We sincerely thank the reviewers for their feedback. We appreciate they find our work is **"well-motivated"** (R1, R4) and **"valuable"** (R3); the approach is **"effective"** (R1, R4) and **"efficient"** (R3) in implementation and computation; the empirical methodology is **"convincing"** (R1); the experiment design is **"extensive"** (R1) and **"sufficient"** (R4); the results are **"convincing"** (R1) and **"promising"** (R3, R4). We also appreciate they find the paper is **"well-written"** and **"clear"** (R1, R3, R4). We address the main concerns raised by the reviewers in the rebuttal, and will incorporate all suggestions for changes in the final version. We sincerely hope this will help the reviewers to finalize their judgments.

Q1. Responses to R1 on differences and relation to prior work of our method.

(1) Differences with [62]. Our work differs from [62] in two aspects. (i) We are the first work to investigate adversarial data argumentation from an information theory perspective, and address the problem of generating "hard" adversarial perturbations from IB principle which has not been studied yet. (ii) We theoretically show that IB principle can be bounded by a maximum-entropy regularization term in the maximization phase of data argumentation, which results in a notable improvement over [62]: $\geq 5\%$ in accuracy on Digits and CIFAR datasets. *(2) "The derived bounds are special cases of existing theorems".* Our bounds are general in the scenario of DNNs, since they hold under deterministic and non-deterministic conditions which cover almost all types of DNNs. Thus they can be widely applied in the training of DNNs and are valuable for further research in this field. *(3) "How the domain generalization and data augmentation are connected and evolved"?* Our work studies a more challenging setting where networks are learned using **one single** source domain compared with conventional domain generalization that uses multiple training source domains. Adversarial data augmentation are employed to synthesize virtual target domains during training so that the generalization and robustness of the learned networks can be improved. We will clarify this part in the final version.

Q2. Clarification and correction on the necessity and clarity of the theorems. (R2)

We thank R2 for the detailed comments on our theoretical results. We address the necessity of Theorems 1 and 2 and will improve the clarity of Theorem 3 and Lemma 1 in the final version. *(1) Necessity of Theorem 1.* Theorem 1 is necessary because it reveals the connection of IB principle and the proposed maximum-entropy regularization during adversarial data argumentation. It is an important theoretical contribution of our work. *(2) Clarification on Theorem 2.* Our bound in Theorem 2 is valuable. First, it is simple and sufficient for optimizing our model. Second, for non-optimal estimators (which is the condition of networks in the maximization phase during data argumentation), the bound in [1*][1] (Proposition 1) is not stronger than ours. Both [1*] and our bound are dominated by $\mathcal{O}(|\mathcal{Y}|/N)$ in this case. We will cite and discuss [1*] in the final version. *(3) "Theorem 3 is an obvious Corollary of [31]".* We will reformulate this theorem as a corollary of [31] in the final version. *(4) Correction on Lemma 1.* We apologize for the miscitation. Lemma 1 should be a trivial adaptation of [62] (Lemma 1). We will correct and cite it in the final version.

Q3. "The motivation for the proposed regularization is somewhat lacking". (R2)

We do not agree. Our motivation has been recognized by R1, R3 and R4. R2 might misunderstand the scope of our work. We do not study the problem of adversarial defence where "neural networks can be confidently wrong with **adversarial examples**" as indicated by R2. Instead, this work improves the generalization and robustness of networks for unseen **real images** by adversarial data augmentation. To achieve this, we propose to train networks with "hard" perturbations against source data by incorporating the IB principle (lines 29-35). Then we theoretically show our regularization term which favors points with high uncertainty can bound the IB principle in Eqs. (6) and (7). The results on four **real-world datasets** demonstrate the proposed regularization improves the usual adversarial training [62] by a large margin.

Q4. Correctness and clarity of the paper. (R2)

(1) Correctness of lines 190-191. We will remove this statement in the final version, because the formal and accurate description of the problem setup has been defined in Theorem 3. *(2) In line 82, authors say they will approximate $I(X;Z)$ by ℓ_2 penalty, but then never discuss it again.* As stated in lines 94-96, this is a standard technique for training DNNs in the literature. We have clearly cited [18, 20] and indicated that we followed their implementation in our work.

Q5. Responses to R3.

(1) Add more explanations of "hard" adversarial perturbations. We will provide more detailed explanations of "hard" adversarial perturbations in the abstract and a paragraph within the method in the final version. We will also add figures of training losses used for measuring them quantitatively. *(2) Add DomainNet or Office dataset.* Our setting is different from domain adaptation since no image from the target domain is used during training. Thus our method is not comparable to domain adaptation methods on these two datasets. *(3) Add more competing methods for Digits dataset.* On this dataset, AGG [42] degrades to [60]. PAR [63] is open-source and we will add its result in the final version.

At last, we will rephrase the corresponding parts (R2, R3) and cite related references (R2). The remaining questions about writing will be carefully addressed. We thank the reviewers for their careful feedback and consideration.

[1] [1*] Y. Wu and P. Yang. "Minimax Rates of Entropy Estimation on Large Alphabets via Best Polynomial Approximation".