First of all, we would like to thank all the reviewers' valuable comments and their recognition (mainly from R #3 and R #5) on the contributions of our work in terms of strong motivation, technical originality, wide applicability to various clustering frameworks, and convincing experiments with a comprehensive ablation study. Our response is as follows, which mainly addresses the concerns and perhaps some potential misunderstandings of R #4. Our source code will be released and more details will be added by the 9th extra page in the revised version. We understand that the reviewers may have missed some critical points in such a short review period, while we do hope that R #4 could reconsider the contributions and novelty of our work based on our clarification.

# 1 To Reviewer 3

**Clarity**. Thanks for your questions. $\Phi$ and $\Psi$ denote the parameters of the attack network and the discriminator respectively, and $\boldsymbol{\delta}$ is the output of the attack network. Notation $T$ is the mapping function of the discriminator (line 128-130) constructed by the neural network, which is often used to discriminate the relevances of the inputs [11], and $\sigma$ represents the activation function of the discriminator. Negative sampling estimation [11] is adopted to optimize the mapping function $T$. We will improve the writing and release the source code.

# 2 To Reviewer 4

**Understanding Eq. 1 and Eq. 3**. These equations refer to the clustering model introduced in Section 2, which is a basic model of an existing method [30] used as a testbed to show how our proposed adversarial algorithm can attack the network and improve its robustness, and our technique can be widely applied to more existing clustering methods which are essentially attack-agnostic. The loss function is $\mathcal{L}_C = KL(p(\mathbf{x},\mathbf{z},\mathbf{y})||q(\mathbf{x},\mathbf{z},\mathbf{y}))$, and Eq. 3 is an approximate loss function of the model in the actual implementation which aims to indicate the physical meaning of the model. The first and third terms represent the reconstruction loss and the satisfied distribution of $\mathbf{y}$, respectively. The second item is utilized to make $\mathbf{z}$ align with the 'exclusive' distribution of a certain cluster. Due to the space limit, we refrain ourselves from expanding too many details of this method, which is not our technical purpose. Certainly, we will proofread carefully and add more details in the revised version given additional one page to improve the readability.

**Its applicability to other deep clustering networks**. The reviewer has expressed the concern about how our technique can be applied to other clustering networks, as it is supposed to be coupled with the clustering model presented in Section 2 (while in fact NOT). Our adversarial learning algorithm is **model-agnostic** and can therefore be applied to the deep clustering model that follows the $\mathbf{x} \rightleftarrows \mathbf{z} \rightarrow \mathbf{y}$ structure. The model in Section 2 (the Conv in the experiments) is one of such models leveraged to introduce one adversarial learning algorithm and afterwards verify its effectiveness encountering an attack algorithm. $\mathcal{L}_C$ changes as the clustering model changes. For other deep clustering models, e.g., VaDE and IDEC, $\mathcal{L}_C$ is their corresponding objective function. To be precise, our defense algorithm is to integrate a set of perturbation-based contrastive constraints into the original clustering model, which can use the learned perturbations to improve both the clustering performance and robustness. Particularly, VaDE [12] assumes a mean-field approximation in the generative process, which is different from the model introduced in Section 2.

**Hyper-parameter test**. We aim to select the hyper-parameters that can achieve a certain difference in the result (the 3rd term of Eq.6) by a moderate perturbation (not too much). The roles of $\gamma$ and $\beta$ are mutually exclusive, so we typically fix $\beta$ and tune $\gamma$. The experiments show that the difference in the result will increase abruptly as $\gamma$ gradually increases, and the critical value $\gamma$ is an ideal hyper-parameter. For example, the difference in the result (not clustering accuracy) is 0.05 (from 0.88 to 0.83) when $\gamma$=0.05 for the Conv with 128-D features, while the difference is 0.21 (from 0.88 to 0.67) when $\gamma$=0.06. Obviously, $\gamma$=0.05 is relatively suitable, and the final clustering result also verifies this choice. The clustering accuracy drops to 0.592 when $\gamma$=0.06, which has almost killed the basic capability of the network.

**Novelty and contributions**. We re-summarize our contributions in the context of R #4's concerns in three aspects: **1)** To the best of our knowledge, this paper is the first work for learning unsupervised adversarial clustering networks, which is in fact very important due to the vulnerability nature of both deep clustering and unsupervised learning (see the comments from R#3 and R#5) but has rarely been studied in the literature. Our research may inspire more efforts from the machine learning community. **2)** We propose both novel attack and defense strategies which can be readily applied to most existing deep clustering methods, owing to our technical design. Our attack technique can also be used to identify the unreliable samples from unlabeled data, permitting data mining applications. **3)** We perform extensive experiments to corroborate the effectiveness of our method as well as the importance of adversarial learning for existing deep clustering models, through the three parts in the experiments on attack, defense, and re-attack strategies.

**Points for clarification**. We will try our best to improve the writing and release the code. Here are some specific explanations: **1)** The description of Table 3 is shown in line 236-239 in the main paper, where we re-attack the network after our defense strategy to further verify the effectiveness of the defense strategy. **2)** Please refer to our response to **Reviewer** #3 for the details about the discriminator $T$. **3)** The attack strategy will learn the corresponding perturbation according to each sample. **4)** The example of Section 1 is the results of the basic clustering model (Conv) with 128-D features. The clustering accuracy dropped from 0.849 to 0.772.

# 3 To Reviewer 5

**1**) For the MIE and Graph modules, their corresponding $\mathcal{L}_C$ adds mutual information $I(\mathbf{x},\mathbf{z})$ and graph constraints $\sum_{i,j} \mathbf{W}_{ij} \parallel \mathbf{y}_i - \mathbf{y}_j \parallel_2^2$ to the basic model. Due to the space limit, we will add a complete formula in the revised version. **2**) The generated images and their corresponding $\mathbf{y}$ are combined via feature reshaping. The combined result and its latent representation $\mathbf{z}$ together form a sample pair as the input of the discriminator.