We thank all the reviewers for recognizing the contribution of our work and providing their valuable comments. We address the questions below and will release the trained model together with code as suggested by reviewers to contribute to the community.

Shared comment on gradient-masking. We first address this by conducting analysis following reviewers' suggestions from four complementary perspectives: §1 black-box attack, §2 stronger white-box attack, §3 loss variations and §4 gradient-free attack. The results have verified that *the improved robustness is indeed due to model improvement instead of gradient masking*.

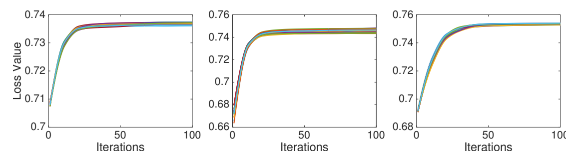
§1 Results under black-box attacks. For the black-box attacks, the results for PDG100, CW20 and CW100 in the original submission are incorrect due to our own fault: *they are results under white-box attacks*. We spot this error after submission deadline but were unable to upload the correct version. *We sincerely apologize for this misleading mistake and any additional efforts required from all the reviewers because of it*. The correct black-box results are shown in the table above. This black-box results, together with the white-box results (Table 1 in the main paper and §2 below) suggest that gradient masking is not present in the model and the improved robustness is indeed due to the inherent improvement of the model itself.

B-Attack	PDG20	PDG100	CW20	CW100
`Undefended`	89.0	88.7	88.9	88.8
`Siamese`	81.6	81.0	80.3	79.8

§2 Results against stronger white-box attacks. We have used random starts for all the white-box evaluations here as well as in the main paper. Here we run the evaluations 5 times and report the lowest performance and ($\max$–$\min$) as in the table on the right. The accuracy of our model under very strong white-box attacks still outperforms Madry model (44.8/45.4) by a large margin.

Attack Iteration	PGD500	PGD1000	CW500	CW1000
min over 5 runs	66.8	66.4	59.0	58.8
max–min	0.3	0.3	0.3	0.2

§3 Loss plots. The loss achieved with PGD adversary against our model increases in a fairly consistent way and plateaus rapidly for 20 different runs with random starts, with relatively small variance (right).



§4 Results against gradient-free attacks. We evaluated our model using a gradient-free black-box attack based on greedy local search,[1] and result is 89.9, further assuring the absence of gradient masking.

To Reviewer 1:

Clarity and miscellaneous comments. ***i)*** Your understanding of Alg 1 is correct. We will add more explanations on this for clarity. ***ii)*** $\boldsymbol{\mu}=\{u_i\}$ represents the "empirical probability vector" over the support of $\{\mathbf{x}_i\}$. They are set to be uniform as shown in the Sinkhorn and IPOT algorithms in supplementary file, in the absence of prior knowledge. We have used this representation to make the presentation general enough to incorporate prior knowledge when available. We will make this clear and mention the Sinkhorn and IPOT sooner in revision. ***iii)*** Label smoothing parameter 0.5 is set without hyper-parameter tuning. The investigation on smoothing parameter was done afterwards. We didn't aim to achieve the best performance by hyper-parameter search but instead to show the general applicability of our approach under a broad range of hyper-parameter settings. ***iv)*** The gap between PGD and the CW-variant is smaller under stronger attacks (§2) and we attribute the remaining gap to the nature of our model, where a one step unsupervised adversary is used for training, different from the multi-step supervised adversary typically used in Madry model. ***v)*** "Emphasizing 1 attack iteration earlier" is a great suggestion. We will update this to avoid confusions as happened to **Reviewer 2**.

Disentangling of distance and coupling. Thanks for the great suggestion. We have preliminarily investigated the disentangling of distance and inter-sample coupling in our main paper as you have already noticed in Sec. 5.2 using the `identity matching`. Further investigation on it (esp. the coupling) as suggested by the reviewer is interesting and we plan to work on it as our next steps.

To Reviewer 2:

Computational concern. The number of iterations $T=1$ is used for our model as mentioned in line 225. We apologize for the confusion and will make it more clear as also suggested by **Reviewer 1**. Given that T is typically set to 7 in conventional PGD adversarial training (e.g. Madry), our approach does not take advantage of extra computation compared to conventional PGD training.

Random targeted baseline. We have experimented with random-targeted adversarial training as suggested by the reviewer. It achieves accuracy of 49.9/48.5 under PGD100/CW100 (`min` over 5 runs), outperformed by our model with a large margin (§2).

Understanding of feature-scattering. Conventional adversarial examples are *decision boundary oriented* (Fig.2), making the effective manifold for training deviate from the original due to *tilting* and *shrinking*, hindering performance (line 36-52), with label leaking as one manifesting phenomenon. Feature scattering is *inter-sample structure* oriented and promotes data diversity without drastically altering the structure of the manifold. We plan to conduct rigorous theoretical analysis of the proposed model as next step.

To Reviewer 3:

Batch size. As shown in the table on the right, larger batch size leads to better performance as it facilitates feature matching. Batch size of 60 is used for our model in the paper. Batch sizes larger than 60 lead to similar results. This observation is similar to other applications with embedded OT matching such as OT-GAN [48].

batch size	40	50	60	70	80
PDG100	58.7	62.7	68.4	68.2	68.4

Label smoothing. Label smoothing is necessary for our model. Our model (with 1-step adversary) achieves compromised results without it, compared to standard PGD adversarial training (e.g. Madry) with 7-steps adversaries. This is an expected result as feature scattering makes the feature distributions more diffused (see Fig 1 in supplementary file), thus the corresponding label should ideally be "diffused" as well, in a spirit similar to *mixup* [70], which is achieved with label smoothing approximately in this work. Better schemes for joint treatment of feature and label scattering is an interesting topic and is left as our future work.

smooth para.	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
Madry	44.8	46.4	46.7	46.1	46.2	46.1	45.6	46.8	47.4
Ours	35.3	56.2	59.2	61.8	58.9	68.4	70.1	71.2	69.7

Choice of distance. Using cosine distance avoids introducing additional tuning parameter as the features are normalized before computing the distance. This and the usage of logits are just design choices. Other distance measures and intermediate features can be used together with our framework as well. We will explain this in the updated paper. As suggested by the reviewer, we will also introduce label leaking earlier in the introduction for clarity. We will release our trained model together with code as suggested.

[1] N. Narodytska and S. Kasiviswanathan. Simple black-box adversarial attacks on deep neural networks, CVPRW17