We thank the recognition from reviewers on the value of research problem, novelty of the proposed method and our results. We address major raised concerns below. One missing reference (R1) will be added in the revised version.

R1: fine-tuned or trained from scratch. The work of [42] has shown that given enough data, training from scratch or fine-tuning from a pretrained model could both learn a good generative model for the target domain, while the fine-tuning simply gets faster convergence. As stated on L2-4, for the few-shot scenario, while it is unlikely to train the model from scratch with a few data, we design our pipeline in the fine-tuning fashion. This is why we identify the trend from the fine-tuned model because there are more correspondences on weights changes under the same way of learning.

R1: FID on other shots. We additionally evaluate our method against existing approaches on other shots in the following table. The performance of NST will not be improved obviously because the style transfer method cannot capture the target style well (L200-202). As stated on L244-246, the BSA's performance drops when increasing the number of shots. The MineGAN and our work behave similarly as both are distribution learning based methods. The more data, the better performance and the closer these two methods are.

Table 1: Quantitative comparisons between different few-shot generation methods (FID↓).

Number of shots	NST [6]	BSA [28]	MineGAN [41]	Ours
1	212.23 ± 9.77	102.34 ± 5.70	102.57 ± 4.76	84.36 ± 3.91
10	204.16 ± 9.28	105.56 ± 5.79	86.44 ± 4.38	74.87 ± 3.75
100	199.52 ± 9.02	110.24 ± 5.87	76.23 ± 4.05	67.55 ± 3.48
1,000	196.43 ± 8.83	119.31 ± 5.96	69.20 ± 3.59	62.40 ± 3.12
10,000	194.88 ± 8.71	131.20 ± 6.04	58.69 ± 3.14	55.74 ± 2.88

R2: clarification on FI. The Fisher Information is computed for each pretrained parameter $\theta_{S,i}$ in the **source** model, and used as an importance weight to regularize the changes of each parameter θ_i during the adaptation on target domain. The i is the index of each parameter in the model. Note that the FI is computed for each single parameter instead of each layer. What Figure 2(right) shows is the **average** FI of all parameters at each layer for easier visualization.

R2: effect of each layer. Thanks for the suggestion to ablate the importance of each layer and we will include it in the revised draft. We mainly focus on the individual parameter because the FI is computed for each parameter.

R2: source domain in Table 3. As stated on L261, we select the FFHQ face dataset as the source domain.

R3: other datasets. In addition to results (256×256) on face datasets, we also presented higher-resolution (512×512, 1024×1024) face generation in Figure 19-20 of the supplementary material and the landscape generation in Figure 4 of the paper. We think our approach can be also used for other datasets with more structures details based on two premises: i) a decent pretrained model on the source domain and ii) some level of similarity between the source and target domain.

R3: more baselines on F_i. The baseline of removing F_i (i.e., without EWC used) is shown in Figure 3 of the paper and discussed on L161-173. We also add another baseline by giving some fixed weights (i.e., the average FI in Figure 2 (right)) for all parameters in each layer. We observe that it will not lead to obvious over-fitting and the performance is slightly worse (FID: 77.16 vs. Ours 74.87 for 10-shot). Comparing with treating all parameters at each layer equally with the same weight, ours by regularizing each single parameter with its own FI still works better.

R4: limited contribution. Technically, our work on regularizing the model's own parameters is to **avoid** introducing additional new parameters as did in existing approaches [28,41] which involves many tedious manual designs (e.g., the number of parameters, the position to embed those parameters). The proposed method is simple and effective, and may shed light on more future understandings of the learned parameters. Practically, we proposed a solution to the challenging few-shot generation problem (e.g., 10 examples) as generative models often struggle in the low-data regime.

R4: later layers are important. Thanks for pointing out this improper statement. We shall not conclude parameters in the later layer are all important based on the **average** FI. It should be some important parameters with much larger FI that results in the higher average FI. We will redesign Figure 2 with the standard deviation for better visualization.

R4: analysis of the regularization weight. We assume R4 is referring to λ in Eq. (3) as this is the only parameter we empirically set the value for. We show its effect on the performance (FID) in the following table (10-shot). A large λ tends preserve more style of the source and a small λ may result in the some level of over-fitting. In addition, we find that (i) if the source and target are more similar, select a larger λ, and (ii) if more target data is given, select a smaller λ.

λ	5×10^6	5×10^7	5×10^8	5×10^9	5×10^{10}
FID ↓	78.21 ± 3.80	75.62 ± 3.74	$\mathbf{74.87 \pm 3.75}$	77.79 ± 3.78	80.21 ± 3.82