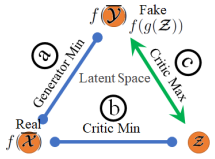
General comments: We thank all the reviewers for their insightful comments, and their unanimous positive comments on the superior performance and high quality of our work. Our novelty has also been affirmed by R1, R2 and R4. We believe that R3 might have misunderstood our work to be an extension of MMD-GANs (e.g., Refs [3,4] suggested by R3). However, we should clarify that (1) our work differs completely from MMD-GANs, and (2) although Ref [4] remains unpublished before the submission deadline, our work also achieves superior performances over Refs [3,4]. We are confident that our work is novel because *our critic operates as semantic embeddings and learns a fruitful latent space*, instead of being a component to build complete metrics as the existing GANs (e.g., MMD-GANs and W-GANs) do. Theoretically, this avoids the Lipschitz constraint but it requires a reciprocal between the generator and critic, where the *auto-encoder is a necessity* instead of just a plug-in. Our CF design (even given the unpublished work), is novel in its • *triangle anchor design with* l_1*-norm* (to stabilise convergence), • *meaningful analysis of amplitude and phase* (to favour other distribution alignment tasks), • t*-net of outputting scales* (to optimise $\mathcal{T}$ distribution types), and • *useful theory*, e.g., Prop. 1 for ill-posed optimisation (a potential problem in Ref [3]) and Lemma 2 for reducing sets. Thus, our RCF-GAN *seamlessly combines the auto-encoder and GANs by using only two neat modules while achieving s.o.t.a. generation and reconstruction*, whereas MMD-GANs (and many others) typically use three modules, but reconstruct and interpolate blurred images. Our supplementary material includes the *s.o.t.a.* results under adv-DCGAN and ResNet structures on more datasets. Below we discuss the reviewers' comments and will address all of them in the revision.

R1: [Expressivity]: This is indeed an interesting topic. Although having proved the effectiveness in our RCF-GAN, the unimodal setting may limit the expressivity on complex datasets and a worthy investigation could be using mixture models (or learning models) to further improve the performance. **[t-net]**: Yes, exactly. In our code (to be released upon acceptance) we put both $f(\cdot)$ and t-net in the critic, as they are optimised simultaneously. As t-net is optional (we can directly use Gaussian samples), we separate it from $f(\cdot)$ but it belongs to the critic; this will be clarified in the revision.

R2: [Q0]: (1) Fig.2-b shows some results on MNIST by directly training $g(\mathbf{z})$. For other datasets, solely training $g(\mathbf{z})$ via the CF typically performs inferior, which in our preliminary trials on CelebA, obtained a 165 FID score (rough faces). This also verifies the benefit of latent space comparison via our critic. Tailored t-net could be of help and we will provide more results in the revision. (2) Yes, other metrics can be seamlessly adopted in our framework. However, we found that our CF loss works better in practice because our t-net automatically optimises $\mathcal{T}$ distribution types. Our CF loss also allows for more complex distributions (e.g., mixture models) to further improve model expressibility. **[Q1]**: (1) The anchor refers to $\mathcal{Z}$ and the critic loss (minimising) is $-(C_{\mathcal{T}}(f(\overline{\mathcal{Y}}),\mathcal{Z})-C_{\mathcal{T}}(f(\overline{\mathcal{X}}),\mathcal{Z}))$. In the dynamic training process, $\mathcal{Z}$ provides a pivot for the critic; please see the figure. This way, the critic can quickly map real data $\overline{\mathcal{X}}$ to the support of $\mathcal{Z}$ via ⓑ (as Lemma 4 requires). (2) Lemma 4 ensures $\mathbb{E}_{\mathcal{Z}}[||\mathbf{z}-f(g(\mathbf{z}))||_2^2]=0 \leftrightarrow \mathbb{E}_{\overline{\mathcal{Y}}}[||\overline{\mathbf{y}}-f(g(\overline{\mathbf{y}}))||_2^2]=0$, i.e., the equivalence between reconstructing in the latent space (on generated images) and the pixel domain (on real images). (3) As GANs, we adversarially train $f(\overline{\mathcal{Y}})$ away from $\mathcal{Z}$ by ⓒ. **[Q2]**: $C_{\mathcal{T}}(\cdot,\cdot)$ is then not a valid metric for the degenerated $\mathcal{T}$ since its support is not $\mathbb{R}^m$ as Lemma 1 requires (may have $C_{\mathcal{T}}(\mathcal{X},\mathcal{Y})=0$ but $\mathcal{X}\neq\mathcal{Y}$). **[Q3]**: We apologise for the typos as pointed out by the reviewer. For the second question, the reviewer might have missed the need of identical supports on f and g. If $\mathcal{Z}\in\mathbb{R}^m$, the support of $\mathcal{Y}=(\mathcal{Z},-1)$ on g is $\mathbb{R}^m$, which does not equal to the support of $(\mathcal{Z},b)$ on f (i.e., $\mathbb{R}^{m+1}$), leading to $g(f(\mathcal{Z},1))\neq(\mathcal{Z},1)$. **[Q4]**: $(\mathcal{X},\mathcal{Y})$ are for the random variables in the latent space and $(\overline{\mathcal{X}},\overline{\mathcal{Y}})$ for the pixel domain. **[Q8]**: In Fig.2-b, using the amplitude only results in generating "wrong" numbers ("1" for digit 4, "6" for digit 5), uneven characters, disconnected artefacts, etc, all of which can be relieved by adding the phase in training.



R3: [Novelty]: Please refer to lines 3-17 of this rebuttal. **[Ablation and anchor design]**: During the rebuttal period we did the ablation study on CelebA and report the FIDs for the original (15.9), $\lambda=0$ (59.3) and without anchor (17.5). Thus, • without reconstruction loss, the generation largely degrades, *validating the correctness and necessity of our reciprocal theory*; • our motivation of anchoring is to find a pivot (i.e., $\mathcal{Z}$) during dynamic training process, where our critic can quickly map real data to the support of $\mathcal{Z}$, as Lemma 4 requires. This is evidenced by the 1.6 FID gain. We will provide more results in the revision. **[Lipschitz]**: Similar to other GANs (e.g., Fisher GAN and Sphere GAN), the Lipschitz constraint is not a necessity in our RCF-GAN. Please refer to our proof. **[Consistency]**: The inconsistency on CelebA with [4] is because [4] rescaled images in ALL datasets to a small size 32×32. For CIFAR10, our result is consistent with [4], whereby both our and [4] are inconsistent with [3] (Table 2). Compared to [3,4], our RCF-GAN also achieved the best performances. **[Notation]**: t_{norm} uses fixed Gaussian and t_{net} learns the t-net for ablation study.

R4: [Model parameters]: Our critic and generator are evaluated under the almost SAME number of model parameters as W-GANs, whereas MMD-GANs need an extra decoder net. The only extra cost in our optional t-net is negligible because it is a 3-layer FC net with the dimension of each layer less than 128. **[Reconstruction and datasets]**: Fig.4 in the paper shows the image reconstruction and interpolation, validating our superior performances on *clear reconstructions and semantic interpolations*. Our supplementary material further validates our scalable and consistent superior performances on two extra structures (adv-DCGAN and ResNet) and one extra dataset (LSUN Church). We will also release our code upon acceptance. **[Phase and amplitude]**: As the amplitude weight measures the model diversity, the mode collapse can be efficiently relieved in our results. We will elaborate more upon this in the revision.