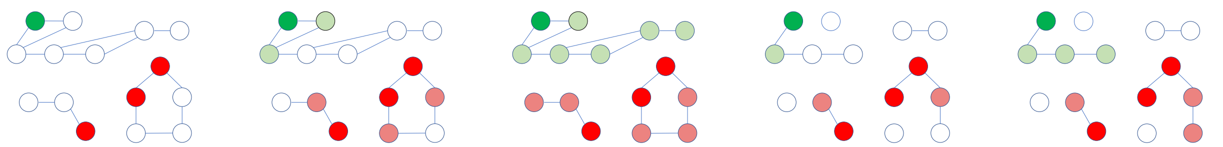
We thank the reviewers for constructive feedback. We are delighted that the reviewers find the paper well-written and appreciate the strong empirical results as well as the theoretical analysis.

To Reviewer 1: **[Intuition of benefits of advanced data augmentation]** In line 198, we explained the theoretical connection between advanced data augmentation and better semi-supervised learning performance. We stated that "Importantly, the number of components is actually decided by the quality of the augmentation operation: an ideal augmentation should be able to reach all other examples of the same category given a starting instance. This well matches our discussion of the benefits of state-of-the-art data augmentation methods in generating more diverse examples. Effectively, the augmentation diversity leads to more neighbors for each node, and hence reduces the number of components in a graph."



(a) Supervised learning (4/15) (b) Advanced supervised augmentation (9/15) (c) UDA with advanced augmentation (15/15) (d) Simple supervised augmentation (7/15) (e) UDA with simple augmentation. (10/15)

Figure 1: Prediction results of different algorithms, where green and red nodes are labeled nodes, white nodes are unlabeled nodes whose labels cannot be determined and light green nodes and light red nodes are unlabeled nodes whose labels can be correctly determined. The accuracies of different algorithms are shown in $(\cdot)$.

Since supervised data augmentation only propagates the label information to the directly connected neighbors of the labeled nodes. Advanced data augmentation that has a high accuracy must lead to a graph where each node has more neighbors. Effectively, such a graph has more edges and better connectivity. Hence, it is also more likely that this graph will have a smaller numbers of components. To further illustrate this intuition, in Figure 1, we provide a comparison between different algorithms. In contrast, the neighbors in nearest neighbor and label propagation are determined by Euclidean distances, which may not have the same labels and may violate the label-preserving assumption used in our analysis. We will include this detailed explanation in the future version.

To Reviewer 2: **[The traversal of the entire sub-graph]** The traversal means that consistency training can propagate labels from labeled nodes to directly connected unlabeled nodes, and then to all connected unlabeled nodes in a component. Please see Figure 1 for an illustration.

[Citation style] Thank you for pointing this out! We will refine the citation style in the future version.

[Adaptive variant of AutoAugment] We agree that adaptively refining the data augmentation can provide more valid noise. We will include this insight in the future version.

To Reviewer 3: **[Contributions]** Our contributions are not only a simple change that leads to better performance, but also an effective framework that is applicable to many tasks, the theoretical insight on why advanced data augmentation works, the state-of-the-art performance and the consistency between theory and practice. As the reviewer has noted, the proposed method is simple and widely applicable, which will attract attention from many researchers of different areas.

[Lacking analysis] We acknowledge that no ablation studies are included in the main paper due to the space limits. The ablation studies are available in the supplementary material B.2. We show that the success of RandAugment should be credited to the diversity of the augmentation transformations, since the model's performance gradually improves as we use more augmentation transformations.

For effective augmentation techniques, if we only use one data augmentation technique, the best augmentations are Equalize, Color and Brightness for CIFAR-10 and Invert, Equalize and ShearX for SVHN. We have also performed experiments that combine UDA with VAT. We find that UDA+VAT leads to similar performance with UDA as shown in Table 1, which means that the data augmentation noise is good enough and adding extra adversarial Gaussian noise does not help.

Methods / # Sup	250	500	1,000	2,000	4,000
UDA	5.43 ± 0.96	4.80 ± 0.09	4.75 ± 0.10	4.73 ± 0.14	4.32 ± 0.08
UDA + VAT	5.89 ± 1.12	4.86 ± 0.16	4.81 ± 0.13	4.65 ± 0.07	4.27 ± 0.15

Table 1: Comparison between UDA and UDA + VAT

To Reviewer 4: We thank the reviewer for the feedback. We will clarify the our contributions in the future version.