Appendix: Towards Viewpoint-Invariant Visual Recognition via Adversarial Training

Shouwei Ruan[1], Yinpeng Dong[2,3], Hang Su[2,4,5*], Jianteng Peng[6], Ning Chen[2], Xingxing Wei[1*]
[1] Institute of Artificial Intelligence, Beihang University, Beijing 100191, China
[2] Dept. of Comp. Sci. and Tech., Institute for AI, Tsinghua-Bosch Joint ML Center,
THBI Lab, BNRist Center, Tsinghua University, Beijing 100084, China
[3] RealAI [4] Peng Cheng Laboratory [5] Pazhou Laboratory (Huangpu), Guangzhou, China [6] OPPO
{shouweiruan,xxwei}@buaa.edu.cn, {dongyinpeng,suhangss,ningchen}@tsinghua.edu.cn, pengjianteng@oppo.com

A. Overview

In this Appendix, we first summarize our main contributions as follows:

- We propose VIAT, the first framework for enhancing the viewpoint robustness of visual recognition models via adversarial training. Unlike previous methods that rely on data augmentation[2] or incorporate extra regularizers[5], VIAT enhances viewpoint robustness by minimizing the model's loss expectation over the worst-case adversarial distribution of viewpoints. It does not require 3D object information and only uses multi-view images as input, enabling it to be trained using real-world data.

- We contribute **GMVFool**, an efficient viewpoint attack method that optimizes the Gaussian mixture distribution of adversarial viewpoints through multi-view images, which can capture a diversity of adversarial viewpoints simultaneously.

- we contribute a multi-view dataset—**IM3D**, which contains 1k typical synthetic 3D objects from 100 ImageNet classes and have realistic viewpoint images.

- We further construct a new benchmark for viewpoint robustness—**ImageNet-V+**, containing 100K images from the adversarial viewpoints, which is 10× larger than the previous ImageNet-V[1]. We hope to serve it as a standard out-of-distribution (OOD) benchmark for evaluating viewpoint robustness in the future.

Then, we provide the formulas proofs, additional experiments, and datasets introduction to the main paper. Specifically, Sec.B, we give a detailed derivation of Eq. (6), in Sec.C we present more experimental findings, in Sec.D, we introduce the IM3D dataset used by VIAT, and in Sec.E we introduce our proposed ImageNet-V+ benchmark.

B. Proof of Eq. (6)

In this section, we will detail derive the gradient of the internal maximization objective in Eq. (5) w.r.t the parameters of the Gaussian mixture distribution.

For the first term of Eq. (5), i.e., the expectation of classification loss $\mathcal{L}_1 = \mathbb{E}_{p(\mathbf{u},\boldsymbol{\Gamma}|\Psi)}\big[\mathcal{L}(f_{\mathbf{W}}(\mathcal{R}(\mathbf{a}\cdot\tanh(\mathbf{u})+\mathbf{b})),y)\big]$, we first calculate its search gradient:

$$\nabla_{\omega_k}\log p(\mathbf{u},\boldsymbol{\Gamma}|\Psi) = \gamma_k; \tag{B.1}$$

$$\nabla_{\boldsymbol{\mu}_k}\log p(\mathbf{u},\boldsymbol{\Gamma}|\Psi) = \frac{\mathbf{u}-\boldsymbol{\mu}_k}{\boldsymbol{\sigma}_k^2}\gamma_k = \frac{\boldsymbol{r}}{\boldsymbol{\sigma}_k}\gamma_k; \tag{B.2}$$

$$\nabla_{\boldsymbol{\sigma}_k}\log p(\mathbf{u},\boldsymbol{\Gamma}|\Psi) = \frac{(\mathbf{u}-\boldsymbol{\mu}_k)^2-\boldsymbol{\sigma}_k^2}{\boldsymbol{\sigma}_k^3}\gamma_k = \frac{\boldsymbol{r}^2-1}{\boldsymbol{\sigma}_k}\gamma_k, \tag{B.3}$$

where $\boldsymbol{r}$ follows the standard Gaussian distribution $\mathcal{N}(\mathbf{0},\mathbf{I})$. To approximate a more realistic gradient, we follow the suggestion in the **Natural Evolution Strategy** (NES)[4] to further compute the natural gradient, which is defined as:

$$\widetilde{\nabla}_{\omega_k,\boldsymbol{\mu}_k,\boldsymbol{\sigma}_k}\mathcal{L}_1 = \mathbf{F}^{-1}\nabla_{\omega_k,\boldsymbol{\mu}_k,\boldsymbol{\sigma}_k}\mathcal{L}_1, \tag{B.4}$$

where $\mathbf{F}$ is the Fisher information matrix as:

$$\begin{aligned}\mathbf{F} = \mathbb{E}_{p(\mathbf{u},\boldsymbol{\Gamma}|\Psi)}\big[&\nabla_{\omega_k,\boldsymbol{\mu}_k,\boldsymbol{\sigma}_k}\log p(\mathbf{u},\boldsymbol{\Gamma}|\Psi)\\ &\cdot\nabla_{\omega_k,\boldsymbol{\mu}_k,\boldsymbol{\sigma}_k}\log p(\mathbf{u},\boldsymbol{\Gamma}|\Psi)^{\top}\big],\end{aligned} \tag{B.5}$$

by further derivation, we can obtain:

$$\mathbf{F}_{\omega_k} = \omega_k; \tag{B.6}$$

$$\mathbf{F}_{\boldsymbol{\mu}_k} = \frac{\omega_k}{\boldsymbol{\sigma}_k^2}; \tag{B.7}$$

$$\mathbf{F}_{\boldsymbol{\sigma}_k} = \frac{2\cdot\omega_k}{\boldsymbol{\sigma}_k^2}. \tag{B.8}$$

[*] Corresponding author.

Therefore we can derive that:

$$\widetilde{\nabla}_{\boldsymbol{\omega}_k} \log p(\mathbf{u}, \boldsymbol{\Gamma}|\Psi) = \mathbf{F}_{\omega_k}^{-1}\gamma_k = \frac{\gamma_k}{\omega_k}; \tag{B.9}$$

$$\widetilde{\nabla}_{\boldsymbol{\mu}_k} \log p(\mathbf{u}, \boldsymbol{\Gamma}|\Psi) = \mathbf{F}_{\boldsymbol{\mu}_k}^{-1}\frac{\boldsymbol{r}}{\boldsymbol{\sigma}_k}\gamma_k = \frac{\boldsymbol{\sigma}_k\mathbf{r}}{\omega_k}\gamma_k; \tag{B.10}$$

$$\widetilde{\nabla}_{\boldsymbol{\sigma}_k} \log p(\mathbf{u}, \boldsymbol{\Gamma}|\Psi) = \mathbf{F}_{\boldsymbol{\sigma}_k}^{-1}\frac{\boldsymbol{r}^2 - 1}{\boldsymbol{\sigma}_k}\gamma_k = \frac{\boldsymbol{\sigma}_k(\mathbf{r}-1)^2}{2\omega_k}\gamma_k. \tag{B.11}$$

By integrating Eq. (B.9), Eq. (B.10) and Eq. (B.11) into Eq. (5), we end up with the natural gradient about $\mathcal{L}_1$:

$$\nabla_{\omega_k}\mathcal{L}_1 = \mathbb{E}_{\mathcal{N}(\mathbf{r}|\mathbf{0},\mathbf{I})}\big[\mathcal{L}_{\text{cls}} \cdot \frac{\gamma_k}{\omega_k}\big]; \tag{B.12}$$

$$\nabla_{\boldsymbol{\mu}_k}\mathcal{L}_1 = \mathbb{E}_{\mathcal{N}(\mathbf{r}|\mathbf{0},\mathbf{I})}\big[\mathcal{L}_{\text{cls}} \cdot \frac{\boldsymbol{\sigma}_k\mathbf{r}}{\omega_k}\gamma_k\big]; \tag{B.13}$$

$$\nabla_{\boldsymbol{\sigma}_k}\mathcal{L}_1 = \mathbb{E}_{\mathcal{N}(\mathbf{r}|\mathbf{0},\mathbf{I})}\big[\mathcal{L}_{\text{cls}} \cdot \frac{\boldsymbol{\sigma}_k(\mathbf{r}^2-1)}{2\omega_k}\gamma_k\big]; \tag{B.14}$$

$$\mathcal{L}_{\text{cls}} = \mathcal{L}(f_{\mathbf{W}}(\mathcal{R}(\mathbf{a} \cdot \tanh(\prod_{k=1}^{K}\boldsymbol{\mu}_k^{\gamma_k} + \prod_{k=1}^{K}\boldsymbol{\sigma}_k^{\gamma_k} \cdot \mathbf{r}) + \mathbf{b})), y). \tag{B.15}$$

For the second term of Eq. (5), i.e., the expectation of entropy regularized loss $\mathcal{H} = \mathbb{E}_{p(\mathbf{u},\boldsymbol{\Gamma}|\Psi)}\big[-\log p(\boldsymbol{a}\cdot\tanh(\mathbf{u}) + \boldsymbol{b})\big]$, we first perform the transformation using the random variable approach, rewriting $\mathcal{H}$ as

$$\mathcal{H} = \mathbb{E}_{p(\mathbf{u},\boldsymbol{\Gamma}|\Psi)}\big[-\log p(\boldsymbol{a}\cdot\tanh(\prod_{k=1}^{K}\boldsymbol{\mu}_k^{\gamma_k} + \prod_{k=1}^{K}\boldsymbol{\sigma}_k^{\gamma_k}\cdot\mathbf{r}) + \boldsymbol{b})\big]. \tag{B.16}$$

Next, the log density of the distribution can be analytically calculated as follows. Note that the dimensions of the random variables are independent of each other, we consider here the case of one dimension. For the random variable r, its probability density is $p(r) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{r^2}{2})$. The probability density of u is $p(u) = \prod_{k=1}^{K}\omega_k^{\gamma_k}(\frac{1}{\sqrt{2\pi}\sigma_k}\exp(-\frac{r^2}{2}))^{\gamma_k}$. For the viewpoint $v = a \cdot \tanh(u) + b$, its inversion is $u = \tanh^{-1}(\frac{v-b}{a}) = \frac{1}{2}(\frac{a+v-b}{a-v+b})$, The derivative of u w.r.t. v is $\frac{\mathrm{d}u}{\mathrm{d}v} = \frac{1}{a(1-\tanh(u)^2)}$. By applying the transformation of variable approach, we can derive the probability density of v as:

$$p(v) = \prod_{k=1}^{K}(\omega_k\frac{1}{\sqrt{2\pi}\sigma_k}\exp(-\frac{r^2}{2}))^{\gamma_k} \cdot \frac{1}{a(1-\tanh(\prod_{k=1}^{K}\mu_k^{\gamma_k} + \prod_{k=1}^{K}\sigma_k^{\gamma_k}\cdot r)^2)}. \tag{B.17}$$

Therefore, the negative log-likelihood of v is

$$-\log p(v) = \sum_{k=1}^{K}\gamma_k\Big[-\omega_k + \frac{r^2}{2} + \frac{\log(2\pi)}{2} + \log\sigma_k + \log(1-\tanh(\mu_k+\sigma_k r)^2) + \log a\Big]. \tag{B.18}$$

Sum over all dimensions into Eq.(B.16), we can simply calculate the gradients of $\mathcal{H}$ w.r.t. ω_k, $\boldsymbol{\mu}_k$, and $\boldsymbol{\sigma}_k$ as:

$$\nabla_{\omega_k}\mathcal{H} = \mathbb{E}_{\mathcal{N}(\mathbf{r}|\mathbf{0},\mathbf{I})}[-\gamma_k]; \tag{B.19}$$

$$\nabla_{\boldsymbol{\mu}_k}\mathcal{H} = \mathbb{E}_{\mathcal{N}(\mathbf{r}|\mathbf{0},\mathbf{I})}\big[-2\gamma_k\tanh(\boldsymbol{\mu}_k + \boldsymbol{\sigma}_k\mathbf{r})\big]; \tag{B.20}$$

$$\nabla_{\boldsymbol{\sigma}_k}\mathcal{H} = \mathbb{E}_{\mathcal{N}(\mathbf{r}|\mathbf{0},\mathbf{I})}\big[\gamma_k\frac{1-2\mathbf{r}\tanh(\boldsymbol{\mu}_k+\boldsymbol{\sigma}_k\mathbf{r})\boldsymbol{\sigma}_k}{\boldsymbol{\sigma}_k}\big]. \tag{B.21}$$

Finally, we can obtain the gradient in Eq. (6) by combining Eq.(B.12), Eq.(B.19) ; Eq.(B.13), Eq.(B.20); and Eq.(B.14), Eq.(B.21).

C. Additional Experimental Result

C.1. Evaluation for More VIAT-Trained Models

We utilize our VIAT to perform viewpoint invariance enhancement for more models with different structures and use the same experimental setup as Sec.4.1. The evaluation results of each model in ImageNet-V+ are shown in Fig. C.1

C.2. More Prediction examples

Fig.C.2 shows the the prediction of standard-trained and VIAT-trained ResNet-50 on the ObjectNet dataset. It can be found that the standard-trained model is prone to prediction errors for some uncommon viewpoint images, while it can predict successfully using the VIAT-trained model.

C.3. Visualization of the adversarial viewpoints

We compare the adversarial viewpoint images generated by the two attack method, ViewFool and our GMVFool, and Fig. C.3 displays the viewpoint images sampled and rendered from the optimum adversarial viewpoint distribution. It can be discovered that GMVFool is able to capture a wider variety of adversarial viewpoints.

D. The IM3D Dataset

D.1. Synthetic 3D Objects

IM3D consists of 1000 synthetic 3D objects belonging to 100 classes in ImageNet and are grouped into 7 superclasses. We use blender to render multi-view images from the upper hemisphere for each 3D object, as well as to obtain the corresponding camera poses for each image. Table.D.1 shows the category information of these objects. Fig. D.1 shows the visualization results of a part of the objects in the 3D dataset that we created. As one of the contributions of our work, we will make this 3D dataset and multi-view image data publicly available in the future.

D.2. Instant-NGP Result

The first step of our method is to train the Instant-NGP[3] for each object using multi-view images, to obtain the 3D

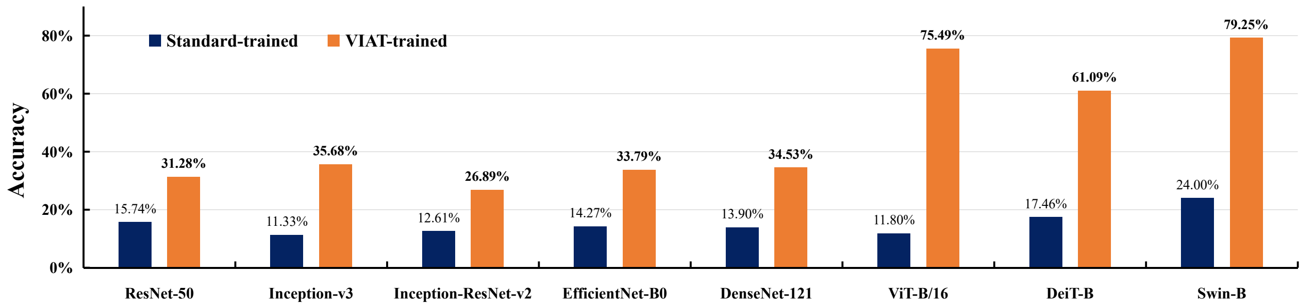

Figure C.1. The accuracy of standard-trained and VIAT-trained models under ImageNet-V+ dataset

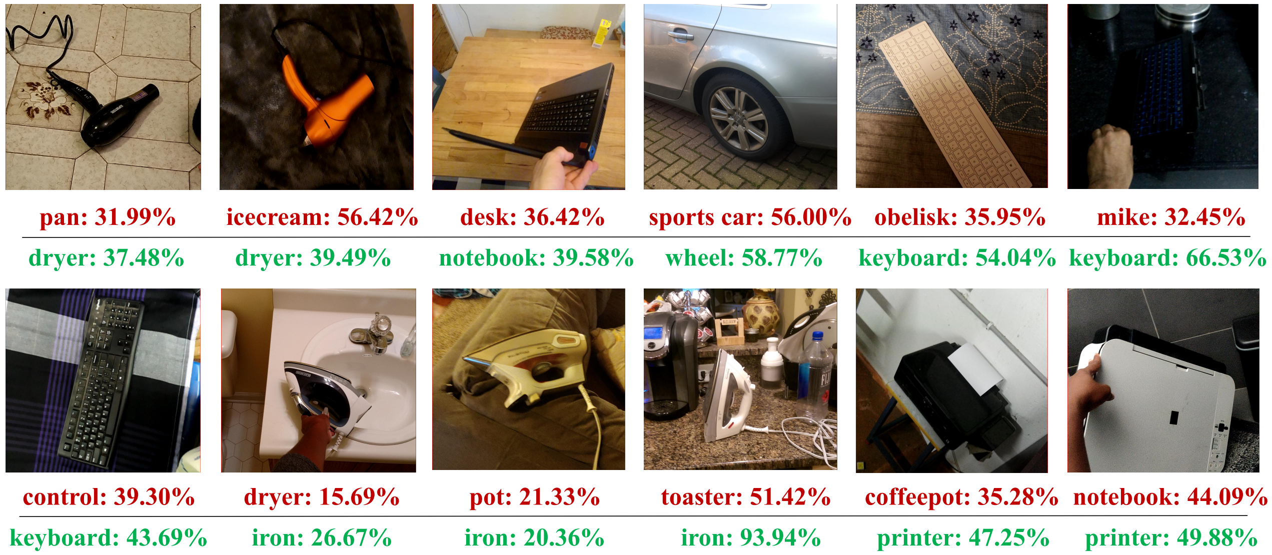

Figure C.2. The prediction of ResNet-50 (standard-trained) and ResNet-50 (VIAT-trained) under the ObjectNet dataset.

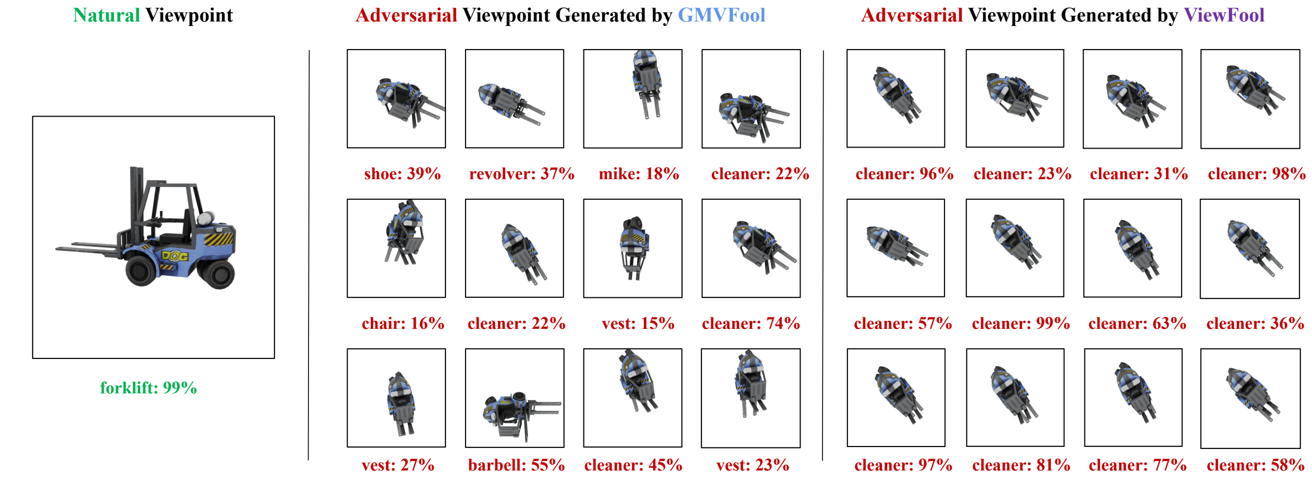

Figure C.3. The comparison of adversarial viewpoints generated using GMVFool and ViewFool, and prediction results from ResNet-50.

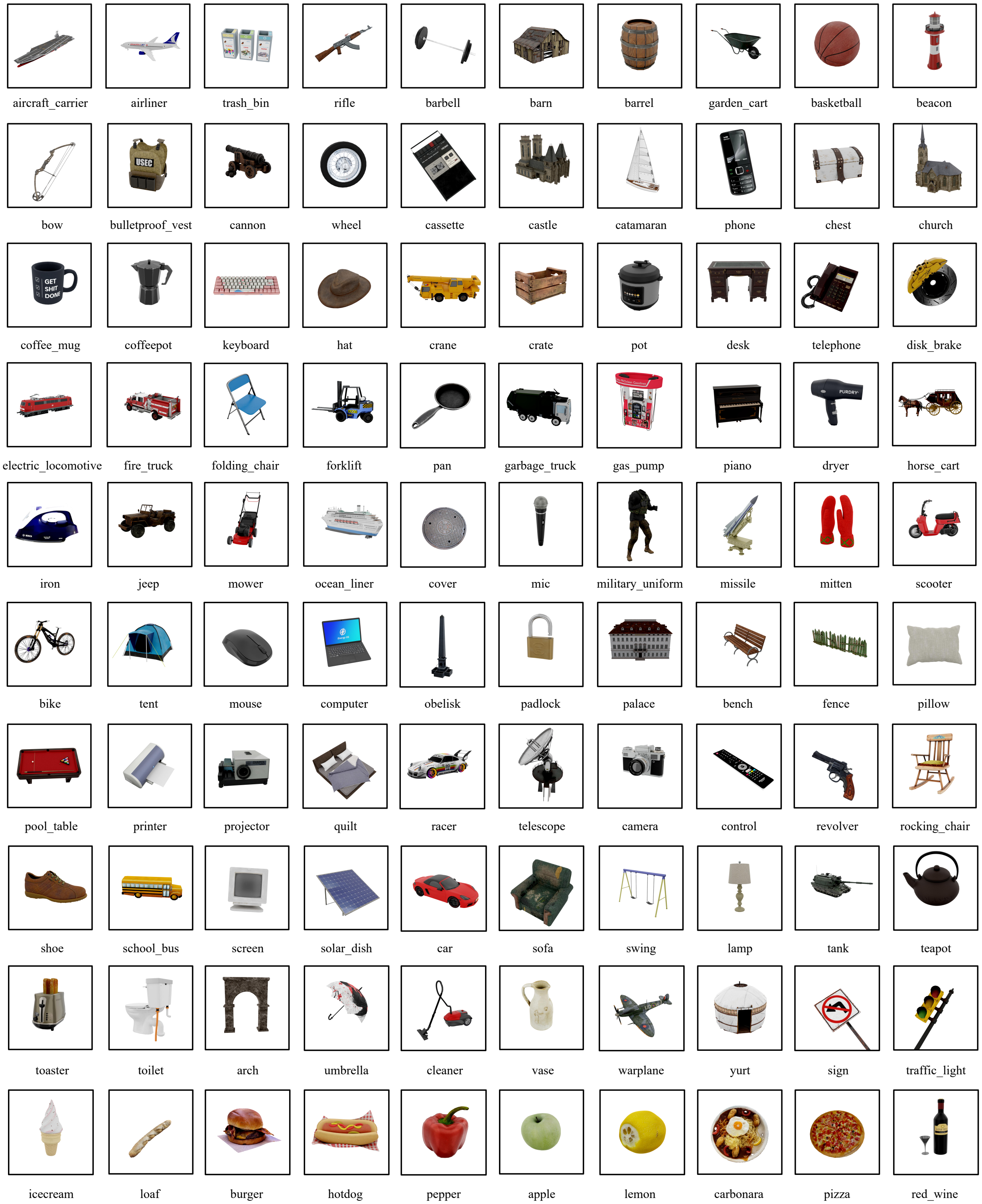

Figure D.1. Visualization for a portion of the objects in our dataset.

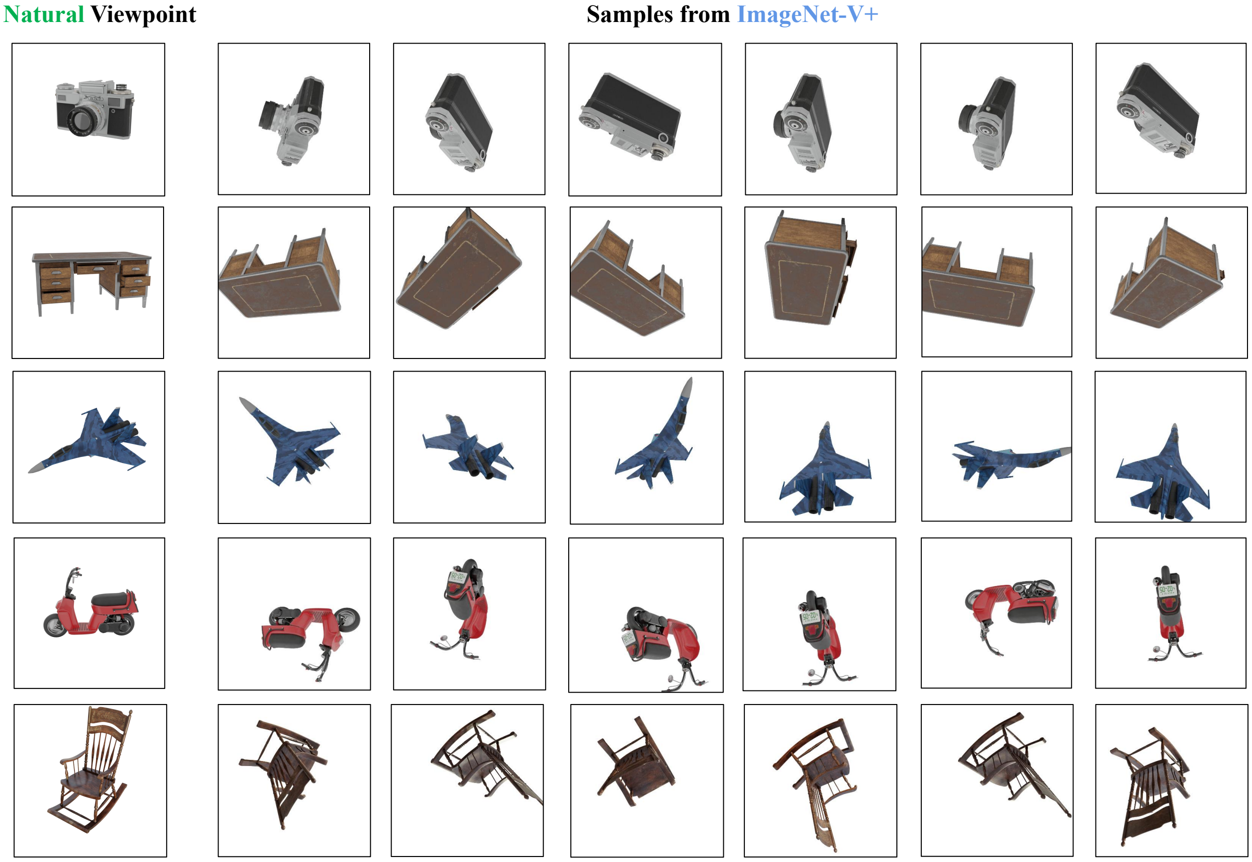

Figure E.1. Sampled images from the ImageNet-V+ dataset.

representations. Therefore, the reconstruction quality of Instant-NGP is crucial to our method. We use the Peak Signal to Noise Ratio (PSNR) to measure the reconstruction quality, a higher PSNR means that the less gap between the rendering of the object and the ground-truth image. We provide the average PSNR metrics for each category (each category contains 10 different objects) in Table. D.1.

E. More about ImageNet-V+

ImageNet-V+ consists of 100,000 images of 1,000 different objects. For each object, we adopt 100 images from varying viewpoints sampled from the adversarial distributions. we generate diversity viewpoints by GMVFool, and Fig. E.1 shows some samples in ImageNet-V+. We can observe that the sampled images for the same object are very different from each other, thus the diversity of ImageNet-V+ is improved.

We will make the ImageNet-V+ dataset publicly available and hope that it will contribute to future research on viewpoint robustness for visual recognition.

References

[1] Yinpeng Dong, Shouwei Ruan, Hang Su, Caixin Kang, Xingxing Wei, and Jun Zhu. Viewfool: Evaluating the robustness of visual recognition to adversarial viewpoints. In *Neural Information Processing Systems (NeurIPS)*, 2022. 1

[2] Spandan Madan, Timothy Henry, Jamell Dozier, Helen Ho, Nishchal Bhandari, Tomotake Sasaki, Frédo Durand, Hanspeter Pfister, and Xavier Boix. On the capability of neural networks to generalize to unseen category-pose combinations. Technical report, Center for Brains, Minds and Machines (CBMM), 2020. 1

[3] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, July 2022. 2

[4] Daan Wierstra, Tom Schaul, Tobias Glasmachers, Yi Sun, Jan Peters, and Jürgen Schmidhuber. Natural evolution strategies. *Journal of Machine Learning Research*, 15(1):949–980, 2014. 1

[5] Fanny Yang, Zuowen Wang, and Christina Heinze-Deml. Invariance-inducing regularization using worst-case transformations suffices to boost accuracy and spatial robustness. *Advances in Neural Information Processing Systems*, 32, 2019. 1

Superclass	Label	Average PSNR	Superclass	Label	Average PSNR
Traffic	airliner	47.59	Indoors	quilt	44.76
	garden cart	42.24		control	45.07
	wheel	42.29		rocking chair	42.62
	catamaran	43.55		shoe	44.12
	crane	40.54		lamp	42.63
	disk brake	41.06		teapot	44.16
	electric locomotive	41.73		toaster	41.64
	fire truck	40.90		toilet	46.16
	forklift	40.21		cleaner	40.77
	garbage truck	41.37		vase	43.78
	horse cart	38.68	Outdoors	trash bin	43.11
	jeep	40.72		basketball	43.08
	ocean liner	41.22		mower	42.10
	scooter	38.67		cover	46.74
	bike	38.85		tent	43.45
	racer	39.82		bench	42.21
	school bus	43.81		fence	42.49
	car	42.22		swing	38.58
	street sign	32.90		umbrella	44.05
	traffic light	28.82		gas pump	43.49
Military	aircraft carrier	41.68	Building	barn	43.28
	rifle	36.55		beacon	42.00
	bow	32.13		castle	41.14
	bulletproof vest	41.96		church	42.36
	cannon	40.81		obelisk	42.10
	military uniform	42.78		palace	36.09
	missile	42.02		telescope	37.10
	revolver	40.59		solar dish	42.32
	tank	39.34		arch	40.26
	warplane	41.84		yurt	36.21
Indoors	barbell	40.55	Electronic	cassette	42.39
	barrel	42.37		phone	44.58
	chest	42.68		keyboard	44.54
	coffee mug	45.58		mic	39.12
	coffeepot	45.21		mouse	44.41
	hat	44.11		computer	44.16
	crate	42.47		printer	43.41
	pot	43.73		projector	36.90
	desk	43.36		camera	41.24
	telephone	37.80		screen	45.12
	folding chair	45.45	Food	hotdog	46.07
	pan	44.66		burger	43.12
	piano	44.19		lemon	47.44
	dryer	44.26		apple	48.85
	iron	43.82		loaf	42.32
	mitten	46.94		carbonara	43.84
	padlock	42.11		pizza	48.19
	sofa	43.89		red wine	39.94
	pillow	45.50		pepper	38.75
	pool table	43.42		icecream	39.70

Table D.1. The categories of the synthetic 3D dataset we construct and the results of Instant-NGP rendering.