We would like to thank you for your thorough evaluation, helpful suggestions, and comments. We here address the key concerns and note that the paper will be updated accordingly. We conducted new experiments to reinforce the evidence for this response, as reviewers suggested. Before we begin, however, we emphasize that the paper provides the first efficient instance-aware test-time augmentation method resulting in significant gains over previous approaches.

Table 1: ImageNet(-C) result of ResNet-50 with the standard training-time augmentations.

Test-time Augmentation	Relative Cost	Clean Test-set	Corrupted set mCE	Corrupted Test-set mCE
Center-Crop	1	24.14	78.93	75.42
Horizontal-Flip	2	23.76	77.91	74.32
5-Crops	5	23.91	77.52	73.87
10-Crops	10	23.04	76.69	72.98
Random($k = 1$)	1	26.89	82.86	79.81
Random($k = 2$)	2	25.14	79.91	77.00
Random($k = 4$)	4	24.29	78.24	75.38
GPS($k = 1$)	1	24.86	82.13	79.43
GPS($k = 2$)	2	23.78	76.45	73.32
GPS($k = 4$)	4	23.44	77.27	73.87
GPS†($k = 1$)	1	27.39	77.21	75.07
GPS†($k = 2$)	2	27.04	76.48	74.27
GPS†($k = 4$)	4	26.88	76.09	73.84
Ours($k = 1$)	1	24.14	75.52	74.29
Ours($k = 2$)	2	24.10	75.00	73.61
Ours($k = 2$) + Flip	4	23.74	74.00	72.59

*GPS : Greedy Policy Search on the clean dataset.
*GPS†: Greedy Policy Search on the corrupted dataset.

Figure 1: Comparison for the same 5 Crop candidates on the clean ImageNet set using ResNet-50. Top-1 accuracies by the number of ensembles. We trained our loss predictor for five crop areas. Compared to the 5-crop ensemble, choosing one transform by our method gives almost the same performance, and selecting the two transforms achieves even better performance with less computational cost.

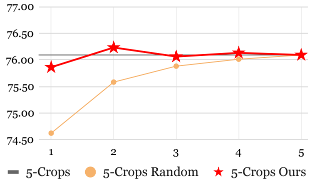
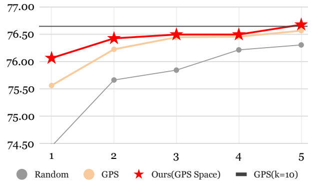


Figure 2: Comparison for the same GPS transforms on the clean ImageNet set using ResNet-50. Top-1 accuracies by the number of ensembles. We trained our loss predictor on the searched GPS policies to choose ones specific for each test instance. Our method properly selects valid transforms from the candidates chosen greedily by GPS, and therefore further improves the performances over static ensemble from GPS.



Comparison study with GPS (Greedy Policy Search) [30] (R1+R2+R3): The official code for GPS (released after our submission) is used for comparison. In Table 1, we show that the proposed method outperformed both GPS and GPS† on ImageNet-C. This means that the performances of GPS on both seen and unseen corruptions lagged behind our proposed method. In particular, the GPS policies found on the corrupted dataset produced poor results in the clean set, while our method prevented the performance degradation on the clean set. We confirmed by the GPS code that the search space of GPS includes all our augmentation policies such as "auto-contrast" and "sharpness"; our performance gains come from the proposed instance-specific transformation. A detailed comparison will be included.

Test-time augmentation for the clean set (R2+R3): We conducted experiments with loss predictor trained for the clean set. In Figure 1, our loss predictor picks out promising one out of five crop regions. Even if only one crop region is selected using the loss predictor, the obtained performance is comparable to the existing 5-crop ensemble. This is clear proof that our method is also effective on the clean set and separated from our search space, our loss predictor itself contributes to enhancing the classification performance. We will include the result of the clean test-set.

Validating loss predictor (R2+R3): Firstly, we add random baselines with $k \geq 1$. As k increases, the random baselines' performance marginally increases, but our method using loss predictor instead of random selection shows significant improvements. Secondly, in Table 1, our method uses augmentation space, which is a subset of GPS's space. Nevertheless, our performance is better since we select the best one for the test instance with the loss predictor. Lastly, in Figure 1 and 2, it is also a critical rationale that performance increases consistently when the order of the policy is dynamically determined with the loss predictor. We will elaborate more on this in the revised paper.

Details of the loss function (R1+R4): We used the surrogated ranking loss proposed in [8], as described in L171. Specifically, to optimize the non-differentiable Spearman correlation between relative losses and predictions, we trained a recurrent neural network that approximates the correlation using the official implementation. This surrogate loss function has been chosen after an extensive comparison with others. We will revise the paper with the details.

R1: We will add the missing related works and add calibrated log-likelihood to our revised paper. **R2:** As the reviewer pointed out, our test-time transformations consist of basic operations that may restore a corrupt image close to normal. However, transformations for a given test image is selected by the loss predictor. As [17] shows, manually targeted image restoration can be harmful to robustness when the corruption of each test image is unknown at test-time. In addition, as shown in Table 2 and 3 in the manuscript, taking into account our transformations at training-time of the target network leads to performance degradation on some corruptions and (most importantly) clean set. The proposed loss predictor contributes to picking the most proper one that not distorts more but may restore a corrupted image, which improves the robustness of the target network in a consistent way. **R3:** In Table 1, we compare the performance of baselines and our method on the ResNet-50 trained with the standard train-time augmentation. We will update the experimental results with various train-time augmentations and more baselines, and revise the manuscript to reflect your additional comments. **R4:** As the number of transformations increases, the cost of transforming and inferencing the input linearly increases. But this is highly parallelizable. Also, in this study, we prepared a small setting to focus on demonstrating the potentials of instance-aware test-time augmentation. Although the augmentation space is limited, the experiment results show the superiority of our methods against previous approaches. Also, applying augmentations repeatedly to expand transformation space in a combinatorial way is promising in our experiment. We expect future works to be conducted in the direction of using less cost while expressing more augmentations.