## A Societal Impacts

Our paper is closely related to Responsible AI (RAI), especially in enabling qualitative assessments of models. Our approach provides visual and intuitive explanations of a model's decision-making criteria, offering insights that are both explanatory and responsible. Our approach of utilizing DSVs for RAI enables global explanations, surpassing traditional Explainable AI (XAI) methodologies, which usually focus on local explanations for individual inputs and cannot provide a global decision criterion. Furthermore, since our method is based on model inversion, it ensures safety and privacy. While the synthesized sets in Fig. 7 might appear similar to the selected sets, they do not replicate specific sample features. This is because DSVs represent a more generalized decision boundary, avoiding the inclusion of image-specific features. Consequently, DSVs enable all models using logistic loss to be more responsible.
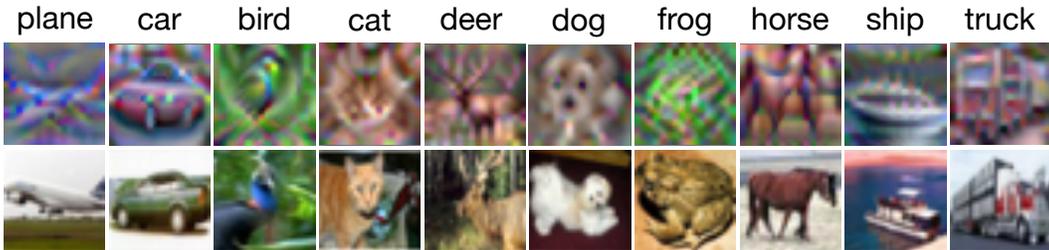
Figure 7: Comparison of synthesized images (first row) created using the DeepKKT condition initiated from noise, and selected images (second row) from the CIFAR-10 training dataset. The selected images were chosen based on $\lambda$ values, *i.e.*, each image has the highest $\lambda$ in each class. Both synthesized and selected images demonstrate similarity at the pixel level sharing common features.

## B Limitations and Future work

In this paper, we propose the DeepKKT condition, which can be applied universally to any deep models to generate deep support vectors (DSVs) that function similarly to support vectors in SVMs. However, it should be noted that the equivalence between DSVs in deep learning models and support vectors in SVMs is only described intuitively, not rigorously. We have shown experimentally in Fig. 7 and intuitively in Sec. C why the DeepKKT condition should be as we suggested, but we have not derived it with rigorous math. Proving this rigorously would be a meaningful research topic.

## C Intuitive explanation of DeepKKT condition

In DeepKKT, many conditions make sense, except for one. For instance, the primal feasibility condition and the manifold condition are reasonable, and the dual feasibility condition can be regarded as importance sampling. However, the most counterintuitive part is the stationarity condition:

$$L_{\text{stat}} = D(\theta^*, -\sum_{i=1}^{n} \lambda_i \nabla_\theta L(\Phi(x_i; \theta^*), y_i)) \tag{12}$$

In this section, we will explain the dynamics of DSVs in an overparameterized deep network and how it is connected to deep learning. Below is a quick analogy of [28] to illustrate this connection.

A deep learning model follows the following ODE:

$$w_{t+1} = w_t - \eta \nabla L(x, y; w_t). \tag{13}$$

Here, $\eta$ is the learning rate and $t$ is the optimization step. The loss $L$ does not go to zero since deep learning models usually exploit a loss function with a logistic tail, such as the cross-entropy loss, and the gradient of the least confident sample (support vector) dominates overall gradient. Thus, there exists a convergence of the gradient direction $g_\infty := \hat{\nabla} L$. There also exists a time $T$ where the

12
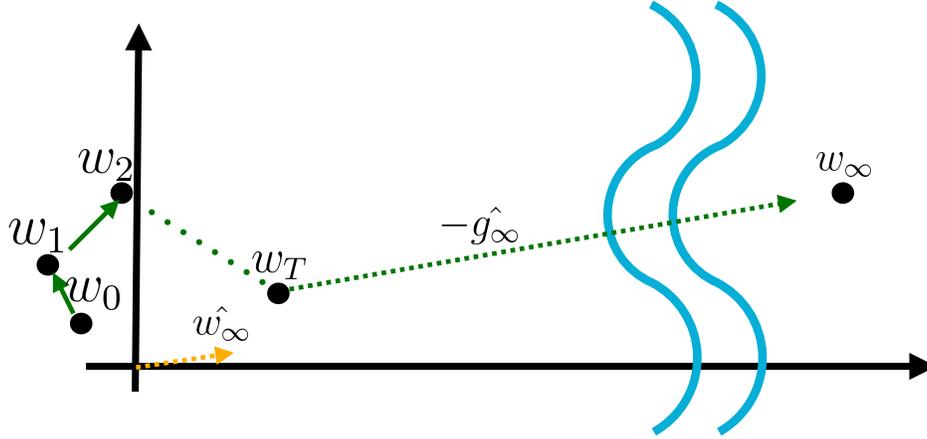
Figure 8: The stationarity condition with a logistic loss. Even though the direction of the gradient $\hat{g}$ converges, the size of the gradient does not go to zero. Therefore, the direction of the converged gradient weight $\hat{w_\infty}$ aligns with $\hat{g}$.

gradient direction converges to $g_\infty - \varepsilon$ for a sufficiently small $\varepsilon$. As illustrated in Fig. 8, $w$ moves toward the direction of $-g_\infty$. Therefore, $\hat{w}_\infty \approx -g_\infty$.

This is for what stationarity condition wants to seek. The direction of $g_\infty$, by using only a few support vectors.

# D   Implementation Details

To obtain the results in Table 2 and Fig. 4, the ConvNet architecture [5] was used for pretraining $\Phi(\cdot; \theta)$ on the SVHN dataset [20], a digit dataset with dimensions similar to CIFAR-10 [11]. For ImageNet, we used the ResNet50 model [7] with the original setting in the paper. Specifically, we used the pretrained model in torchvision library in pytorch [22]. For visualizing synthesized DSVs in ImageNet, we increased the contrast in 224x224 dimensions. When calculating $L_{\text{stat}}$, we averaged the distance per parameter. In Alg. 1, $\eta$ was set to 5.

To synthesize DSVs in ImageNet, we used translation, crop, cutout, flip, and noise for augmentation, with hyperparameters set to 0.125, 0.2, 0.15, 0.5, and 0.01, respectively. In Eq. (9), we set $\alpha$ to 2e-5, $\beta$ to 40, and $\gamma$ to 1e-6. When calculating $L_{\text{stationarity}}$, we averaged the distance per parameter.

For dataset distillation in Table 2, we used translation, crop, flip, and noise for augmentation, with hyperparameters set to 0.125, 0.2, and 0.5, respectively. In Eq. (9), we set $\alpha$ to 2e-3, and both $\beta$ and $\gamma$ to 0. For retraining models with synthesized images, we used a learning rate of 1e-4 while the other parameters set to the default values of the Adam optimizer [9].

To obtain the pretrained weight $\theta^*$ for CIFAR10 and CIFAR100, we chose the ConvNet architecture [5], a common choice in deep learning. This architecture includes sequential convolutional layers followed by max pooling, and a single fully-connected layer for classification. The learning rate was set to $10^{-3}$ with a weight decay of $0.005$ using the Adam optimizer. Additionally, we employed flipping and cropping techniques, with settings differing from those used for DSVs reconstruction to ensure fair comparison. For pretraining $\Phi$ on the Street View House Numbers (SVHN) dataset [20], a digit dataset with dimensions similar to CIFAR-10 [11], we exclusively trained the fully-connected layer of the CIFAR-10 pre-trained ConvNet. This approach resulted in a training accuracy of 80%.

# E   DSVs by Selection

Fig. 9 shows the selected images with large Lagrangian multipliers $\lambda$'s, which correspond to the candidates used in Fig. 2b. Surprisingly, there is a meaningful match between the selected DSVs and the synthesized DSVs in the CIFAR-10 dataset, as shown in Fig. 7. This implies that synthesizing DSVs corresponds to reviving training data that lie on the boundary manifolds.

13
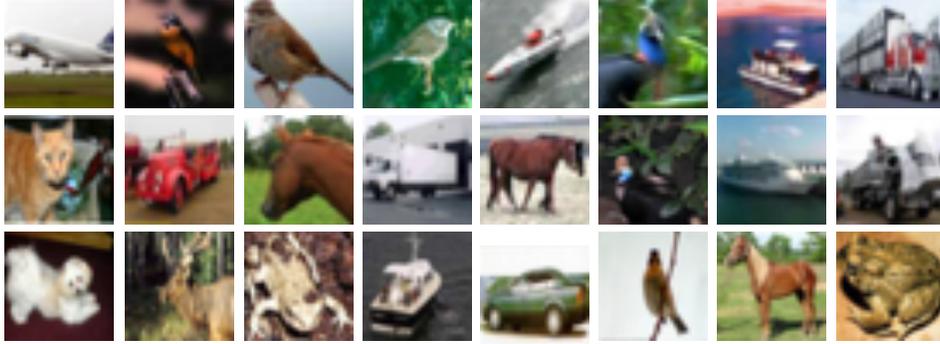
Figure 9: Images of DSV candidates (Selected in the CIFAR-10 dataset).

{'plane'} + {'truck'} = {'truck'}

{'cat'} + {'bird'} = {'bird'}
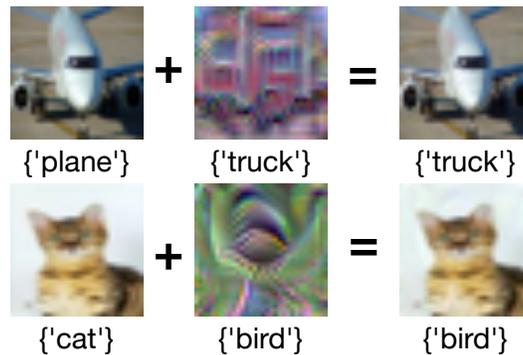
Figure 10: Examples of adversarial attack in the CIFAR-10 dataset

## F    More Characteristics of DSVs

**DSVs Are Full of Discriminative Information**    In Fig. 10, we conducted an experiment by mixing a randomly sampled image from the real dataset with an image from the DSVs. Upon observation, the mixed image is virtually indistinguishable from an image obtained solely from the real dataset. It is noteworthy to highlight this situation resembles that of an adversarial attack [32, 19], yet we did not apply gradient descent to the image; we simply mixed two images. This suggests that the discriminative informational density in a single DSV image is substantially greater than that in a randomly sampled image. The fact that the DSV's characteristics remained dominant in the classification, underscores the significant role of DSVs in explaining the model's classification ability.

## G    More Examples

In Fig. 11, examples of latent interpolations between target labels are presented. The smoothness of these interpolations within the latent space indicates that the semantic information learned from the training data has been effectively applied during the DSV generation process. This observation provides evidence that the DeepKKT optimization successfully conducts the generative process.

Fig. 12 and  13 provide examples of deep support vectors generated using the CIFAR-100 and ImageNet datasets, respectively. Fig. 14 presents additional examples related to image editing.

Fig. 15 empirically supports on our assertion on decision criterion. Starting from CIFAR100 random images and CIFAR10-pretrained models, we edited the image CIFAR10 labels as latents. The edited images changes the image following decision criterions in generated DSVs. 1) For editing images to deer, antler grows. 2) For dog editing, facial dots are generated. 3) For cat editing, pointed traiangler features are generated.

14
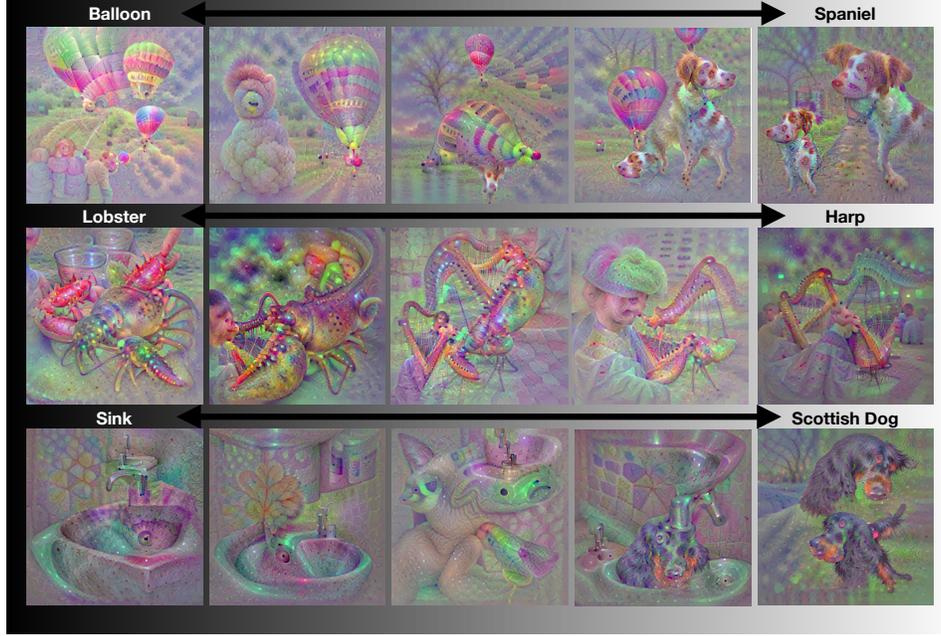
Figure 11: More examples of latent interpolation in the ImageNet dataset

## H Algorithm

Alg. 1 presents our algorithm of generating Deep Support Vectors (DSVs). Initialized either from a noise $x_i^s \sim \mathcal{N}(0, I)$ or a real sample, it iterates to obtain the primal $X^S$ and dual $\Lambda^S$ variables.

---

**Algorithm 1** Support Vector Refinement for Deep Learning Model

---

**Require:** Pretrained classifier $\Phi(\cdot; \theta)$, loss function $L$, augmentation function set $A$, number of DSV candidate $N$, number of class $C$, hyperparameters $\alpha, \beta$
**Ensure:** Freeze classifier $\Phi(\cdot; \theta)$
 1: Initialize $N \times C$ number of support vector candidates
 2: **for** $i = 1$ to $C$ **do**
 3:     sample $N$ number of $(x_i^s, \lambda_i^s)$ for label $y_i^s = i$
 4: **end for**
 5: Define $X^S = \{x_i^s \mid i \in [C], s \in [N]\}$
 6: Define $\Lambda^S = \{\lambda_i^s \mid i \in [C], s \in [N]\}$
 7: **repeat**
 8:     $L_{\text{primal}}(X^S) = \sum_{s=1}^{N} \sum_{i=1}^{C} L(\Phi(x_i^s; \theta), y_i^s)$
 9:     $L_{\text{stationary}}(X^S) = \|\theta + \sum_{s=1}^{N} \sum_{i=1}^{C} \lambda_i^s y_i^s \nabla_\theta \Phi(x_i^s; \theta)\|_2^2$
10:     $L_{\text{kkt}}(X^S) = \beta_1 \cdot L_{\text{primal}}(X^S) + L_{\text{stationary}}(X^S)$
11:     $L_{\text{prior}} = \beta_2 \cdot L_{\text{tot}}(X) + \beta_3 L_{\text{norm}}(X)$
12:     Sample $f_A \in A$
13:     Define $AX^S = \{f_A(x_i^s) \mid x_i^s \in X^S\}$
14:     $L_{\text{akkt}}(X^S) = L_{\text{kkt}}(AX^S)$
15:     $L_{\text{total}}(X^S) = L_{\text{kkt}}(X^S) + \eta \cdot L_{\text{akkt}}(X^S) + L_{\text{prior}}$
16:     Update $X^S \leftarrow X^S + \nabla_{X^S} L_{\text{total}}(X^S)$
17:     Update $\Lambda^S \leftarrow \Lambda^S + \nabla_{\Lambda^S} L_{\text{total}}(X^S)$
18:     Remove $x_i^s$s for corresponding $\lambda_i^s < 0$
19: **until** $X^S$ converges
20: **return** Set of DSV : $X^S$

---

15
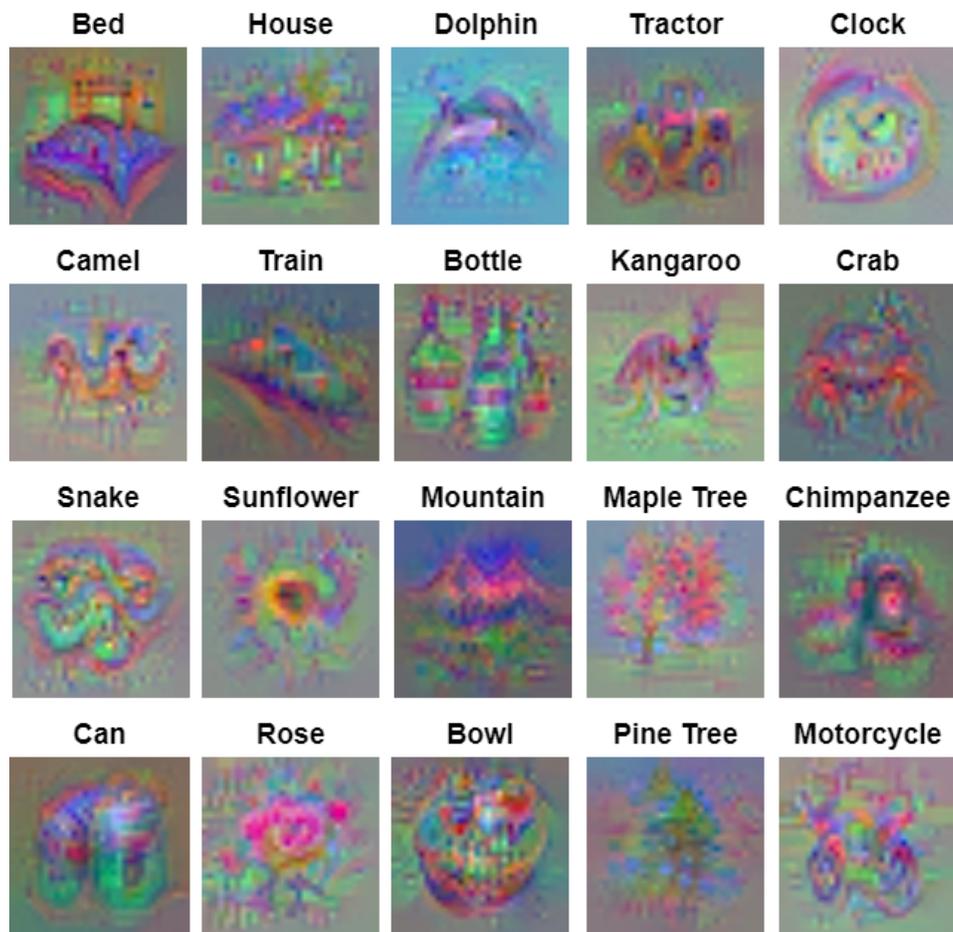
Figure 12: More examples of generated images with CIFAR-100 dataset

16
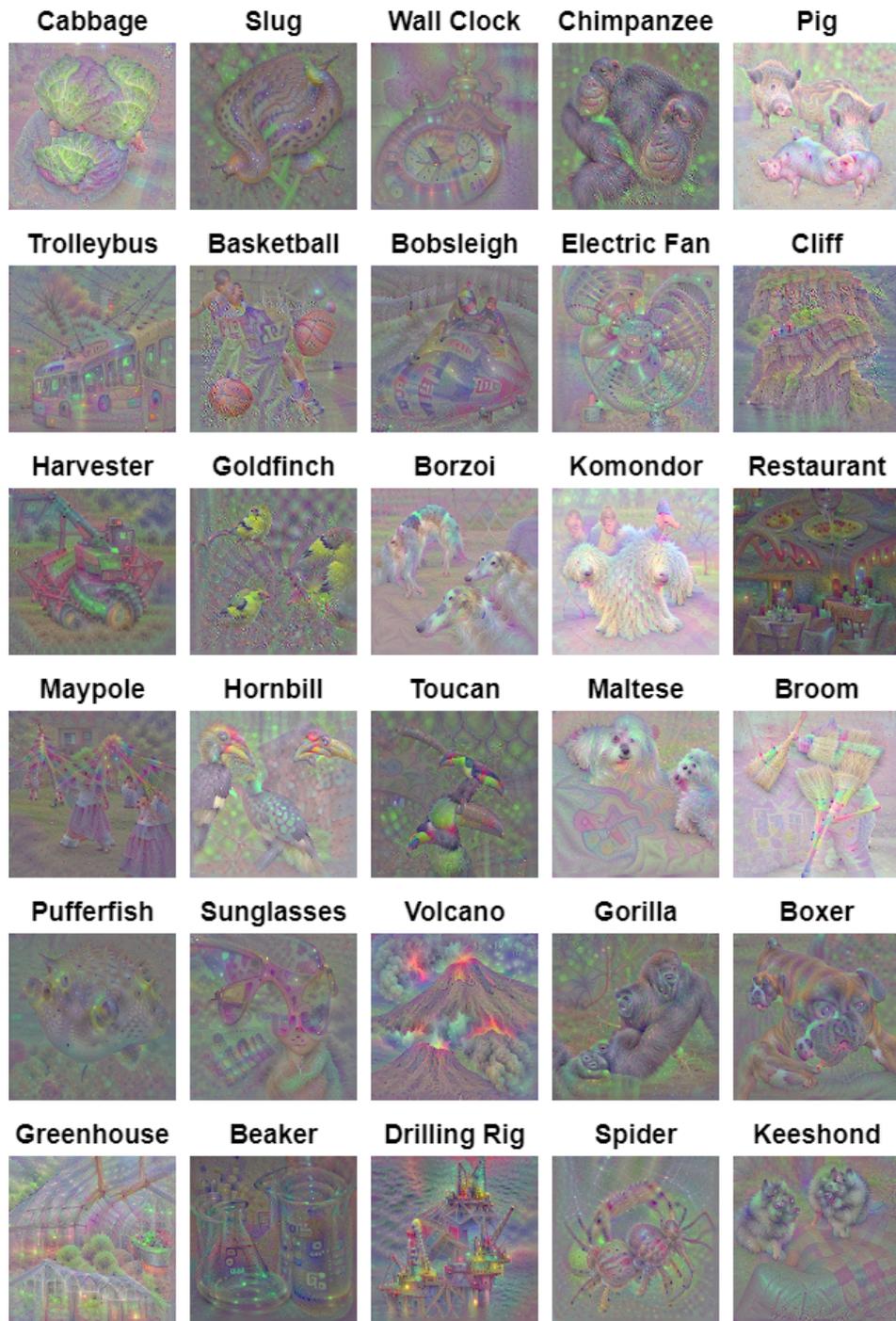
Figure 13: More examples of generated images with ImageNet dataset

17
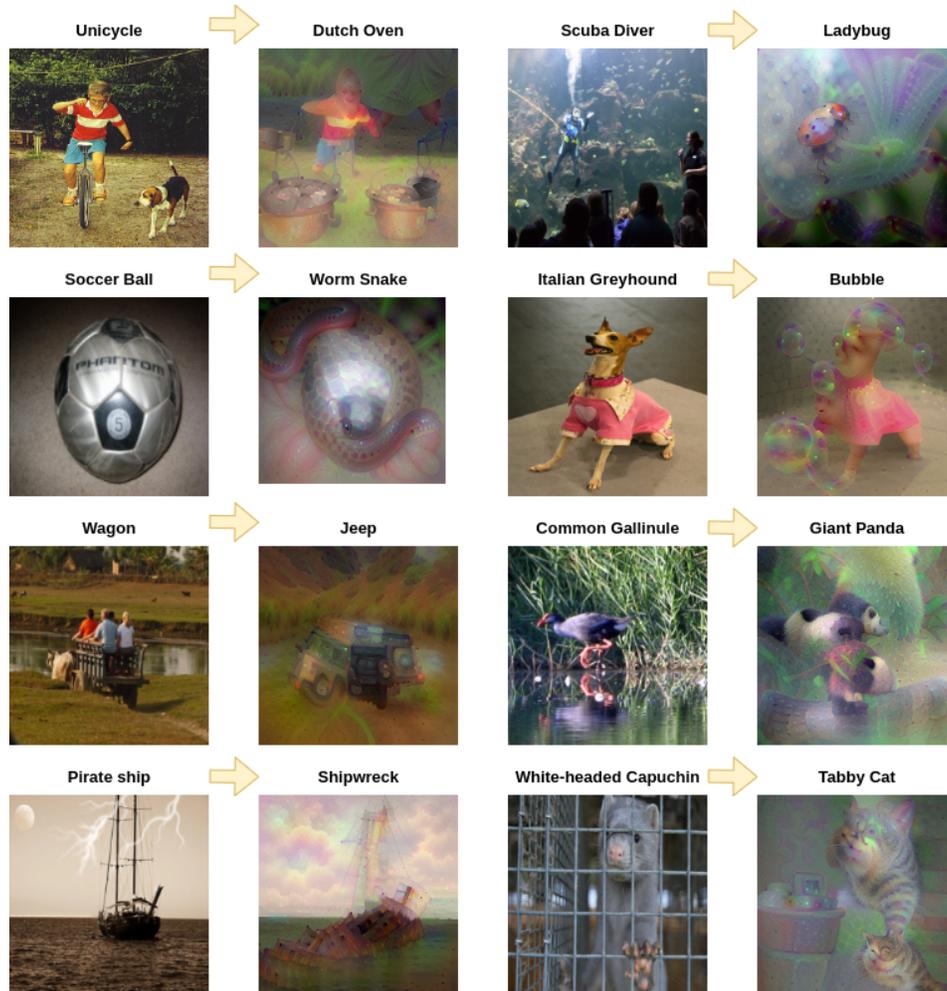
Figure 14: More examples of image editing. The images to the left of the arrows represent the initial images before training, while those to the right depict the edited images after training.
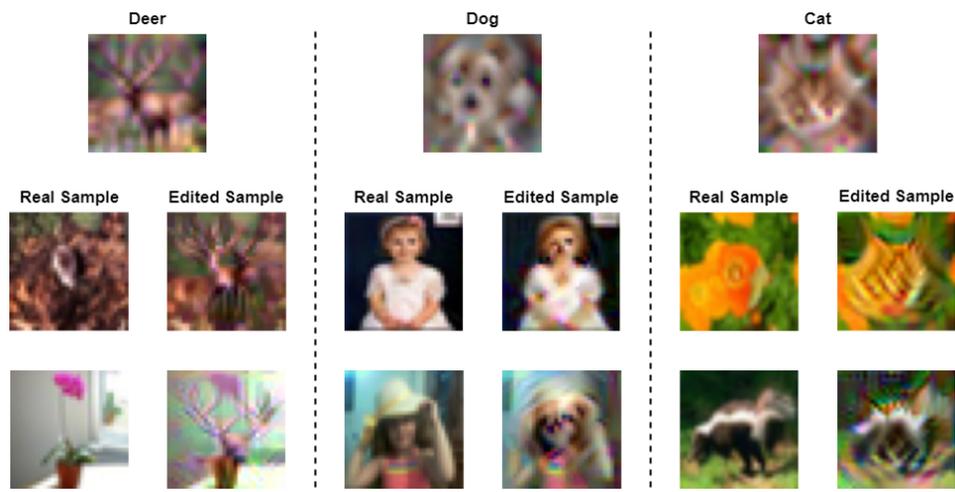
Figure 15: More examples of image editing. The images to the left of the arrows represent the initial images before training, while those to the right depict the edited images after training.

18

# NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Introduction and abstract effectively explains the paper's contributions and scope

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See sec B.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: There are no theorems or lemma in this paper.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide code

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide code and its setting

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes, we leave full details in Sec **??**

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Yes, we offer standard deviation.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Yes, see Sec D

9. **Code Of Ethics**

19

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification:

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: See Sec. A

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification:

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We explained used library and environments with citations

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification:

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification:

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification:

20