Enhance the Visual Representation via Discrete Adversarial Training

Xiaofeng Mao[†] **Yuefeng Chen**[†] **Ranjie Duan**[†] **Yao Zhu**[‡] **Gege Qi**[†]
Shaokai Ye[§] **Xiaodan Li**[†] **Rong Zhang**[†] **Hui Xue**[†]
[†]Alibaba Group, [‡]Zhejiang University, [§]EPFL
{mxf164419,yuefeng.chenyf,ranjie.drj}@alibaba-inc.com

NeurIPS 2022

Reporter: 王启迅

2022/12/14

Main idea

Using adversarial training (AT) and discrete representation to strengthen OOD robustness.

- Problem of AT (in CV tasks): an adversarial image perturbed in continuous pixel space actually differs with the truly "hard" examples appeared in real world. -> poor OOD generalization
- Converse phenomenon in NLP: AT is benefit for both generalization and robustness. Explanation: text space is discrete and symbolic, where adversarial text is practically existing when a typo is made by humans.

Main idea

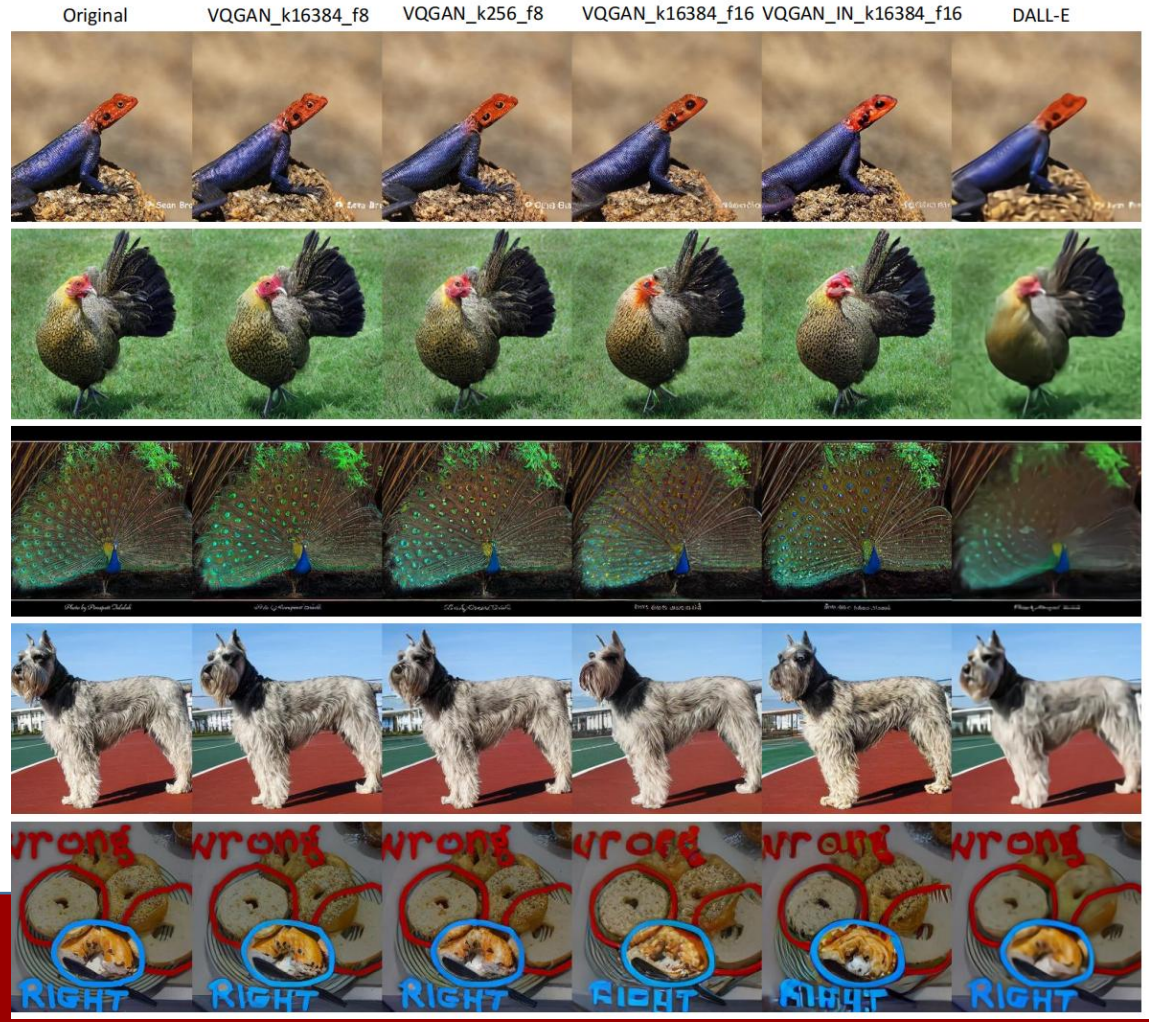
Borrow the symbolic nature of languages, and apply it on CV tasks.

- Method: discretizing continuous images into a more meaningful symbolic space, and conducting AT on such text-like inputs.

- Propose DAT (discrete AT), affects the discretization process to produce diverse adversarial inputs beyond l_p bound for training (while AT works only within a small perturbation radius)

Understanding

Visualization of the reconstructed image from discrete representations:

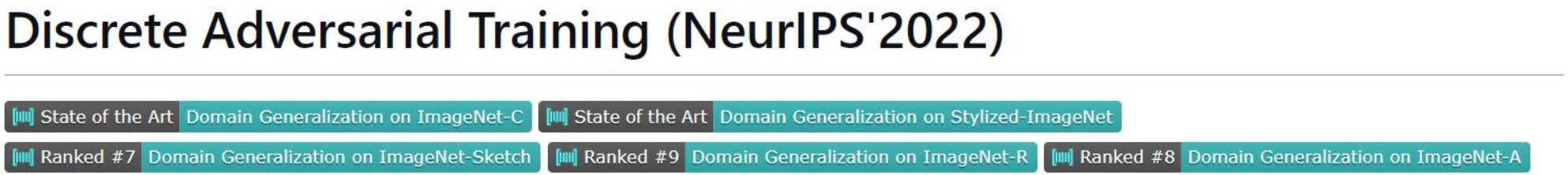


Images augmented by discrete representations maintain global structures, with local features and textures changed.

From the experiments we know this can lead to better OOD generalization.

Contributions

- First transfer the merit of NLP-style adversarial training to vision models.
- Propose DAT, where images are presented as discrete visual words, and the model is training on example which has the adversarially altered discrete visual representation.
- DAT achieves significant improvement on multiple tasks including image classification, object detection and self-supervised learning. (building new robustness recognition SOTA)



Related method

AT (adversarial training)

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\delta} \mathcal{L}(x+\delta, y, \theta) \right] \quad \text{s.t. } \|\delta\|_p < \epsilon$$

- **Remark:** the perturbation is added on pixel level.

Proposed method

DAT (discrete AT)

- **Step 1.** learn an expressive visual codebook and represent the training image set in discrete space.
- Use VQGAN (Esser et al., CVPR 2021) for image discretization

- For continuous input $x \in \mathbb{R}^{H \times W \times 3}$,
- VQGAN learns an encoder $\mathrm{Enc}_\phi(\cdot)$, a decoder $\mathrm{Dec}_\psi(\cdot)$ and quantization $\mathrm{q}_\mathcal{Z}(\cdot)$
- $\mathrm{Enc}_\phi(\cdot)$ maps x to a latent vector $v = \mathrm{Enc}_\phi(x) \in \mathbb{R}^{(h \times w) \times d}$, where h, w are height and width of the feature map, d is the latent dimension)
- $\mathrm{q}_\mathcal{Z}(\cdot)$ learns a codebook $\mathcal{Z} = \{z_k \mid z_k \in \mathbb{R}^d\}_{k=1}^K$ such that each latent vector $v_{ij} \in \mathbb{R}^d$ can be quantized onto its closest codebook entry z_k:

$$v_{\mathbf{q}} = \mathbf{q}_\mathcal{Z}(v) := \left(\underset{z_k \in \mathcal{Z}}{\arg\min} \|v_{ij} - z_k\| \right) \in \mathbb{R}^{h \times w \times d},$$

Proposed method

Then the decoder reconstruct the original input from the discrete codebook item

$$\hat{x} = \mathrm{Dec}_{\psi}(v_{\mathbf{q}})$$

Training of VQGAN:

$$\mathcal{L}_{\mathrm{VQGAN}} = \min_{\mathrm{Enc},\mathrm{Dec},\mathcal{Z}} \max_{D} \mathbb{E}_{x\sim p(x)}[\mathcal{L}_{\mathrm{VQ}}(\mathrm{Enc}, \mathrm{Dec}, \mathcal{Z}) + \mathcal{L}_{\mathrm{GAN}}(\{\mathrm{Enc}, \mathrm{Dec}, \mathcal{Z}\}, D)]$$

$$\mathcal{L}_{\mathrm{VQ}}(\mathrm{Enc}, \mathrm{Dec}, \mathcal{Z}) = \|x - \hat{x}\|_{\mathrm{precept}} + \|\mathrm{sg}[\mathrm{Enc}(x)] - v_{\mathbf{q}}\|_2^2 + \|\mathrm{sg}[v_{\mathbf{q}}] - \mathrm{Enc}(x)\|_2^2$$

$$\mathcal{L}_{\mathrm{GAN}}(\{\mathrm{Enc}, \mathrm{Dec}, \mathcal{Z}\}, D) = [\log D(x) + \log(1 - D(\hat{x}))]$$

Remark:

1. $\|x - \hat{x}\|_{\mathrm{precept}}$ is the perceptual reconstruction loss instead of L2 loss
2. $\|\mathrm{sg}[\mathrm{Enc}(x)] - v_{\mathbf{q}}\|_2^2$ makes the codebook entry moves towards the output of the encoder
3. $\|\mathrm{sg}[v_{\mathbf{q}}] - \mathrm{Enc}(x)\|_2^2$ makes the encoder moves towards the codebook
4. $\mathcal{L}_{\mathrm{GAN}}(\{\mathrm{Enc}, \mathrm{Dec}, \mathcal{Z}\}, D) = [\log D(x) + \log(1 - D(\hat{x}))]$ is the GAN loss

Proposed method

The above image discretization process is denoted as $\hat{x} = \mathcal{Q}(x)$

Introducing AT into this discretization process:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\delta} \mathcal{L}(\mathcal{Q}(x+\delta), y, \theta) \right] \qquad \delta \simeq \alpha \nabla_x \mathcal{L}(\mathcal{Q}(x), y, \theta)$$

How to update δ:

$$\nabla_x \mathcal{L}(\mathcal{Q}(x), y, \theta) = \frac{\partial \mathcal{L}}{\partial \hat{x}} \cdot \frac{\partial \hat{x}}{\partial v_{\mathrm{q}}} \cdot \frac{\partial v_{\mathrm{q}}}{\partial v} \cdot \frac{\partial v}{\partial x}$$

Problem: $\frac{\partial v_{\mathrm{q}}}{\partial v}$ is hard to solve, because $v_{\mathbf{q}} = \mathbf{q}_{\mathcal{Z}}(v) := \left(\arg\min_{z_k \in \mathcal{Z}} \|v_{ij} - z_k\| \right) \in \mathbb{R}^{h \times w \times d}$, is non-differentiable

Straight-through estimator: $\nabla_x \mathcal{L}(\mathcal{Q}(x), y, \theta) = \frac{\partial \mathcal{L}}{\partial \hat{x}} \cdot \frac{\partial \hat{x}}{\partial x} \simeq \frac{\partial \mathcal{L}}{\partial \hat{x}}$ since $\hat{x} \simeq x$

Algorithm

Algorithm 1: Pseudo code of DAT

Input: Classifier F; Pre-trained discretizer $\mathcal{Q}$; A sampled mini-batch of clean images x with labels y; attack magnitude α.

Output: Learned network parameter θ of F

1: Fix the network parameters of $\mathcal{Q}$
2: **for** each training steps **do**
3: $\hat{x} \leftarrow \mathcal{Q}(x)$ //Get the discrete reconstruction $\hat{x}$
4: $\delta \leftarrow \alpha \nabla_{\hat{x}} \mathcal{L}(\hat{x}, y, \theta)$ //Estimate the adversarial perturbations
5: $x_{adv} \leftarrow \mathcal{Q}(x + \delta)$ //Generate discrete adversarial examples
6: Minimize the classification loss w.r.t. network parameter $\arg\min_\theta \mathcal{L}(x_{adv}, y, \theta)$
7: **end for**

Overall pipeline

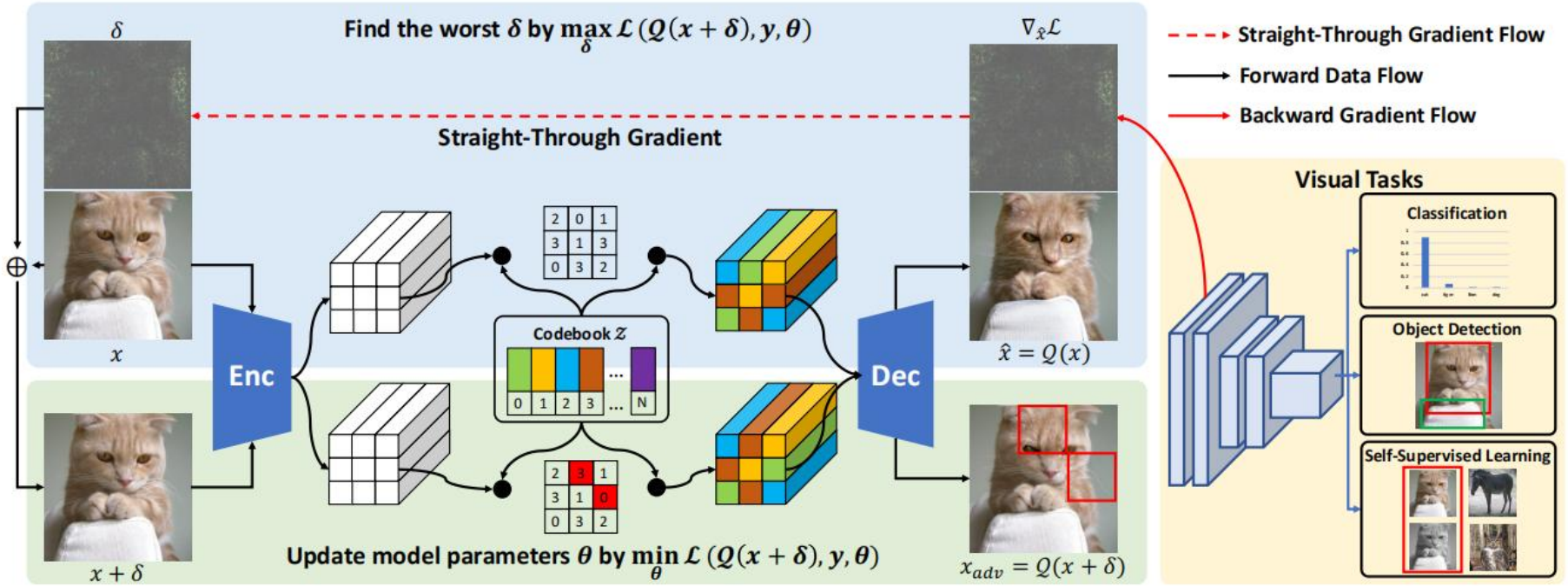


Figure 1: The overall pipeline of Discrete Adversarial Training (DAT).

Experiments

Image classification

- Improve both adversarial robustness and OOD generalization

Methods	ImageNet	Adversarial Robustness		Out of Distribution Robustness					
		FGSM	DamageNet	A	C↓	V2	R	Sketch	Stylized
ResNet50 [46]	76.13	12.19	5.94	0.0	76.70	63.20	36.17	24.09	7.38
+ DAT (Ours)	**76.52**	**30.66**	**14.42**	**4.38**	**74.16**	**65.02**	**41.90**	**27.27**	**10.8**
DeepAugment [47]	76.66	21.61	11.94	3.46	60.37	**65.24**	42.17	29.50	14.68
+ Augmix [34]	75.82	27.05	19.60	3.86	53.55	63.63	46.77	32.62	21.23
+ DAT (Ours)	**77.10**	**35.32**	**22.86**	**6.86**	**50.82**	65.14	**47.88**	**34.98**	**21.89**
ViT [20]	72.00	23.30	28.99	6.44	77.61	57.34	25.69	15.56	5.82
DrViT [41]	79.48	45.76	44.91	17.20	46.22	68.05	44.77	34.59	19.38
AugReg-ViT [48]	79.91	44.32	45.24	19.03	54.50	67.90	39.46	29.16	16.62
+ DAT (Ours)	**81.46**	**51.82**	**45.70**	**30.15**	**44.65**	**70.83**	**47.34**	**34.77**	**23.13**
MAE-H [19]	86.90	60.16	64.36	68.18	33.92	78.47	64.12	49.08	26.36
+ DAT (Ours)	**87.02**	**63.77**	**70.42**	**68.92**	**31.40**	**78.82**	**65.61**	**50.03**	**32.77**

Experiments

Image classification (comparing with other algorithms)

Training Strategies	ImageNet	Adversarial Robustness		Out of Distribution Robustness					
		FGSM	DamageNet	A	C↓	V2	R	Sketch	Stylized
Normal [48]	79.91	44.32	45.24	19.03	54.50	67.90	39.46	29.16	16.62
Advprop [26]	79.54	**72.38**	45.48	18.53	51.46	68.74	43.51	31.68	19.24
Fast Advprop [27]	79.02	70.52	44.87	17.86	53.31	67.09	41.84	29.42	18.39
Pyramid AT [28]	**81.68**	50.36	45.53	23.18	44.95	70.32	47.30	**36.87**	20.02
Debiased [57]	79.33	46.85	44.99	18.32	49.82	67.55	40.32	29.43	22.37
DAT (Ours)	81.46	51.82	**45.70**	**30.15**	**44.65**	**70.83**	**47.34**	34.77	**23.13**

Table 2: Comparison of DAT with other training strategies. We use AugReg-ViT as the base model.

Training Strategies	ImageNet	Adversarial Robustness		Out of Distribution Robustness					
		FGSM	DamageNet	A	C↓	V2	R	Sketch	Stylized
Normal [48]	76.13	12.19	5.94	0	76.70	63.2	36.17	24.09	7.38
Advprop [26]	**77.59**	28.65	**15.58**	4.33	70.53	**65.47**	38.75	25.51	7.99
Fast Advprop [27]	76.6	17.33	7.45	2.19	73.31	64.24	38.17	25.03	8.3
Pyramid AT [28]	75.46	30.35	14.22	3.01	76.42	62.46	38.85	23.76	10.41
Debiased [57]	76.91	20.4	6.66	3.51	**67.55**	65.04	40.8	**28.42**	**17.4**
DAT (Ours)	76.52	**30.66**	14.42	**4.38**	74.16	65.02	**41.9**	27.27	10.8

Table 6: Comparison of DAT with other training strategies. We use ResNet50 as the base model.

Remark: DAT achieves better performances on ViT than on CNN. (On CNN, Debiased seem to be the best.

Understanding

Why DAT is better than AT on OOD?

1. The adversarial image produced by DAT is more close to the clean sample than that produced by AT.

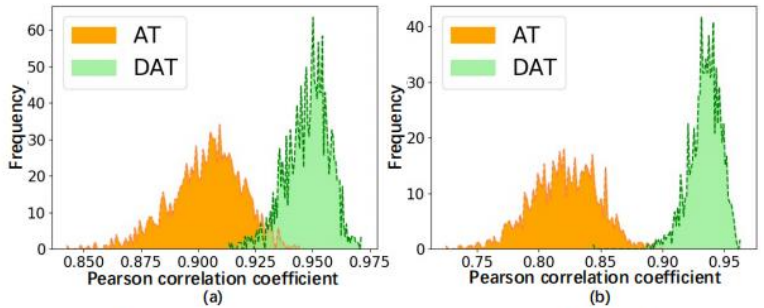


Figure 2: The frequency histogram of the Pearson correlation coefficient (PCC) between BN statistics on clean and adversarial images. Larger PCC value means smaller distributional difference with clean images. (a), (b) present the difference on mean and variance statistics respectively.

Understanding

Why DAT is better than AT on OOD?

2. Comparing to AT, DAT avoid introducing high-frequency, meaningless noise.

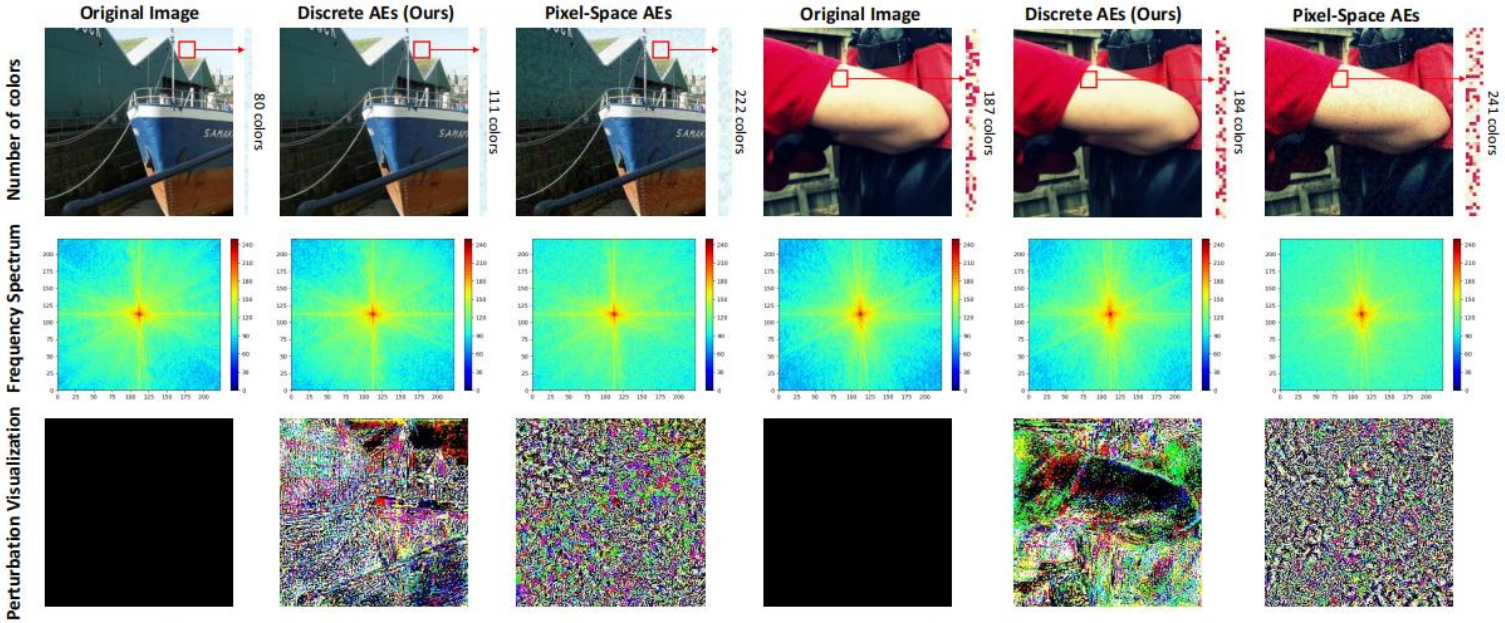
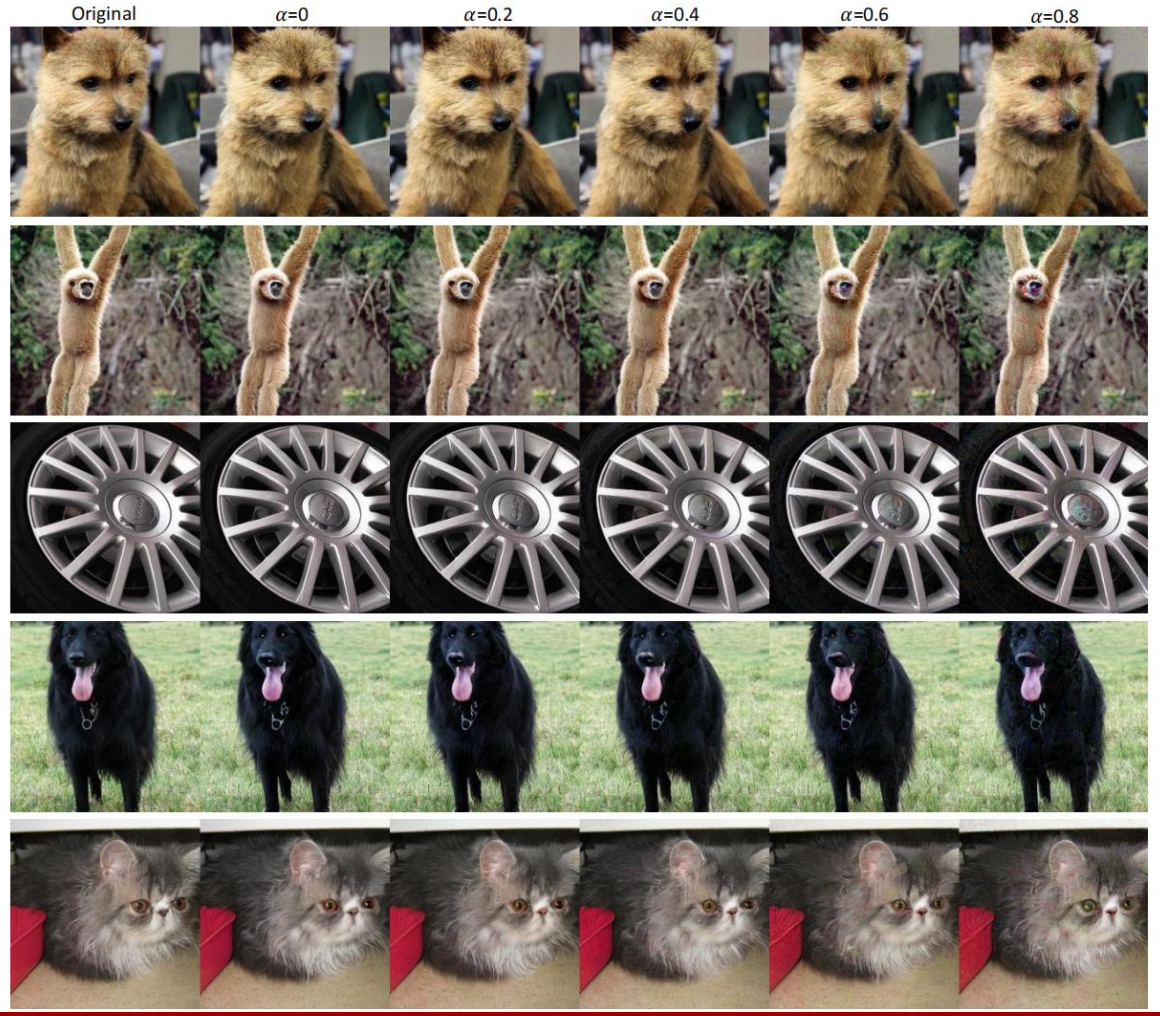


Figure 3: Comparison of discrete perturbations and pixel-space perturbations.

Understanding

Why DAT is better than AT on OOD?

2. Comparing to AT, DAT avoid introducing high-frequency, meaningless noise.

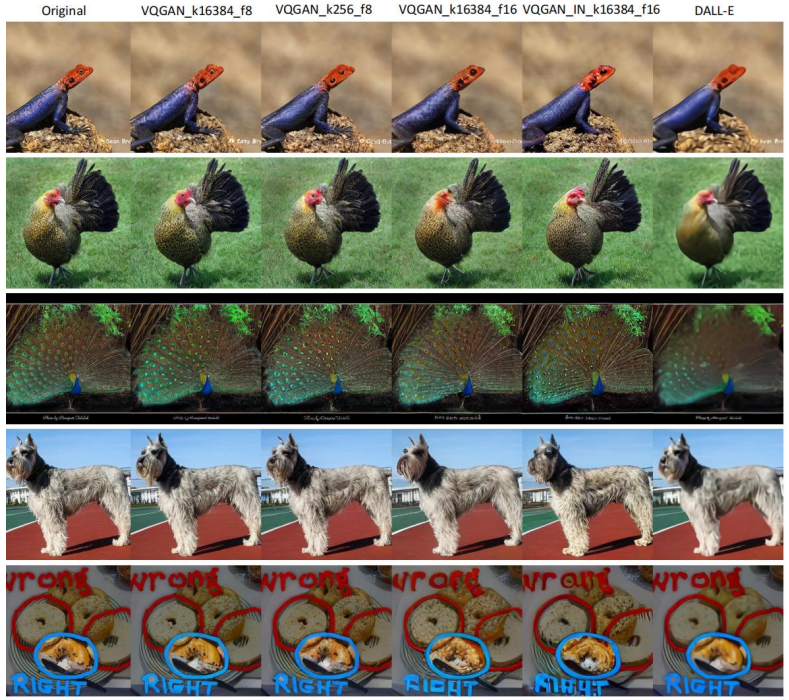


Samples reconstructed by DAT.

- These samples changes local features.
- The changes are far beyond the l_p-ball constraint.
- This seems to cause better OOD generalization

Understanding

Discrete representation as data augmentation itself is good for OOD



Training Strategies	ImageNet	Adversarial Robustness FGSM	Adversarial Robustness DamageNet	Out of Distribution Robustness A	Out of Distribution Robustness C↓	Out of Distribution Robustness V2	Out of Distribution Robustness R	Out of Distribution Robustness Sketch	Out of Distribution Robustness Stylized
Normal [48]	76.13	12.19	5.94	0	76.70	63.2	36.17	24.09	7.38

Types	α	Modified Codes	ImageNet	Adversarial Robustness FGSM	Adversarial Robustness DamageNet	Out of Distribution Robustness A	Out of Distribution Robustness C↓	Out of Distribution Robustness V2	Out of Distribution Robustness R	Out of Distribution Robustness Sketch	Out of Distribution Robustness Stylized
Random	-	3.8%	76.47	29.01	10.7	3.00	74.71	64.75	40.19	26.17	9.89
Adv.	0.0	0.0%	76.38	23.94	9.12	3.2	76.31	64.71	38.41	24.62	8.77
Adv.	0.1	3.8%	76.52	30.66	14.42	4.38	74.16	65.02	41.9	27.27	10.8
Adv.	0.2	7%	75.93	34.47	15.21	3.11	75.09	64.38	40.27	26.33	10.14
Adv.	0.4	13%	74.28	36.2	17.76	1.96	77.25	62.5	38.75	24.31	8.61

Table 8: DAT with different perturbations. We use ResNet50 as the base model.