A. CIFAR10-B Statistics

Classes	Green	Gray	Blue	White	Black	Brown	Red	Others
cat	171	130	137	57	84	299	63	59
dog	271	90	124	39	81	297	58	40
truck	121	180	278	187	24	177	2	31
bird	453	58	179	32	32	237	0	9
airplane	85	112	609	81	7	102	4	0
ship	71	88	711	45	15	56	0	14
frog	417	64	75	56	49	238	15	86
horse	493	29	108	41	50	251	7	21
deer	604	27	99	8	25	227	4	6
automobile	177	262	147	130	34	230	15	5

Table 5: **CIFAR10-B Statistics.** The number of instances that belongs to each background color from each class.

B. More discussion on chosen baselines and comparison to AugMix & AugMax

We choose the DAs that are strong and commonly studied in recent works such as [48, 43, 44]. PyTorch's official "SOTA" training method [44] also uses the combinations of our chosen baselines along with other perfected hyperparameters such as longer training; however, we isolated each baseline to measure the individual contribution of a single method. In addition, we compare with two recent works AugMix and AugMax, which are recent methods evaluated on plain Resnet18, and our generalization results are superior, as per reported in [45] (see Tab.6-*left*). AugMix and AugMax are not a single method but a combinations of different ones and were designed for other robustness purposes. Lastly, we want to re-iterate that generalization performance is not our main focus but an additional benefit when reducing the model bias toward specific colors.

C. ImageNet-10

We labeled ImageNet-10 and found similar subgroup degradation aspect on ImageNet, and our method can still mitigate the issue using standard training procedure without fine-tuning (Tab.6-*right*); therefore, we believe FlowAug will work on high-resolution/larger datasets

D. Selection of Color Groups

While labeling the datasets, we add one color only if it has a couple of images. This prevents having many colors with 0 images. For example, in CIFAR10, classes like "bird" and "ship" do not contain images with a red background, but it has some presence in other classes, so we added red as a color group.

E. Generalization on CIFAR10-C and CIFAR100-C

Aside from generalization on *i.i.d.* data, we are interested in FlowAug's capabilities to generalize to out-of-domain (OOD) data, which is another aspect of robustness. We use models of the last epoch from Table 2 to test on CIFAR10-C and CIFAR100-C. Although FlowAug does not explicitly add corruptions such as various kinds of blurring, contrast and so on to training data, we observe comparable performances (Fig. 8) with augmentation methods that have corruption effects, such as *MixUp*, and FlowAug is better than *Cutout* and *Cutmix* (+1.35% and +1.26%, respectively).

	CIFAR10	CIFAR100
AugMix	95.79	78.23
AugMax	95.76	78.96
Ours	**96.58**	**79.99**

	Acc.	MacroStd	W-Std
Standard	91.80	9.57	11.17
Autoaug	93.40	8.84	9.99
Ours-Alg1	**94.80**	**8.39**	**9.96**

Table 6: ***Left*: Accuracy of our FlowAug compared with AugMix & AugMax. *Right*: ImageNet-10 results.** Our FlowAug show superior results.

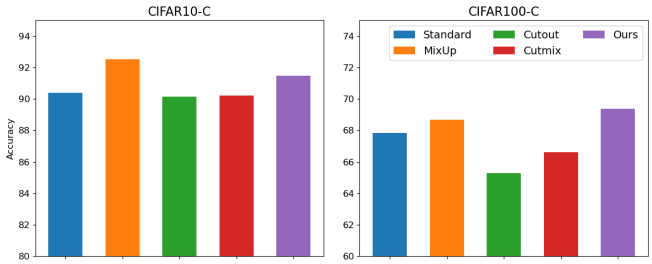

Figure 8: **CIFAR10-C and CIFAR100-C results (severity=1).** FlowAug's results are comparable with common DA methods that have corruption effects such as *Mixup*, even though FlowAug does not add corruptions to training. Note that *Mixup* (Figure 2(b)) produces a similar effect to blurrings.

F. Switch Operation

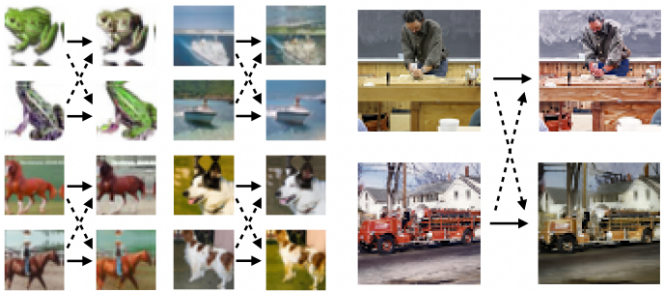

Figure 9: **Examples of switch operation on decoupled representations.** [28] can perform *switch* operations on global and local representations of images on various datasets (figure used with the author's permission).

G. Label Quality

The background color labels are labeled by a person with an experienced computer vision background for consistency and are verified twice. As a quality check, two people with strong technical backgrounds checked 500 random images. The rate of agreement is 91.8 percent, and 5.4 percent of images that did not agree in the first round agree with the labels we used for experiments after discussion. The disagreement rate is smaller than the average error rate in modern datasets [32]. We will release the dataset and welcome the community to update the background labels.

H. An Additional Protected Attribute

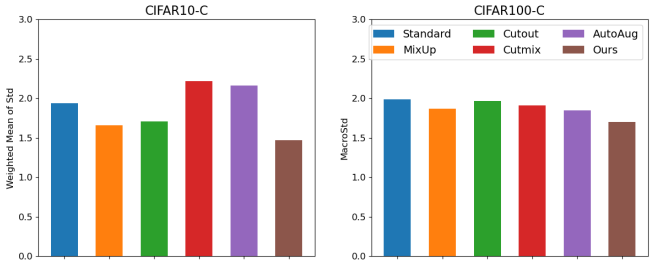

Figure 10: **Additional protected attribute on CIFAR10 and CIFAR100.**

In this work, we studied color as a bias, and we can also study another protected attribute in a hand-wavy fashion. For example on CIFAR10, we can group the ten classes into vehicles and animals and apply our MacroStd to measure sensitivity. A similar study can be conducted on CIFAR100 with its original 20 super-classes. On CIFAR10, since there are only two "super-classes" so we report the weighted average of the standard deviation of vehicle/animal group; on CIFAR100, we report our MacroStd across the 20 superclasses. The results are summarized in Fig. 10, and it shows FlowAug is again superior. We emphasize that this is a conceptual study on an additional protected attribute and is not within the scope of our work.

I. CIFAR100-B Statistics

Classes	Green	Gray	Blue	White	Black	Brown	Red	Others
apple	23	6	9	40	11	11	0	0
aquarium fish	30	5	12	0	38	9	4	2
baby	18	16	25	7	7	21	5	1
bear	56	15	9	1	2	16	1	0
beaver	36	8	18	8	6	24	0	0
bed	8	21	7	18	3	42	0	1
bee	24	8	10	6	4	33	8	7
beetle	29	13	8	15	1	28	3	3
bicycle	24	30	17	7	3	17	2	0
bottle	10	24	9	17	7	28	4	1
bowl	6	19	19	14	22	18	2	0
boy	21	13	16	13	8	21	5	3
bridge	11	9	67	5	5	2	1	0
bus	14	19	23	18	8	14	0	4
butterfly	53	12	3	7	4	17	2	2
camel	31	10	25	3	9	19	2	1
can	5	24	12	25	7	27	0	0
castle	4	12	65	15	1	2	0	1
caterpillar	63	6	7	0	6	15	2	1
cattle	47	5	13	8	5	20	1	1
chair	3	10	4	71	4	6	0	2
chimpanzee	65	3	4	3	1	20	1	3
clock	3	21	12	34	4	20	1	5
cloud	0	2	80	2	8	2	2	4
cockroach	2	19	13	44	1	14	5	2
couch	5	14	19	24	6	27	3	2
crab	9	23	16	16	15	16	3	2
crocodile	39	13	11	2	5	28	2	0
cup	9	27	18	17	14	11	1	3
dinosaur	23	14	12	26	7	17	0	1
dolphin	14	11	74	0	0	1	0	0
elephant	48	6	17	3	3	19	2	2
flatfish	15	14	34	13	6	12	6	0
forest	4	7	22	7	0	11	3	46
fox	32	11	19	1	7	26	2	2
girl	15	11	15	10	13	26	7	3
hamster	11	23	22	4	8	22	9	1
house	20	4	42	20	3	7	0	4
kangaroo	44	14	4	2	3	31	1	1
keyboard	8	15	23	8	5	23	3	15
lamp	5	20	23	14	12	20	5	1
lawn˙mower	32	6	6	45	2	9	0	0
leopard	31	14	20	1	8	22	1	3
lion	36	4	19	1	3	31	3	3
lizard	13	14	21	5	8	30	5	4
lobster	14	12	18	17	11	15	6	7
man	15	19	14	9	18	24	1	0
maple˙tree	11	6	48	29	2	4	0	0

Table 7: **CIFAR100-B Statistics (part 1).** The number of instances that belongs to each background color from each class.

Classes	Green	Gray	Blue	White	Black	Brown	Red	Others
motorcycle	11	28	10	30	2	16	3	0
mountain	0	7	83	6	1	1	2	0
mouse	15	13	13	12	6	34	1	6
mushroom	57	9	5	3	7	19	0	0
oak˙tree	6	2	72	15	1	4	0	0
orange	13	11	17	20	20	6	3	10
orchid	35	4	7	4	35	9	1	5
otter	28	14	27	3	5	22	1	0
palm˙tree	3	7	62	13	5	5	0	5
pear	19	13	10	28	7	20	0	3
pickup˙truck	31	27	16	6	5	14	0	1
pine˙tree	7	11	61	13	1	7	0	0
plain	0	13	70	14	0	3	0	0
plate	3	22	15	26	17	13	2	2
poppy	57	4	7	4	14	4	2	8
porcupine	51	11	8	1	11	17	0	1
possum	27	19	10	1	16	22	2	3
rabbit	34	6	12	2	16	28	2	0
raccoon	30	6	14	5	21	23	1	0
ray	23	11	49	3	6	8	0	0
road	47	7	34	2	2	7	1	0
rocket	7	14	66	5	4	4	0	0
rose	54	3	10	9	11	5	1	7
sea	2	6	74	2	0	12	1	3
seal	17	14	43	4	6	12	1	3
shark	12	3	68	1	13	2	1	0
shrew	30	8	17	5	5	31	3	1
skunk	42	13	5	3	6	30	1	0
skyscraper	1	5	80	9	2	2	0	1
snail	45	7	12	3	5	26	1	1
snake	21	15	13	7	4	30	7	3
spider	37	18	15	2	9	11	7	1
squirrel	36	6	14	5	8	30	0	1
streetcar	22	10	32	11	4	13	7	1
sunflower	44	1	26	9	7	3	2	8
sweet˙pepper	15	7	9	27	14	14	6	8
table	10	17	18	17	7	29	2	0
tank	16	13	30	17	1	23	0	0
telephone	5	11	5	59	4	13	3	0
television	4	24	10	26	10	21	2	3
tiger	45	3	14	8	9	16	0	5
tractor	33	8	34	10	1	14	0	0
train	18	16	42	13	4	7	0	0
trout	20	6	15	28	14	16	1	0
tulip	53	2	11	4	16	13	1	0
turtle	14	9	37	10	6	22	2	0
wardrobe	10	22	10	20	9	20	2	7
whale	1	4	85	7	0	3	0	0
willow˙tree	13	3	43	26	2	5	0	8
wolf	17	10	22	5	22	18	6	0
woman	12	21	16	19	7	15	8	2
worm	13	13	26	8	17	15	7	1

Table 8: **CIFAR100-B Statistics (part 2).** The number of instances that belongs to each background color from each class.