Appendices

A. k-means++ Seeding Algorithm

As shown in Algorithm 2, the core idea of k-means++ seeding algorithm is to sample S centers sequentially, where each new center is sampled with probability proportional to the squared distance to its nearest center. The set of centers returned by Algorithm 2 is theoretically guaranteed to far away from each others [1].

Algorithm 2: k-means++seeding Algorithm [1]

```
Input: $G := \{p_i : p_i \in \mathbb{R}^D\}$; Target size $S$
Output: Center set $C$ of size $S$
$C_1 = \{c_1\}$, where $c_1$ is sampled uniformly at random from $G$
for $t = 2, \cdots, S$ do
    $E_t(x) := \min_{c \in C_{t-1}} ||x - c||_2$
    $c_t \leftarrow$ sample $x$ from $G$ with probability $\frac{E_t^2(x)}{\sum_{x \in G} E_t^2(x)}$
    $C_t \leftarrow C_{t-1} \cup c_t$
end
return $C_S$
```

B. The details about the benchmark datasets

The detailed statistic and the default data augmentation for the benchmark datasets are listed as belows.

- **CIFAR-10 & CIFAR-100 [18]:** The training sets of the two datasets are composed of 50,000 colored images with 10 and 100 classes, respectively. Each image in these two datasets is in size of 32 × 32. For CIFAR datasets, the default augmentation crops the padded image at a random location, and then horizontally flips it with the probability of 0.5. Then, it applies Cutout [8] to randomly select a 16 × 16 patch of the image, and set the pixels within the selected patch as zeros.

- **SVHN [21]**: This dataset contains color house-number images with 73,257 core images for training and 26,032 digits for testing. The default augmentation crops the padded image at a random location. Then it applies Cutout to randomly select a 16 × 16 patch of the image, and set the pixels within the selected patch as zeros.

- **ImageNet [7]:** ImageNet includes colored images of 1,000 classes. The training set has roughly 1.2M images, and the validation set has 50,000 images. The default augmentation randomly crops and resizes images to a size of 224 × 224, and then horizontally flips it with a probability of 0.5. Subsequently, it performs ColorJitter and PCA to the flipped image [19].

C. Ablation Study

C.1. Case Study

In Figure 6, DivAug's candidate images are obtained by only applying the single transform `Rotate` with fixed probability parameter p (the magnitude parameter remains random). As shown in Figure 6, Variance Diversity and model performance are highly correlated.

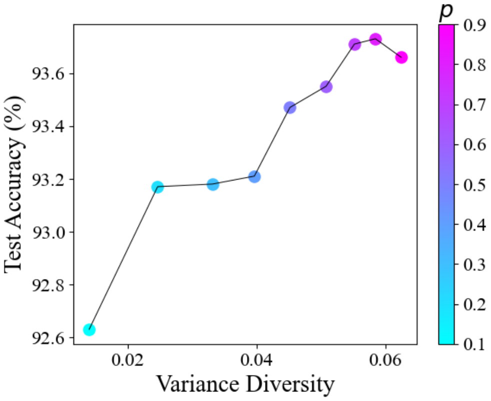

Figure 6: The case study of applying single transformation `Rotate`. To control Variance Diversity of augmented images, the fixed p is varied from 0.1 to 0.9.

C.2. Comparison between DivAug and the Random Baseline

To check the effect of k-means++ in DivAug, we compare the performance of Wide-ResNet-28-10 with DivAug and that with the random baseline on CIFAR-10 and CIFAR-100 in Table 6. For a fair comparison, the random baseline here randomly picks four augmented images from eight candidates for training. Also, the magnitude m and probability p are also randomly picked. As shown in Table 6, DivAug is significantly better than the random baseline.

D. Detailed Analysis For The Correlation between Variance Diversity and Generalization

Recently, two measures, Affinity and Diversity, are introduced in [11] for quantifying distribution shift and augmentation diversity, respectively. Across several benchmark datasets and models, it has been observed that the performance gain from data augmentation can be predicted not by either of these alone but by jointly optimizing the two [11]. Specifically, Affinity quantifies how much a sub-policy shifts the training data distribution from the original one. For a set of augmented data, our proposed diversity measure is calculated based on the variance of their probability vectors. Meanwhile, the diversity measure proposed in [11] is defined as the training loss of a given model over the augmented data. Below, we give the formal definition of Affinity and Loss Diversity:

Table 5: Training hyperparameters of CIFAR-10, CIFAR-100 and ImageNet under the supervised settings. LR represents learning rate, and WD represents weight decay. We do not specifically tune these hyperparameters, and all hyperparameters are consistent with those reported in Adversarial AutoAugment [28].

Dataset	Model	Batch Size	LR	WD	Epoch	LR Schedule
CIFAR-10	Wide-ResNet-40-2	128	0.1	5e−4	200	cosine
	Wide-ResNet-28-10	128	0.1	5e−4	200	cosine
	Shake-Shake (26 2x96d)	128	0.2	1e−4	600	cosine
	PyramidNet+ShakeDrop	128	0.1	1e−4	600	cosine
CIFAR-100	Wide-ResNet-40-2	128	0.1	5e−4	200	cosine
	Wide-ResNet-28-10	128	0.1	5e−4	200	cosine
	Shake-Shake (26 2x96d)	128	0.1	5e−4	1200	cosine
ImageNet	ResNet-50	512	0.2	1e−4	120	cosine

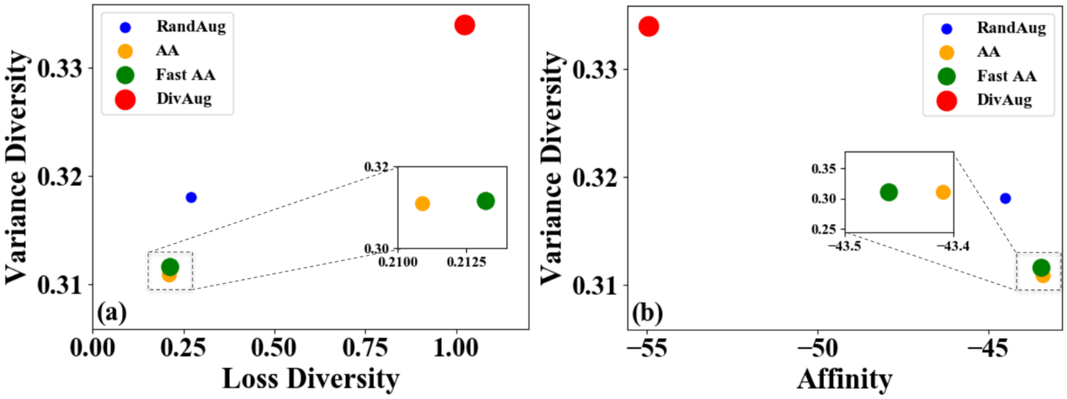

Figure 7: **The performance gain is positively correlated to Variance Diversity. Also, the Loss Diversity and Variance Diversity are highly correlated.** The marker size in the legend indicates the relative gain in test accuracy of different methods. (a) The Loss Diversity and the Variance Diversity of augmented data generated by different methods. All points lies near the diagonal of the Figure. In general, the relative gain in test accuracy increases with larger Variance Diversity (b) The Affinity and Variance Diversity of augmented data generated by different methods.

Table 6: The performance of Wide-ResNet-28-10 with DivAug and with the random baseline.

Dataset	Method	Accuracy
CIFAR10	Random ($S = 4$)	$97.7 \pm .1$
	DivAug	$98.1 \pm .1$
CIFAR100	Random ($S = 4$)	$83.3 \pm .2$
	DivAug	$84.2 \pm .2$

Definition 1 (Affinity [11]). *Let D_{train} and D_{val} be training and validation datasets drawn i.i.d. from the same clean data distribution, and let D'_{val} be derived from D_{val} by applying a stochastic augmentation strategy, a, once to each image in D_{val}, $D'_{val} = \{(a(x_i), y) : \forall (x_i, y) \in D_{val}\}$. Further let m be a model trained on D_{train} and $\mathcal{A}(m, D)$ denote the model's accuracy when evaluated on dataset D. The affinity $\tau[a; m; D_{val}]$ is defined as:*

$$\tau[a; m; D_{val}] = \mathcal{A}(m, D'_{val}) - \mathcal{A}(m, D_{val}) \tag{7}$$

Definition 2 (Loss Diversity [11]). *Let D_{train} be the training set, and D'_{train} be the augmented training set resulting from applying a stochastic augmentation strategy α. For a set of augmented data $\mathcal{S} = \{x'_i\}$, where x'_i is obtained by applying α to x_i, stochastically. Further, given a model m which is trained on D'_{train}, let $\hat{L}_i$ be the training loss corresponding to x'_i. The Loss Diversity between $\{x'_i\}$, $\mathcal{D}_{\text{loss}}(\{x'_i\})$, is defined as:*

$$\mathcal{D}_{\text{loss}}(\mathcal{S}) = \mathbb{E}_{x'_i \in \mathcal{S}} \hat{L}_i \; ^{\ddagger} \tag{8}$$

As we analyzed, given a set of augmented data which has large Variance Diversity, it is hard for models to give consist predictions for them, which will result in a large training loss. Thus, Loss Diversity and Variance Diversity are highly correlated. The main difference between them is that Variance Diversity is a unsupervised measure, *i.e.*, Variance Diversity is not related to the label information.

We further plot the performance gain from each augmentation methods against the Affinity, Loss Diversity, and Variance Diver-

[‡] The original definition of Loss Diversity is defined for the entire training set. To make it comparable to Variance Diversity, we extend the concept to a set of augmented data generated from a same original data x_i.

sity of the augmented data generated by them in Figure 7. In the legend, the marker size indicates the test accuracy of a Wide-ResNet-40-2 model trained with different automated data augmentation methods (The detailed results are shown in the first row of Table 2). Figure 7 demonstrates the Loss Diversity and Variance Diversity are highly correlated, which is consistent with our theoretical analysis. Following [11], we show the Affinity and Variance Diversity of augmented data generated by different methods in Figure 7 (b). There is a clear trend that the Loss Diversity and Variance Diversity contradict with the Affinity to some extent. We remark that although RA has larger Variance Diversity than AA and Fast AA, the performance gain from RA is smaller. According to the hypothesis in [11], this can be explained by RA has smaller Affinity than those of AA and Fast AA. In contrast, although DivAug has the largest Variance Diversity, largest Loss Diversity, and the smallest Affinity, DivAug performs best in terms of the test accuracy. We hypothesize that there might exist a sweet spot between the Diversity and Affinity, and how to achieve this sweet spot is a interesting future direction for the automated data augmentation methods.

E. Experiment Details

We list the details of training hyperparameters from the experiments in Section 4.3 in Table 5.

For the semi-supervised learning experiment in Section 4.4, we follow the settings in [24] and employ Wide-ResNet-28- 2 [26] as the backbone model and evaluate UDA [24] with varied supervised data sizes. For the experiments on CIFAR-10 with supervised data size 1000, 2000, and 4000, the hyperparameters of them are identical as below: we train the backbone model for 200K steps. We use a batch size of 32 for labeled data and a batch size of 448 for unlabeled data. The softmax temperature τ is set to 0.4. The confidence threshold β is set to 0.8. The backbone model is trained by a SGD optimizer with learning rate of 1e$-$4, weight decay of 5e$-$4, and the nesterov momentum with the momentum hyperparameter set to 0.9. We remark that all hyperparameters are identical to those reported in [24], except two differences: we train the backbone model for 200K steps instead of 500K, and we do not apply Exponential Moving Average to the parameters of backbone model.