Response Summary. We truly appreciate reviewers' valuable comments and positive feedback. We are encouraged reviewers found the problems addressed by our paper are important (Reviewer 3, 4), our motivation is clear (Reviewer 1, 2, 3), and idea is novel (Reviewer 1, 3, 4). We are glad all reviewers found this paper is well written with excellent reproducibility. One primary concern was the results on DARTS search space were not SOTA. Similar as DARTS and other differentiable NAS, we also conducted more architecture search for high-performance models. We have released a new architecture on GitHub (the anonymous link in the original submission file), with 2.50% test error on CIFAR10, 15.84% on CIFAR100, and 24.37% on ImageNet, which are SOTA results under DARTS experimental settings. Thanks Reviewer 3 for the strong support, and we will provide a detailed item-wise response for other reviewers below.

Responses to Review 1:

Q1: As for exploration, there are some other traditional methods like BO. What the difference between BO and your way is. **R1**: BO needs to build probabilistic models (e.g. GP) by training multiple architectures from scratch first, while differentiable NAS only trains a supernet once. It is not intuitive to directly introduce the exploration in BO to differentiable NAS, while our exploration could be easily applied to differentiable NAS to solve rich-get-richer problem.
Q2: Is it difficult to tune hyperparameters? **R2**: No. There are only two hyperparameters to be tuned, and the others are default. The ablation studies in Sec 4.3, 4.4, and Appendix H show our model is robust to the two hyperparameters.
Q3: Some SOTA one-shot NAS methods [1,2] in the NAS-Bench-201 dataset missed. And the results on DARTS search space are not the SOTA. **R3**: The two references are missing in the review. Our best single run can achieve 46.48% for ImageNet on NAS-Bench-201 as described in the bottom of Table 1 and line 246 in the original submission, outperforming [1] (46.34% for ImageNet) provided by the reviewer. Please refer to ***Resp. Summ.*** and GitHub for the results of our new high-performance model on the DARTS search space.

Responses to Review 2:

Q1: The paper is not very novel with limited contribution, the method mitigating catastrophic forgetting is not new and similar with the previous EWC work. **R1**: This paper is the first paper introducing intelligent exploration into NAS, through the probability density function based on a graph autoencoder. We believe it is novel in NAS and Reviewer 1 and 3 also agree on it. Besides, the regularization method to mitigate forgetting in our method is totally different from EWC and WPL (EWC applied to one-shot NAS). EWC and WPL both calculate the joint posterior probability through estimating the Fisher information matrix and assuming the previous models in optimal points, while the two conditions hardly hold in differentiable NAS. We propose an architecture complementation scheme, and theoretically shows it could optimize the joint posterior probability as EWC and WPL, without the assumption of the two conditions.
Q2: NAS-Bench is not a well-established benchmark that not many people are very familiar with. **R2**: The NAS-Bench is a newly established benchmark with a much simpler search space, while the ground-truth test accuracy for all candidates in the search space is reported, helping the NAS methods to conduct reproducible experiments with much less computational requirements. Building a well-established benchmark is becoming a new interesting research direction in NAS, and concurrent NAS-Bench 101, NAS-Bench 201, NAS-Bench 1Shot1, et. al., all help to relieve computational requirements, and recent researches in NAS community prefer these benchmarks for enhancing reproducibility.
Q3: Require a stronger ImageNet result. **R3**: Please refer to ***Resp. Summ.*** and GitHub for our new SOTA results.
Q4: The author should not treat supernet training as a multi-task learning problem. The reviewer agrees that it is a multi-model optimization problem, not a multi-task problem. **R4**: This paper focuses on the multi-model forgetting problem in the supernet training. The multi-model forgetting in NAS is very related to catastrophic forgetting in multi-task learning, as described in Sec. 2. To avoid confusion, we have rephrased "catastrophic forgetting" to "multi-model forgetting" when describing the forgetting in NAS, and rephrased Sec.3.2.
Q5: In relieving forgetting, the selection of three models seems arbitrary and not intuitive. **R5**: It should be noted that our architecture complementation scheme is to select specific architectures for regularization. We theoretically show our method can optimize the joint posterior probability similar as EWC and WPL, but with less constraints. The ablation study in Appendix H shows our method outperforms other naive schemes, including WPL, random selection, and so on.
Q6: The paper directly compares related works copied from the previous paper without hyperparameter tuning. **R6**: The results of peer algorithms are from the original paper (NAS-Bench-201) since we adopt the same experimental settings as that paper. Furthermore, our model outperforms most peer algorithms under all hyperparameter settings.

Responses to Review 4:

Q1: In Table 2, γ = 0.2 performs the worst with huge variance? Should the performance change smoothly with γ?
R1: The performance should change smoothly with γ. As discussed in Sec. 4.3, a dynamic γ seems to achieve better performance, and a small and static γ may lead to local optimal. As we conducted experiments with limited random seeds, the outliers may greatly affect the statistical results. We have conducted experiments with more random seeds and will remove outliers to obtain statistical results to avoid the effects of outliers in the final version.
Q2: The definition of complementary/orthogonal architecture? Should the union of α_{i-1} and α_i^c be the whole search space, or just needs the union of α_{i-1} and α_i^c includes α_i? **R2**: We define that α_m is orthogonal to α_n, so they do not share parameters $\omega_m \cap \omega_n = \emptyset$. As to the complementary architecture, since we first select the α_{i-1} into the replay buffer, the complementary architecture α_i^c is defined as $\omega_i \subseteq \{\omega_i^c \cup \omega_{i-1}\}$ that only needs the union of α_{i-1} and α_i^c includes α_i, and α_i^c is also orthogonal to α_{i-1}.