon projection into Wasserstein ball, the images are clamped such that the pixels are in the range [0, 1]. Due to this clamping, the algorithm overshoots the available budget. Hu *et al.* [5] improve this shortcoming by adding an l_∞ constraint on z and solve the following,

$$\min_{z,\Pi} \frac{\lambda}{2}\|w - z\|_2^2 + \sum_{ij} \Pi_{ij} \ln \Pi_{ij} \tag{5}$$
$$\text{subject to } \Pi\mathbf{1} = x, \Pi^\top \mathbf{1} = z, \langle \Pi, C\rangle < \epsilon, z_j \le \frac{1}{\|w\|_1}.$$

Hu *et al.* also show that l_2 norm based PGD step with large step-size is effective compared to l_∞ norm.

Equations 4 and 5 use a regularized version of OT and in practice compute an approximate Wasserstein distance. In order to compute exact Wasserstein distance, Wu *et al.* [6] propose a dual projection method and apply Frank-Wolfe algorithm to obtain the optimal transport matrix. In addition to the attacks, certified robustness against Wasserstein attacks based on Wasserstein smoothing has also been proposed [42]. Wasserstein distance based feature matching is also demonstrated to be prominent defense mechanism in [43].

Unbalanced Optimal Transport The entropic regularized OT can only be used when the total probability masses are same. This restriction naturally precludes employing entropic regularized OT to pixel domain. In order to address this issue, unbalanced OT has been proposed [44], [45], [10]. Unbalanced OT uses KL divergence instead of marginal equality constraints and solves the formulation using the Fenchel-Legendre dual form [46]. Such relaxed formulation has also been proposed in [13], though solved in the primal form.

III. Methodology

In this section we discuss our proposed objective formulation and analytically derive the solution. We also show a geometric convergence proof. The formulations proposed in Equations 4 and 5 need the marginals (x, z) to be probability vectors [41]. However, images do not inherently lie in probability simplexes and normalizing them to probability vectors leads to information loss [47], [44], [10]. To overcome this, we relax the equality constraints as,

$$\min_{\Pi} \langle \Pi, C\rangle + \eta\langle \Pi, \ln \Pi\rangle + \tau\Phi(\Pi\mathbf{1}, x) + \tau\Phi(\Pi^\top\mathbf{1}, z), \tag{6}$$

where, $\Phi(a, b)$ is a divergence measure. In this work, we use $\Phi(a, b) = KL(a\|b) = \sum_{i=1}^n a_i \log\left(\frac{a_i}{b_i}\right) - a_i + b_i$. Note that the generalized KL divergence definition follows from [48]. Smooth measures such as l_2 can also be applied [49], [50]. We provide a discussion on l_2 regularization in Section IV.

By applying Fenchel-Legendre conjugate dual [46], the formulation in Equation 6 can be re-written as,

$$\max_{\alpha,\beta} -F^*(-\alpha) - G^*(-\beta) - \eta \sum_{ij} \exp(\frac{\alpha_i + \beta_j - C_{ij}}{\eta}),$$

where,

$$F^*(\alpha) = \max_{\Pi} \Pi^\top \alpha - \tau KL(\Pi\mathbf{1}||x)$$
$$G^*(\beta) = \max_{\Pi} \Pi^\top \beta - \tau KL(\Pi^\top\mathbf{1}||z)$$

Now, we need that w, obtained in Equation 3, be dropped into Wasserstein ball with respect to x. In other words, we need the output adversarial sample z which is closer to w in l_2 sense and lies in a given Wasserstein ball with respect to clean sample x. Thus, the objective can be written as,

$$\min_{\alpha,\beta,z} h\big(\alpha, \beta, z\big) = \eta \sum_{ij} \exp(\frac{\alpha_i + \beta_j - C_{ij}}{\eta}) + \tag{7}$$
$$\tau\langle \exp(-\alpha/\tau), x\rangle + \tau\langle \exp(-\beta/\tau), z\rangle + \frac{\gamma}{2}\|z - w\|^2,$$

where, $\|.\|$ denote l_2 norm. We solve for each variable independently by taking derivative with respect to single variable and setting to zero.

Solving for α

To solve for α, we minimize the following equation,

$$\min_{\alpha} \eta \sum_{ij} \exp(\frac{\alpha_i + \beta_j - C_{ij}}{\eta}) + \tau\langle \exp(-\alpha/\tau), x\rangle \tag{8}$$

Taking derivative of Equation 8 wrt. α_i,

$$\nabla_\alpha h(\alpha, \beta, z) = e^{\frac{\alpha_i}{\eta}} \sum_j e^{(\beta_j - C_{ij})/\eta} - x_i e^{-\frac{\alpha_i}{\tau}} \tag{9}$$

Setting Equation 9 to zero gives,

$$\frac{\alpha_i}{\eta} + \ln\big(\sum_j e^{(\beta_j - C_{ij})/\eta}\big) = \ln x_i - \frac{\alpha_i}{\tau} \tag{10}$$

$$\alpha_i^{k+1} = \big[\ln x_i - \ln\big(\sum_j e^{(\beta_j^k - C_{ij})/\eta}\big)\big]\frac{\eta\tau}{\eta + \tau}, \tag{11}$$

where k denotes the iteration index. We can further manipulate Equation 11 to obtain,

$$\alpha_i^{k+1} = \big[\frac{\alpha_i^k}{\eta} + \ln x_i - \ln\big(\sum_j e^{(\alpha_i^k + \beta_j^k - C_{ij})/\eta}\big)\big]\frac{\eta\tau}{\eta + \tau}. \tag{12}$$

Solving for β

To obtain β, we solve for

$$\min_{\beta} \eta \sum_{ij} \exp(\frac{\alpha_i + \beta_j - C_{ij}}{\eta}) + \tau\langle \exp(-\beta/\tau), z\rangle + \frac{\gamma}{2}\|z - w\|^2.$$

Similar to solution for α, we obtain,

$$\beta_j^{k+1} = \big[\frac{\beta_j^k}{\eta} + \ln z_j - \ln\big(\sum_i e^{(\alpha_i^k + \beta_j^k - C_{ij})/\eta}\big)\big]\frac{\eta\tau}{\eta+\tau}. \quad (13)$$

Solving for z

To obtain z, we solve for

$$\min_{z} \tau\langle \exp(-\beta/\tau), z\rangle + \frac{\gamma}{2}\|z - w\|^2.$$

Then,

$$\nabla_z h(\alpha, \beta, z) = \tau e^{-\beta_i/\tau} + \gamma(z_i - w_i) \quad (14)$$

which gives,

$$z^{k+1} = w + \frac{\tau}{\gamma} e^{-\beta^{k+1}/\tau} \quad (15)$$

We iteratively solve for α, β and z for a fixed number of iterations. Since the iterations alternatively update α and β, the algorithm can be considered to perform Sinkhorn-like iterations [44], [10]. We present these steps in Algorithm 1.

Algorithm 1: Modified Sinkhorn iterations for computing adversarial sample

```
Input: $k = 0, \alpha^0 = \beta^0 = 0, \eta = 0.001, \tau = 0.01, \gamma = 0.5, B(\alpha^k, \beta^k) = diag(e^{\alpha^0/\eta})e^{-C/\eta}diag(e^{\beta^0/\eta})$
Output: $z, B(\alpha^k, \beta^k)$
while convergence do
    $r^k = B(\alpha^k, \beta^k)\mathbf{1}_n = \sum_j e^{(\alpha_i^k - C_{ij}/\eta + \beta_j^k)/\eta}$
    $q^k = B(\alpha^k, \beta^k)^\top \mathbf{1}_n = \sum_i e^{(\alpha_i^k - C_{ij}/\eta + \beta_j^k)/\eta}$
    For even $k$
        $\alpha^{k+1} = \big[\frac{\alpha^k}{\tau} + \ln(x) - \ln(r^k)\big]\frac{\eta\tau}{\eta+\tau}$
        $\beta^{k+1} = \beta^k$
    For odd $k$
        $\beta^{k+1} = \big[\frac{\beta^k}{\tau} + \ln(z) - \ln(q^k)\big]\frac{\eta\tau}{\eta+\tau}$
        $\alpha^{k+1} = \alpha^k$
        $z^{k+1} = w + \frac{\tau}{\gamma}\exp(-\frac{\beta^{k+1}}{\tau})$
    $B(\alpha^k, \beta^k) = diag(e^{\alpha^k/\eta})e^{-C/\eta}diag(e^{\beta^k/\eta})$
    $k = k + 1$
end
```

Lemma 1 Let (α^*, β^*) be the optimal solution of Equation 7. Then, the sup norms of the optimal solution $\|\alpha^*\|_\infty$, $\|\beta^*\|_\infty$ and $\|z\|_\infty$ are bounded by,

$$\max\{\|\alpha^*\|_\infty, \|\beta^*\|_\infty\} \leq \tau R \quad (16)$$

$$\|z\|_\infty \leq \|w + \tau/\gamma e^{-\min(\beta^*/\tau)}\|_\infty \quad (17)$$

$$= \|w\|_\infty + \tau/\gamma e^{\|\beta^*\|_\infty/\tau}, \quad (18)$$

where $R = \max(\|\ln(x)\|_\infty, \|\ln(z)\|_\infty) + \max(\ln(n), \frac{1}{\eta}\|C\|_\infty - \ln(\eta))$. The proof for $\|\alpha^*\|_\infty$ and $\|\beta^*\|_\infty$ can be directly obtained from [10]. For the sake of completeness, we provide the complete proof in appendix B. The proof for $\|z\|_\infty$ is also straightforward as $w \in R_+^n$.

Lemma 2 Updates $(\alpha^{k+1}, \beta^{k+1})$ from Algorithm 1 satisfies the following bound

$$\delta^{k+1} \leq \delta^k, \quad (19)$$

where, $\delta^{k+1} = \big(\frac{\tau}{\tau+\eta}\big)^k\|\beta^*\|_\infty$. The proof can be directly extended from [10]. We also provide the proof in appendix B for the self-sufficiency of this article.

Lemma 3 Update z^{k+1} from Algorithm 1 satisfies the following,

$$z^{k+1} - z^* = \frac{\tau}{\gamma}\big(e^{-\frac{\beta^{k+1}}{\tau}} - e^{-\frac{\beta^*}{\tau}}\big) \quad (20)$$

$$= \frac{\tau}{\gamma}e^{-\frac{\beta^*}{\tau}}\big(e^{\frac{\beta^* - \beta^{k+1}}{\tau}} - 1\big) \quad (21)$$

$$\|z^{k+1} - z^*\|_\infty \leq \frac{\tau}{\gamma}e^{\frac{\|\beta^*\|_\infty}{\tau}}\big(e^{\frac{\|\beta^{k+1} - \beta^*\|_\infty}{\tau}} - 1\big) \quad (22)$$

$$\leq \frac{\tau}{\gamma}e^{\frac{\|\beta^*\|_\infty}{\tau}}\big[e^{\tau\left(\frac{\tau}{\tau+\eta}\right)^k\|\beta^*\|_\infty} - 1\big] \quad (23)$$

Lemma 4 The measures α and β are empirically observed. Let the empirical observations be represented as $\hat{\alpha}$ and $\hat{\beta}$. In this lemma, we show that the absolute deviation made by the approximation of α and β with $\hat{\alpha}$ and $\hat{\beta}$ given by,

$$|\mathrm{E}[h(\alpha^*, \beta^*, z^*)] - h(\hat{\alpha}, \hat{\beta}, \hat{z})| \leq 2B, \quad (24)$$

where (α^*, β^*, z^*) is the optimal solution, and $B = \max(\eta e^{(\|\alpha\|_\infty + \|\beta\|_\infty)/\eta}, \gamma\|z\|_\infty + \tau e^{\|\beta\|_\infty/\tau})$.

We first show that $h(.)$ is Lipschitz continuous.

$$\nabla_\alpha h(\alpha, \beta, z) = e^{\frac{\alpha_i}{\eta}}\sum_j e^{(\beta_j - C_{ij})/\eta} - x_i e^{-\frac{\alpha_i}{\tau}} \quad (25)$$

$$\leq e^{\|\alpha\|_\infty}\sum_j e^{\|\beta\|_\infty/\eta} \quad (26)$$

$$\leq n e^{(\|\alpha\|_\infty + \|\beta\|_\infty)/\eta} \quad (27)$$

$$\nabla_z h(\alpha, \beta, z) = \tau e^{-\beta_i/\tau} + \gamma(z_i - w_i) \quad (28)$$

$$\leq \gamma z_i + \tau e^{-\beta_i/\tau} \quad (29)$$

$$\leq \gamma\|z_i\|_\infty + \tau e^{\|\beta_i\|_\infty/\tau} \quad (30)$$

Thus,

$$\nabla h(\alpha, \beta, z) \leq \max(n e^{(\|\alpha\|_\infty + \|\beta\|_\infty)/\eta}, \gamma\|z\|_\infty + \tau e^{\|\beta\|_\infty/\tau}) \quad (31)$$

Now we show the proof.

$$\mathrm{E}[h(\alpha^*, \beta^*, z^*)] - h(\hat{\alpha}, \hat{\beta}, \hat{z}) \quad (32)$$

$$= \mathrm{E}[h(\alpha^*, \beta^*, z^*)] - h(\alpha^*, \beta^*, z^*) + \quad (33)$$
$$h(\alpha^*, \beta^*, z^*) - h(\hat{\alpha}, \hat{\beta}, \hat{z})$$

The last two components of Equation 33 is non-positive as $\hat{\alpha}, \hat{\beta}, \hat{z}$ are only empirical minimizers. Thus, we obtain,

$$|\mathrm{E}[h(\alpha^*, \beta^*, z^*)] - h(\hat{\alpha}, \hat{\beta}, \hat{z})| \leq \sup_{\alpha,\beta,z} |\mathrm{E}[h(\alpha, \beta, z)] - h(\alpha, \beta, z)| \tag{34}$$

Using [51], [52], we get

$$|\mathrm{E}\{h(\alpha^*, \beta^*, z^*)\} - h(\hat{\alpha}, \hat{\beta}, \hat{z}))| \leq 2B \tag{35}$$

IV. Discussion

In this section, we discuss about an alternate formulation using l_2 measure in place of KL divergence.

$$\min_{\Pi,z} \langle \Pi, C\rangle + \eta\langle \Pi, \ln \Pi\rangle + \tau\|\Pi\mathbf{1} - x\|_2^2 + \tau\|\Pi^\top \mathbf{1} - z\|_2^2 + \frac{\lambda}{2}\|w - z\|_2^2.$$

We can directly solve for Π and z in the following manner. In order to obtain Π, we minimize

$$\min_{\Pi} \langle \Pi, C\rangle + \eta\langle \Pi, \ln \Pi\rangle + \frac{\tau}{2}\|\Pi\mathbf{1} - x\|_2^2 + \frac{\tau}{2}\|\Pi^\top \mathbf{1} - z\|_2^2. \tag{36}$$

The derivative of Equation 36 with respect to Π is given by,

$$C + \tau\Pi\mathbf{1}\mathbf{1}^\top - \tau x\mathbf{1}^\top + \tau\mathbf{1}\mathbf{1}^\top\Pi - \tau\mathbf{1}z^\top + \eta\mathbf{1}\mathbf{1}^\top + \eta\ln\Pi. \tag{37}$$

Here, gradient descent can be performed to obtain Π. To obtain z, we minimize,

$$\frac{\tau}{2}\|\Pi^\top \mathbf{1} - z\|_2^2 + \frac{\lambda}{2}\|w - z\|_2^2. \tag{38}$$

Taking derivative with respect to z gives,

$$\Pi^\top \mathbf{1} + \lambda(w - z), \tag{39}$$

and equating to zero, we get,

$$z = w + \frac{\Pi^\top \mathbf{1}}{\lambda} \tag{40}$$

We can iteratively solve for Π and z using Equations 37 and 40 until convergence.

V. Experimental Results

A. Implementation Details

We implemented the proposed method in Pytorch framework. In case of MNIST and Fashion-MNIST, we use the architecture proposed in [4]. For CIFAR-10, we used the Resnet-18 pre-trained model available at [53]. MNIST model obtains an accuracy of 98.7% on clean examples, Fashion-MNIST model achieves an accuracy of 90.13%, and, CIFAR-10 model obtains an accuracy of 94.7% on clean examples. We use the cost matrix obtained using l_p norm between (i, j) and (k, l) as $(|i - j|^p + |k - l|^p)^{1/p}$ co-ordinates. We use $\gamma = 0.5, \eta = 0.001, \tau = 0.01, \alpha = 0.02, p = 1$. The number of PGD steps and Sinkhorn iterations is set to 10 each. In case of adversarial training, we set PGD steps to 20 as there are no imperceptibility constraints. Since we use an unbalanced OT, the distance is not comparable with that of [4] or [5]. Thus, we normalize $B(\alpha, \beta)$ such that $\mathbf{1}^\top B(\alpha, \beta)\mathbf{1} = 1$. For convergence, we use the number of Sinkhorn iterations and $B \odot C \leq \epsilon = 0.2$. We use an SGD optimizer with a learning rate = 0.01, momentum = 0.9, and weight decay = 5×10^{-4} for adversarial training for MNIST and a learning rate = 3×10^{-4} for Fashion-MNIST. We run the model for 40 and 60 epochs respectively for MNIST and Fashion-MNIST.

We use ResNet-50 [54] for Tiny ImageNet [55]. We use Adam optimizer with a learning rate of 0.0001, 2 PGD steps and train for 50 epochs. The nominal accuracy is 61.6%. In case of adversarial training, we also apply adversarial training practices mentioned in [30], [31]. Due to longer running time per epoch, we limit the adversarial training for 50 epochs. Following [30], we add the regularizer,

$$\mathcal{L}_{KL} = KL(\Phi(x), \Phi(z)),$$

where, KL denotes KL divergence, Φ denotes ResNet-50 feature extractor, x denotes clean sample and z denotes adversarial sample. The overall loss function is given by,

$$\mathcal{L} = \lambda\mathcal{L}_{CE}(\Phi(x), y) + (1 - \lambda)\mathcal{L}_{KL},$$

where $\lambda = 0.9$ and $\mathcal{L}_{CE}$ is the cross-entropy loss. Since random perturbation of x leads to better adversarial training [31], we also perform $x = x + 0.01\mathcal{N}(0, I)$.

We compare the nominal accuracy (NA) and attack accuracy (AA) of [4], [5] and ours. NA is computed over clean test samples and AA is computed over adversarially attacked test samples. NA is the ratio of correctly classified clean test samples to the number of total clean test samples. AA is the ratio of correctly classified adversarial test samples to the number of total adversarial test samples.

B. Adversarial attack performance

In Table I we compare the nominal accuracy (NA) and attack accuracy. It is clearly evident that our attack degrades the performance of all the models including standard and adversarially trained ones. For MNIST and CIFAR-10, we use both standard as well as adversarial trained models provided by Wong *et al.* and Hu *et al.*.

TABLE I: Nominal and attack accuracy against standard model

Method	MNIST		CIFAR-10		Fashion-MNIST	
	NA	AA	NA	AA	NA	AA
Wong *et al.* [4]	97.28	0.0	81.68	0.67	-	-
Hu *et al.* [5]	92.77	4.66	84.38	0.0	-	-
Standard	98.89	1.7	94.76	0.0	90.13	0.0

In case of Tiny ImageNet, the attack accuracy is 3.35%. We also show the examples of clean and attacked images in Figure 2. In addition, our model only takes 3.2 minutes compared to 43 minutes of Hu *et al.* to completely break the MNIST dataset on a Tesla V100 16GB machine.

We compare our proposed attack against state-of-art attacks in Table II. We can observe that our attack performs well in both the datasets.

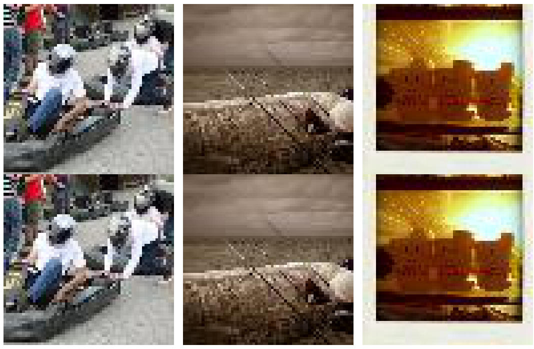

Fig. 2: Original (top row) and attacked (bottom row) images from Tiny ImageNet

TABLE II: Performance of our attack on standard models of MNIST and Fashion-MNIST.

Attack	MNIST	F-MNIST
Wong *et al.*	0.0	-
Hu *et al.*	4.66	-
PGDL2 [20]	37.93	27.24
CW [21]	4.25	10.0
TPGD [30]	5.01	11.95
FFGSM [31]	26.64	12.19
MIFGSM [27]	3.16	10.37
AutoAttack [26]	54.2	25.75
M-DI2FGSM [33]	9.92	17.14
Ours	1.7	0.0

C. Sensitivity Analysis

We show the sensitivity analysis in Figure 3. The image enclosed in red box is the original image. The last three images of first row highlight the effect of increasing τ while keeping η and γ fixed. As observed, changing $\tau \in [.01, 1]$ does not have a significant impact. Second row shows the effect of increasing η. Here, larger value of $\eta = 0.01$ leads to large distortion and the attacked image is severely distorted. Whereas, smaller values generate images closer to original image. Last row demonstrates the effect of γ. Lower value of $\gamma = 0.005$ leads to blurring and reduction in brightness. However, higher values lead to imperceptible distortion. Thus, we see that the attack is not highly sensitive to these parameters. Further, we use the same parameter setting for all the adversarial attack experiments.

D. Adversarial Robustness

In Figure 6, we show the adversarial accuracy vs. 1-Wasserstein distance. Here we compare L-inf robust model, standard model and our adversarially trained model against the attack proposed in Hu *et al.*. We can observe that our model outperforms the L-inf robust model and the margin increases as the distance increases. Thus, we can claim that our model demonstrates a high robustness against Wasserstein space attack. The standard MNIST model performs very poor and the accuracy against adversarial samples drops drastically to about 0% within a distance of 0.5. In case of Fashion-MNIST in Figure 7, we observe a similar phenomenon and our model exhibits much higher robustness compared to standard model.

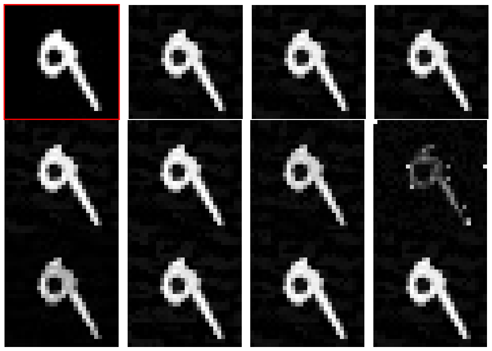

Fig. 3: Original image. First row shows the effect of τ = .01, .1, 1 with η = 0.001, γ = 0.5. Second row is obtained with η=0.0001, .0005, .005, .01 and τ = 0.01, γ = 0.5. Last row is obtained with γ=.005, .05, 5, 50 and τ = 0.01, η = 0.001.

In Table III we show a comparison of the defense against several attacks. We use the code provided in [56] to perform the attacks. In order to compare with L_∞ robust model and adversarial robust models of [4], [5], we perform adversarial attacks with $\epsilon \leq .3$. It can be observed that our model is highly robust against various types of attacks when compared to adversarial robust models of L_∞ and [4], [5]. This is due to the fact that the proposed attack is very strong and thus when the model is trained using adversarial examples, the boundaries learnt resist the attacks robustly. On the other hand, the standard model does not have any resilience against them. We observe a similar phenomenon in case of Fashion-MNIST in Table IV, however, here the model did not perform well against [21]. In case of Tiny ImageNet, the adversarial trained model accuracy is 25.15% for adversarial samples and 61.5% for clean samples. Here, we did not observe a significant drop in the accuracy against clean samples. This is because of high weightage for the loss computed against clean samples compared to the loss for adversarial samples. On the other hand, a small λ may not lead to convergence.

In Table V we present results for natural perturbations like translation and rotation. Here we observe that the standard model performs better compared to other models. While the models are robust to smaller values of translation, higher values lead to severe drop in the accuracy. This observation is consistent with the observations presented by Hu *et al.* [5]. The models trained with samples generated using Wasserstein attack do not exhibit good robustness against natural perturbations like rotation and translation. In order to compare the clean and affine transformed images, we perform Procrustes analysis. We present the results in Table VI. Procrustes analysis shows that the original and transformed images admit a high similarity. Further, with increasing degree of rotation and translation, the dis-similarity increases. This also leads to reduction in classification performance. We can also note that translation has higher dis-similarity and the mis-classification rate is higher in this case. Additionally, we also include Wasserstein distance between the samples and the Procrustes transformed samples. We would like to emphasize that our model takes a global cost matrix, whereas, [4], [5] take

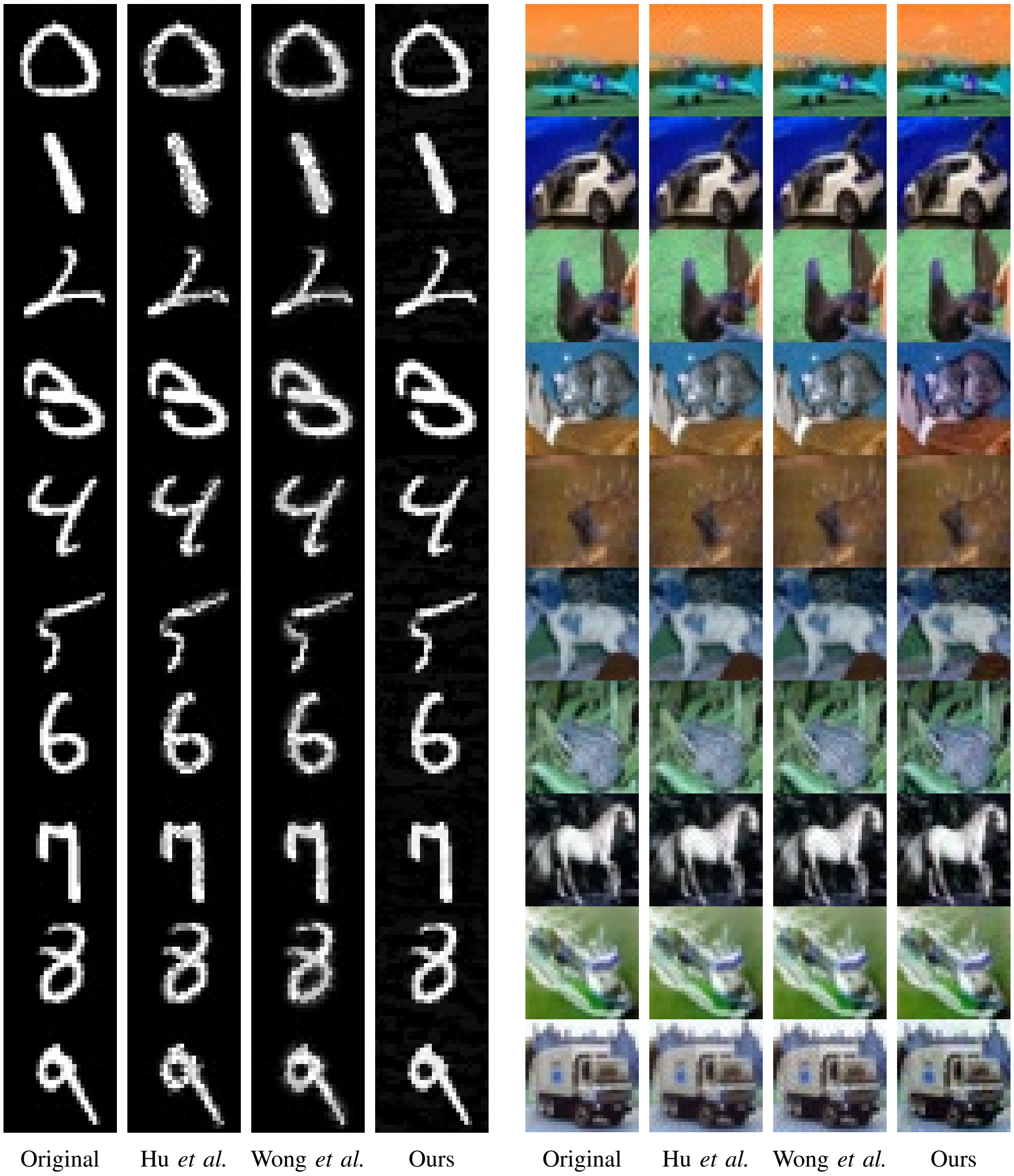

Fig. 4: Original and successful adversarial examples for MNIST and CIFAR-10.

a local cost matrix due to which the distances are distinct.

TABLE III: Performance of our adversarial trained model against adversarial attacks on MNIST.

Attack	Standard	L∞	[4]	[5]	Ours
Before	98.89	98.2	97.28	92.77	98.01
CW [21]	6.2	88.58	94.49	87.49	93.68
TPGD [30]	5.21	39.26	9.57	7.65	96.66
MIFGSM [27]	0.00	0.02	0.00	0.7	90.24
AutoAttack [26]	0.00	0.00	0.00	0.08	85.65
M-DI2FGSM [33]	0.12	0.07	0.03	1.09	91.85

E. Subjective Evaluation

We show the example images obtained by our algorithm in Figure 4 and analyze the magnified images in Figure 5. We observe that attack via Hu *et al.* generates more artefacts while Wong *et al.*' attack reduces the brightness. On the other hand, our attack appears to be the closest to original image.

TABLE IV: Performance of our adversarial trained model against adversarial attacks on Fashion-MNIST.

Attack	Standard	Ours
PGDL2 [20]	27.24	52.2
CW [21]	10.0	10.0
TPGD [30]	11.95	45.41
FFGSM [31]	12.19	46.66
MIFGSM [27]	10.37	46.35
DeepFool [32]	24.0	35.41
AutoAttack [26]	25.75	50.22
M-DI2FGSM [33]	17.14	48.71

We show the example original, adversarial and error images obtained by our algorithm in Figure 8. The first row shows the clean samples, middle row shows adversarial samples and the third row shows the error between clean and adversarial samples. It is evident that the distortion introduced due to attack is imperceptible in adversarial images. Here, we ob-

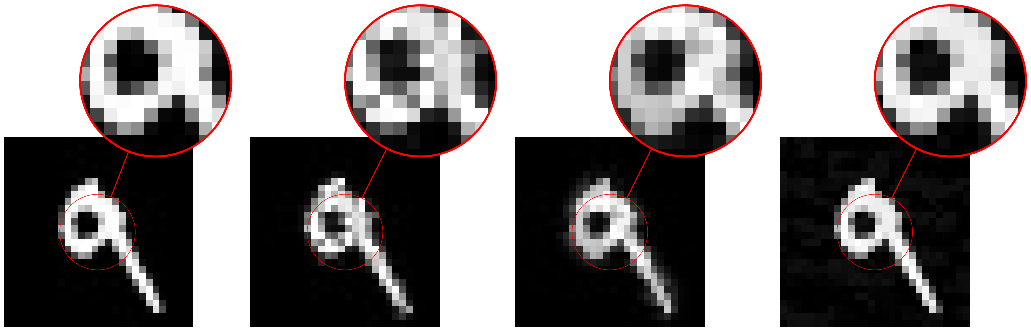

Fig. 5: Magnified images. Original, Hu *et al.*, Wong *et al.* and Ours

TABLE V: Accuracy against natural perturbation

Method	MNIST						Fashion-MNIST					
	Translation			Rotation			Translation			Rotation		
	5%	15%	20%	10°	20°	30°	5%	15%	20%	10°	20°	30°
Wong *et al.* [4]	95.47	82.88	45.83	96.61	95.14	92.09	-	-	-	-	-	-
Hu *et al.* [5]	91.15	82.37	50.018	91.7	89.03	84.64	-	-	-	-	-	-
Standard	97.75	87.03	47.85	98.48	97.39	93.67	84.44	68.41	39.57	84.44	70.32	57.76
Ours	96.09	82.36	42.8	97.51	95.95	92.74	84.15	66.99	38.36	76.15	68.3	58.03

TABLE VI: Procrustes dis-similarity (d) measure for rotation and translation.

Transformation		MNIST					F-MNIST				
		d	Ours	[4]	[5]	AA	d	Ours	[4]	[5]	AA
R	10°	0.005	0.060	0.015	0.011	97.51	0.004	0.023	0.014	0.011	76.15
	20°	0.011	0.105	0.080	0.053	95.95	0.007	0.042	0.032	0.028	68.3
	30°	0.016	0.144	0.137	0.118	92.74	0.01	0.060	0.053	0.044	58.03
T	5%	0.016	0.109	0.068	0.059	96.09	0.021	0.062	0.028	0.027	84.14
	15%	0.094	0.360	0.091	0.087	82.36	0.101	0.184	0.118	0.109	66.99
	20%	0.155	0.497	0.161	0.150	42.8	0.150	0.248	0.169	0.145	38.36

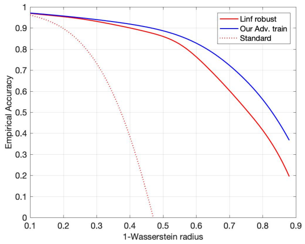

Fig. 6: MNIST

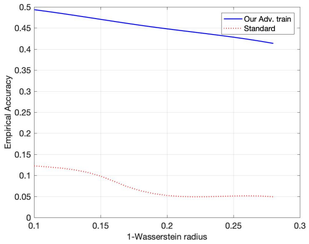

Fig. 7: F-MNIST

serve that attack looks like distortion in shape of localized patches. This is because of inherent nature of attack which is performed by pixel mass movement. The major movement happens locally as the cost of moving to large distance is very high. Thus most of the large distance movement is restricted. Further, the maximum values of error occurs in foreground only. This may be due to the fact that the foreground is much brighter compared to background. A similar phenomenon can be observed in Fashion-MNIST dataset also.

F. Targeted Attack

Targeted attacks can also be performed by presenting the target labels to our model. In Figure 9, we see that for different target labels, the attacked images look similar to the clean image itself. Thus, there are no visible artefacts. In our experiments, we observe that targeted attacks require more iterations compared to untargeted attacks which may also lead to slightly more noise.

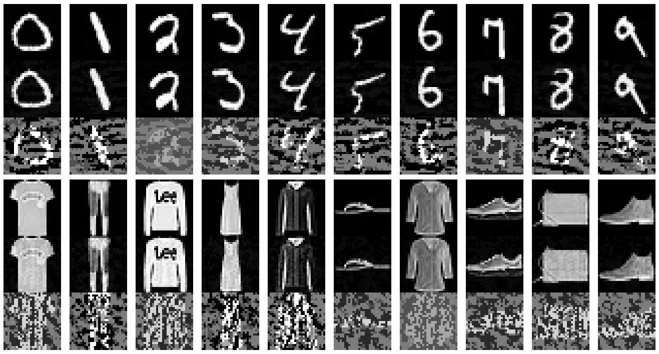

Fig. 8: Clean image, adversarial image and error between clean and adversarial image for MNIST and Fashion-MNIST.

VI. Conclusion

In this paper, we present a novel technique for performing adversarial attack on images. We formulate the objective with the goal to obtain an adversarial sample which is closer to the sample obtained from gradient step in l_2 sense and lies in a Wasserstein ball with respect to the original sample. Towards this, we propose a combined objective function using an unbalanced optimal transport technique. We analytically solve

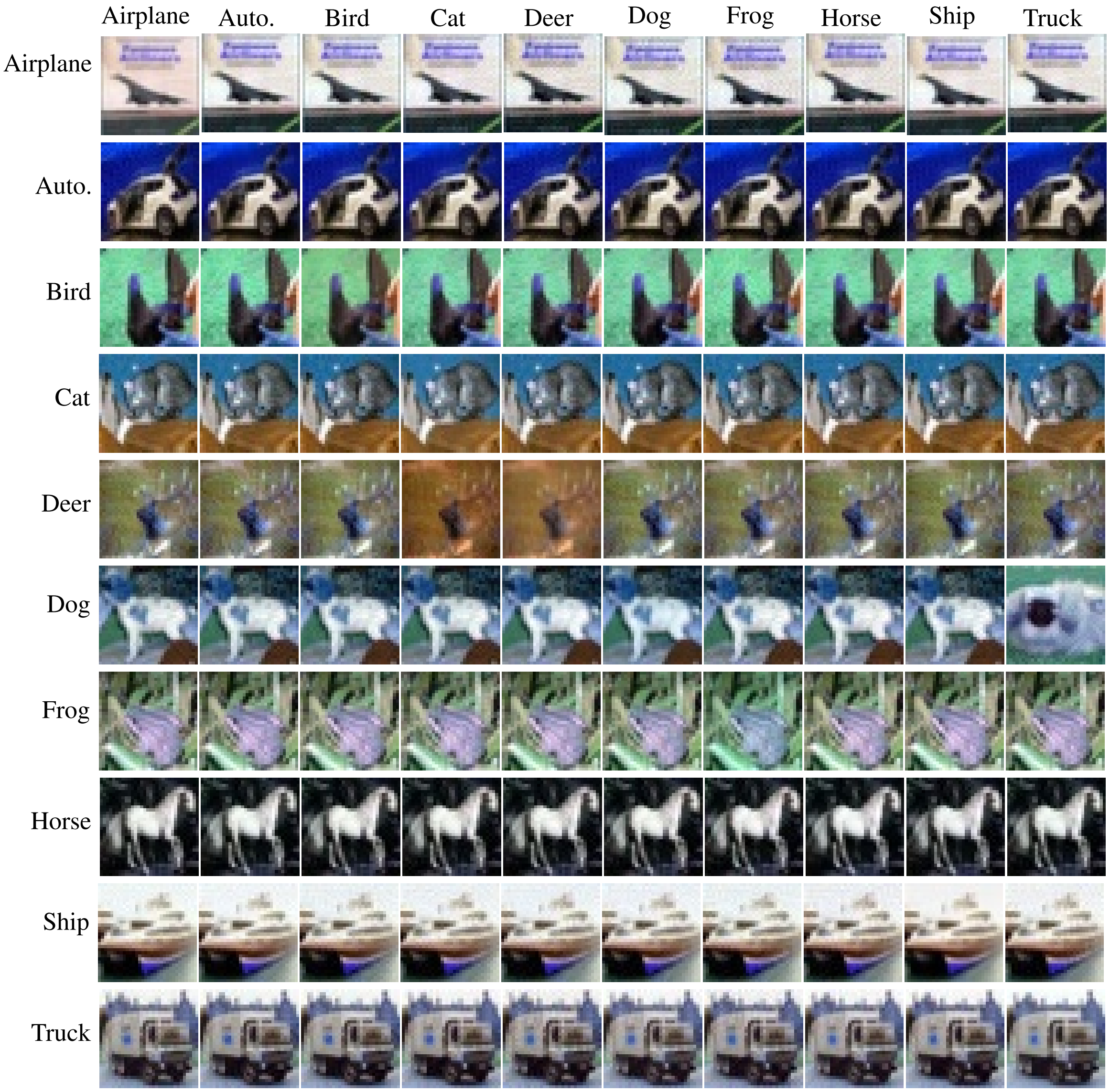

Fig. 9: Targeted attack.

for the adversarial sample in the dual domain. Our method demonstrates that it can easily defeat the standard as well as adversarial robust models against Wasserstein space attacks. Further, our adversarial defense also proves significantly better than other robust models against different set of attacks. We also show that the adversarial images generated by our model does not suffer from artefacts as in case of other Wasserstein space attacks.

References

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[2] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[3] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," *arXiv preprint arXiv:1611.01236*, 2016.

[4] E. Wong, F. Schmidt, and Z. Kolter, "Wasserstein adversarial examples via projected sinkhorn iterations," in *ICML*. PMLR, 2019, pp. 6808–6817.

[5] J. E. Hu, A. Swaminathan, H. Salman, and G. Yang, "Improved image wasserstein attacks and defenses," in *ICLR 2020*, 2020.

[6] K. Wu, A. Wang, and Y. Yu, "Stronger and faster wasserstein adversarial attacks," in *ICML*. PMLR, 2020, pp. 10 377–10 387.

[7] J. Li, J. Cao, S. Zhang, Y. Xu, J. Chen, and M. Tan, "Internal wasserstein distance for adversarial attack and defense," *arXiv preprint arXiv:2103.07598*, 2021.

[8] O. Pele and M. Werman, "A linear time histogram metric for improved sift matching," *ECCV*, pp. 495–508, 2008.

[9] A. Figalli, "The optimal partial transport problem," *Arch. Rational Mech. Anal.*, vol. 195, no. 2, pp. 533–560, 2010.

[10] K. Pham, K. Le, N. Ho, T. Pham, and H. Bui, "On unbalanced optimal transport: An analysis of sinkhorn algorithm," in *ICML*. PMLR, 2020, pp. 7673–7682.

[11] J. Lee, N. Bertrand, and C. Rozell, "Unbalanced optimal transport regularization for imaging problems," *IEEE Trans. on Comp. Imag.*, vol. 6, pp. 1219–1232, 2020.

[12] K. Fatras, T. Sejourne, R. Flamary, and N. Courty, "Unbalanced minibatch optimal transport; applications to domain adaptation," *ICML*, pp. 3186–3197, 2021.

[13] C. Frogner, C. Zhang, H. Mobahi, M. Araya-Polo, and T. A. Poggio, "Learning with a wasserstein loss," in *NIPS*, 2015.

[14] Z. Wang, D. Zhou, M. Yang, Y. Zhang, C. Rao, and H. Wu, "Robust document distance with wasserstein-fisher-rao metric," in *Asian Conference on Machine Learning*. PMLR, 2020, pp. 721–736.

[15] Z. Ma, X. Wei, X. Hong, H. Lin, Y. Qiu, and Y. Gong, "Learning to count via unbalanced optimal transport," in *Proceedings of the AAAI*

Conference on Artificial Intelligence, vol. 35, no. 3, 2021, pp. 2319–2327.

[16] Y. Chen, Y. Lan, R. Xiong, L. Pang, Z. Ma, and X. Cheng, "Evaluating natural language generation via unbalanced optimal transport." in *IJCAI*, 2020, pp. 3730–3736.

[17] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[18] K. Kokomoto, R. Okawa, K. Nakano, and K. Nozaki, "Intraoral image generation by progressive growing of generative adversarial network and evaluation of generated image quality by dentists," *Scientific reports*, vol. 11, no. 1, pp. 1–10, 2021.

[19] R. Sinkhorn, "Diagonal equivalence to matrices with prescribed row and column sums," *The American Mathematical Monthly*, vol. 74, no. 4, pp. 402–405, 1967.

[20] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[21] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Symposium on security and privacy*, 2017, pp. 39–57.

[22] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *Symposium on security and privacy*, 2016, pp. 582–597.

[23] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *European symposium on security and privacy*, 2016, pp. 372–387.

[24] Z. Che, A. Borji, G. Zhai, S. Ling, J. Li, Y. Tian, G. Guo, and P. Le Callet, "Adversarial attack against deep saliency models powered by non-redundant priors," *IEEE Transactions on Image Processing*, vol. 30, pp. 1973–1988, 2021.

[25] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," in *International conference on machine learning*. PMLR, 2018, pp. 284–293.

[26] F. Croce and M. Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," in *ICML*. PMLR, 2020, pp. 2206–2216.

[27] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *CVPR*, 2018, pp. 9185–9193.

[28] J. Su, D. V. Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," *IEEE Transactions on Evolutionary Computation*, vol. 23, no. 5, pp. 828–841, 2019.

[29] R. Storn and K. Price, "Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces," *Journal of global optimization*, vol. 11, no. 4, pp. 341–359, 1997.

[30] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *ICML*. PMLR, 2019, pp. 7472–7482.

[31] E. Wong, L. Rice, and J. Z. Kolter, "Fast is better than free: Revisiting adversarial training," *arXiv preprint arXiv:2001.03994*, 2020.

[32] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," in *CVPR*, 2016, pp. 2574–2582.

[33] C. Xie, Z. Zhang, Y. Zhou, S. Bai, J. Wang, Z. Ren, and A. L. Yuille, "Improving transferability of adversarial examples with input diversity," in *CVPR*, 2019, pp. 2730–2739.

[34] A. Ross and F. Doshi-Velez, "Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients," in *AAAI*, vol. 32, no. 1, 2018.

[35] A. Nayebi and S. Ganguli, "Biologically inspired protection of deep networks from adversarial attacks," *arXiv preprint arXiv:1703.09202*, 2017.

[36] D. Krotov and J. Hopfield, "Dense associative memory is robust to adversarial inputs," *Neural computation*, vol. 30, no. 12, pp. 3151–3167, 2018.

[37] E. Wong and Z. Kolter, "Provable defenses against adversarial examples via the convex outer adversarial polytope," in *ICML*. PMLR, 2018, pp. 5286–5295.

[38] A. Mustafa, S. H. Khan, M. Hayat, J. Shen, and L. Shao, "Image super-resolution as a defense against adversarial attacks," *IEEE Transactions on Image Processing*, vol. 29, pp. 1711–1724, 2019.

[39] Z. Zhao, H. Wang, H. Sun, J. Yuan, Z. Huang, and Z. He, "Removing adversarial noise via low-rank completion of high-sensitivity points," *IEEE Transactions on Image Processing*, 2021.

[40] F. A. Potra and S. J. Wright, "Interior-point methods," *Journal of computational and applied mathematics*, vol. 124, no. 1-2, pp. 281–302, 2000.

[41] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," *NeurIPS*, vol. 26, pp. 2292–2300, 2013.

[42] A. Levine and S. Feizi, "Wasserstein smoothing: Certified robustness against wasserstein adversarial attacks," in *AISTATS*. PMLR, 2020, pp. 3938–3947.

[43] H. Zhang and J. Wang, "Defense against adversarial attacks using feature scattering-based adversarial training," *arXiv preprint arXiv:1907.10764*, 2019.

[44] L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard, "Scaling algorithms for unbalanced optimal transport problems," *Mathematics of Computation*, vol. 87, no. 314, pp. 2563–2609, 2018.

[45] T. Séjourné, J. Feydy, F.-X. Vialard, A. Trouvé, and G. Peyré, "Sinkhorn divergences for unbalanced optimal transport," *arXiv preprint arXiv:1910.12958*, 2019.

[46] G. Peyré, M. Cuturi *et al.*, "Computational optimal transport: With applications to data science," *Foundations and Trends® in Machine Learning*, vol. 11, no. 5-6, pp. 355–607, 2019.

[47] J.-D. Benamou, "Numerical resolution of an "unbalanced" mass transport problem," *ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique*, vol. 37, no. 5, pp. 851–868, 2003.

[48] I. Csiszár, "Axiomatic characterizations of information measures," *Entropy*, vol. 10, no. 3, pp. 261–273, 2008.

[49] M. Blondel, V. Seguy, and A. Rolet, "Smooth and sparse optimal transport," in *AISTATS*. PMLR, 2018, pp. 880–889.

[50] J. Lee, N. P. Bertrand, and C. J. Rozell, "Unbalanced optimal transport regularization for imaging problems," *IEEE Transactions on Computational Imaging*, vol. 6, pp. 1219–1232, 2020.

[51] P. L. Bartlett and S. Mendelson, "Rademacher and gaussian complexities: Risk bounds and structural results," *JMLR*, vol. 3, no. Nov, pp. 463–482, 2002.

[52] A. Genevay, L. Chizat, F. Bach, M. Cuturi, and G. Peyré, "Sample complexity of sinkhorn divergences," in *AISTATS*. PMLR, 2019, pp. 1574–1583.

[53] K. Liu, "Pytorch-cifar," https://github.com/kuangliu/pytorch-cifar.

[54] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[55] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[56] H. Kim, "Torchattacks: A pytorch repository for adversarial attacks," *arXiv preprint arXiv:2010.01950*, 2020.

Appendix A
Parameters of adversarial attacks

We provide the parameter settings for various attacks in Table VII. For CW [21] attack, $c = 5$ and $\kappa = 0$ is used. Other parameters for all the attacks are set to default. For AutoAttack [26] we use L2 norm.

TABLE VII: Parameter settings

Attack	ϵ	α	Steps
PGDL2[20]	0.2	0.2	400
TPGD [30]	0.2	0.2	10
FFGSM [31]	0.2	0.2	-
MIFGSM [27]	0.2	0.2	10
DeepFool [32]	-	-	2
AutoAttack [26]	0.2	-	-
M-DI2FGSM [33]	0.2	0.2	2

Appendix B
Proofs

1) Proof of Lemma 1:: Let (α^*, β^*) be the optimal solution of Equation 7. Then, the sup norms of the optimal solution

$\|\alpha^*\|_\infty$, $\|\beta^*\|_\infty$ and $\|z\|_\infty$ are bounded by,

$$\begin{aligned}\max\{\|\alpha^*\|_\infty, \|\beta^*\|_\infty\} &\le \tau R \qquad (41)\\ \|z\|_\infty &\le \|w + \tau/\gamma e^{-\min(\beta^*/\tau)}\|_\infty\\ &= \|w\|_\infty + \tau/\gamma e^{\|\beta^*\|_\infty/\tau},\end{aligned}$$

where
$R = \max(\|\ln(x)\|_\infty, \|\ln(z)\|_\infty) +$
$\max(\ln(n), \frac{1}{\eta}\|C\|_\infty - \ln(\eta))$.

$$\frac{\alpha_i^*}{\tau} = \ln x_i - \ln\left(\sum_{j=1}^{n} e^{\frac{\alpha_i^*+\beta_j^*-C_{ij}}{\eta}}\right), \qquad (42)$$

$$\begin{aligned}c &= \ln x_i - \ln\left(e^{\frac{\alpha_i^*}{\eta}} \sum_{j=1}^{n} e^{\frac{\beta_j^*-C_{ij}}{\eta}}\right), \qquad (43)\\ &= \ln x_i - \frac{\alpha_i^*}{\eta} - \ln\left(\sum_{j=1}^{n} e^{\frac{\beta_j^*-C_{ij}}{\eta}}\right).\end{aligned}$$

The third term of RHS in Equation 42 can be written as,

$$\begin{aligned}\ln\left(\sum_{j=1}^{n} e^{\frac{\beta_j^*-C_{ij}}{\eta}}\right) &\ge \ln\left(\sum_{j=1}^{n} \min_j e^{\frac{\beta_j^*-C_{ij}}{\eta}}\right), \qquad (44)\\ &\ge \ln(n) + \min_j \frac{\beta_j^* - C_{ij}}{\eta},\\ &\ge \ln(n) - \frac{\|\beta^*\|_\infty}{\eta} - \frac{\|C\|_\infty}{\eta}.\end{aligned}$$

Similarly, we can show that,

$$\ln\left(\sum_{j=1}^{n} e^{\frac{\beta_j^*-C_{ij}}{\eta}}\right) \le \ln(n) + \frac{\|\beta^*\|_\infty}{\eta}. \qquad (45)$$

Using Equations 44 and 45, we obtain,

$$\left|\ln\left(\sum_{j=1}^{n} e^{\frac{\beta_j^*-C_{ij}}{\eta}}\right)\right| \le \frac{\|\beta^*\|_\infty}{\eta} + \max\left(\ln(n), \frac{\|C\|_\infty}{\eta} - \ln(n)\right). \qquad (46)$$

Now, from Equation 42, we can write,

$$\begin{aligned}\left|\frac{\alpha_i^*}{\tau} + \frac{\alpha_i^*}{\eta}\right| &\le |\ln(x_i)| + \left|\ln\left(\sum_{j=1}^{n} e^{\frac{\beta_j^*-C_{ij}}{\eta}}\right)\right| \qquad (47)\\ &\le |\ln(x_i)| +\\ &\frac{\|\beta^*\|_\infty}{\eta} + \max\left(\ln(n), \frac{\|C\|_\infty}{\eta} - \ln(n)\right).\end{aligned}$$

Using the fact that $\ln(x_i) \le \max\{\ln(\|x\|_\infty), \ln(\|z\|_\infty)\}$, we obtain,

$$\begin{aligned}\left|\frac{\alpha_i^*}{\tau} + \frac{\alpha_i^*}{\eta}\right| &\le \max\{\ln(\|x\|_\infty), \ln(\|z\|_\infty)\} + \qquad (48)\\ &\frac{\|\beta^*\|_\infty}{\eta} + \max\left(\ln(n), \frac{\|C\|_\infty}{\eta} - \ln(n)\right).\end{aligned}$$

Now taking $\|\alpha^*\|_\infty \ge \|\beta^*\|_\infty$ without loss of generality, we get,

$$\begin{aligned}\frac{\|\alpha^*\|_\infty}{\tau} &\le \max\{\ln(\|x\|_\infty), \ln(\|z\|_\infty)\} + \qquad (49)\\ &\max\left(\ln(n), \frac{\|C\|_\infty}{\eta} - \ln(n)\right)\\ &= R.\end{aligned}$$

which proves Lemma 1.

2) Proof of Lemma 2:: Updates $(\alpha^{k+1}, \beta^{k+1})$ from Algorithm 1 satisfies the following bound

$$\delta^{k+1} \le \delta^k, \qquad (50)$$

where, $\delta^{k+1} = \left(\frac{\tau}{\tau+\eta}\right)^k \|\beta^*\|_\infty$. The proof can be directly extended from [10].

$$\begin{aligned}\alpha_i^{k+1} &= \left[\frac{\alpha_i^k}{\eta} + \ln(x_i) - \ln(r_i^k)\right]\frac{\eta\tau}{\eta+\tau}, \qquad (51)\\ &= \left[\frac{\alpha_i^k}{\eta} + \ln(x_i) - \ln(r_i^*) + \{\ln(r_i^*) - \ln(r_i^k)\}\right]\frac{\eta\tau}{\eta+\tau}. \qquad (52)\end{aligned}$$

This can be written as,

$$\alpha_i^{k+1} - \alpha_i^* = \left[\eta \ln\left(\frac{r_i^*}{r_i^k}\right) - (\alpha_i^* - \alpha_i^k)\right]\frac{\tau}{\tau+\eta}. \qquad (53)$$

From [10], we know that,

$$|\alpha_i^{k+1} - \alpha_i^*| \le \max_j |\beta_j^k - \beta_j^*| \frac{\tau}{\tau+\eta}. \qquad (54)$$

Thus,

$$\|\alpha^{k+1} - \alpha^*\|_\infty \le \frac{\tau}{\tau+\eta}\|\beta^k - \beta^*\|_\infty. \qquad (55)$$

Similarly,

$$\|\beta^{k+1} - \beta^*\|_\infty \le \frac{\tau}{\tau+\eta}\|\alpha^k - \alpha^*\|_\infty. \qquad (56)$$

Combining equations 55 and 56,

$$\|\alpha^{k+1} - \alpha^*\|_\infty \le \left(\frac{\tau}{\tau+\eta}\right)^2 \|\alpha^{k-1} - \alpha^*\|_\infty. \qquad (57)$$

By induction,

$$\begin{aligned}\|\alpha^{k+1} - \alpha^*\|_\infty &\le \left(\frac{\tau}{\tau+\eta}\right)^k \|\beta^0 - \beta^*\|_\infty, \qquad (58)\\ &= \left(\frac{\tau}{\tau+\eta}\right)^k \|\beta^*\|_\infty \qquad (59)\end{aligned}$$

Similarly, we can show for $\|\beta^{k+1} - \beta^*\|_\infty$. This proves Lemma 2.