Supplementary: On Generating Transferable Target Perturbations

We study the effect of augmentations and ensemble learning by analysing class-wise transferability in Appendix A. We further discuss on why augmentations and ensemble learning leads to more transferable targeted patters in Appendix A.1 and Appendix A.2. We then present the vulnerability of batchnorm to black-box targeted perturbations in Appendix B. In Appendix C, we analyze the effect of linear back-propagation of gradients [3] and using more gradients from skip connections [14] on the targeted attack transferability. For the sake of completeness, we report the drop in clean accuracy caused by different defenses including input processing methods (JPEG, Median Blur, and NRP), adversarial training, and stylized training in Appendix D. Names of 100 target classes are provided in Appendix E. Finally, we present visual illustrations to showcase different targeted adversarial patterns found by our method, TTP (Transferable Targeted Perturbations), in Appendix F.

Appendix A. Effect of Augmentations and Ensemble Learning

We proposed a mechanism to explore augmented adversarial space and ensemble learning to boost transferability of the targeted adversarial perturbations found by TTP. A per-class analysis for 10 targets presented in Table 1 reveals that augmentations and ensemble learning increase the adversarial effect for every target. TTP is trained against naturally trained ResNet50 and ResNet ensemble $\mathtt{R}_{ens}$`:ResNet{18,50,101,152}` and perturbations are transferred to naturally trained VGG16 and stylized VGG16 [2]. In some cases, such as `Hippopotamus`, augmented learning maximizes the transferability from ResNet50 to naturally trained VGG16 by more than 100% (Table 1). Similarly, we observe that ensemble learning proves to be effective e.g., see `Grey-Owl` in Table 1. VGG16 trained on stylized ImageNet showed higher resistance against targeted adversarial attacks. For example, transferability of perturbations found by TTP for `French Bulldog` distribution is around 11% on VGG16 (SIN) as compared to 63% on VGG16 trained on ImageNet (IN) (Table 1).

Appendix A.1. Why Augmentations boost Transferability?

Ilyas *et al*. [5] showed that adversarial examples can be explained by features of the attacked class label. In our targeted attack case, we wish to imprint the features of the target class distribution onto the source samples within an allowed distance (*e.g.* $l_\infty \leq 16$). However, a black-box (unknown) model might apply different set of transformations (from one layer to another) to process such features and reduce the target transferability. Training on adversarial augmented samples allows the generator to capture such targeted features that are robust to transformations that may vary from one model to another.

Figure 1: Unconstrained targeted patterns for Snowmobile are shown to demonstrate how discriminators (models) from the same family can capture different information to classify a certain class. Thus, TTP when trained against ensemble of same family models shows higher transferability than any of the individual model.

Appendix A.2. Why an Ensemble of Weak Models maximizes Transferability?

Different models of the same family of networks can exploit different information to make prediction. One such example is shown in Fig. 1. Generators are trained against Dense121 and Dense169 to target Snowmobile distribution. Unrestricted generator outputs reveal that Dense121 is more focused on Snowmobile's blades while Dense169 emphasis background pine tree patterns to discriminate Snowmobile samples. This complementary information from different models of the same family helps the generator to capture more generic global patterns for a given target distribution.

Appendix B. The Vulnerability of Batchnorm

Batchnorm [8] helps in optimization of neural networks as well as increases their clean accuracy. However, our empirical cross-family (Dense $\rightarrow$ VGG$_{BN}$, Dense $\rightarrow$ VGG, ResNet $\rightarrow$ VGG$_{BN}$, ResNet $\rightarrow$ VGG) analysis presented in Fig. 2 suggests that batchnorm makes the model more vulnerable to the targeted adversarial attacks. Adversarial perturbations found by TTP transfer better against models trained using batchnorm as compared to models trained without it (Fig. 2).

Appendix C. Skip Connections and Linear Back-Propagation of Gradients

Dongxian *et al*. [14] observed that while back-propagating, giving more importance to the gradients coming from skip connections can enhance adversarial transferability. Similarly, Guo *et al*. [3] showed that encouraging

Source	Augmentations	Grey-Owl	Goose	Bulldog	Hippopotamus	Cannon	Fire-Truck	Model-T	Parachute	Snowmobile	Street-Sign	Average
		Target Model: VGG16										
ResNet50	✗	56.5	80.9	49.0	43.9	61.9	82.9	56.5	89.4	41.3	72.9	63.5
ResNet50	✓	56.7	84.1	63.7	94.9	79.5	91.5	76.5	89.8	70.4	80.8	78.8
R_{ens}	✓	85.1	94.5	63.3	97.8	90.5	95.8	90.7	96.1	89.6	90.4	89.1
		Target Model: VGG16 (SIN)										
ResNet50	✗	1.61	43.1	0.50	40.9	14.9	9.6	5.8	36.2	6.2	19.2	17.8
ResNet50	✓	1.30	69.6	11.6	68.7	17.0	15.2	20.5	33.2	35.4	30.9	30.3
R_{ens}	✓	17.6	77.7	11.4	77.0	59.7	48.4	56.1	72.8	74.1	41.2	53.6

Table 1: ***Per Target Transferability of our Method (*****TTP*****):*** Top-1 target accuracy (%) with 49.95K ImageNet val. samples for each target. Perturbation budget: $l_\infty \leq 16$. Adversarial perturbations are transferred from naturally trained ResNet50 and ResNet ensemble to naturally trained VGG16 and stylized VGG16 [2]. Augmentations as well as ensemble learning improves efficiency of TTP.

Figure 2: ***Batchnorm Vulnerability to Targeted Transferability :*** {10-Targets (all source) settings}. TTP (Algorithm 1 in the paper) strength is higher against models trained naturally with batchnorm as compared to without batchnorm. Batchnorm [8] provides better optimization and increase model clean accuracy but these empirical results indicate that it also make the model more vulnerable to blackbox targeted attacks. Each value is averaged across 10 targets (see Section 4 in the paper for details) with 49.95k ImageNet val. samples for each target. Perturbation budget is $l_\infty = 16$.

Source	Attack	Natural Training					Augs.	Stylized		Adversarial	
		VGG19$_{BN}$	Dense121	ResNet152	WRN-50-2	VGG16	Augmix	SIN	VGG16 (SIN)	Adv. ($l_\infty = 0.5$)	Adv. ($l_\infty = 1.0$)
ResNet50	PGD [10]	0.8/2.1	1.9/3.7	3.0/4.7	2.5/4.4	0.3/1.5	0.4/1.3	0.1/0.4	0.1/0.3	0.0/0.0	0.0/0.0
	MI [1]	1.5/1.8	3.2/6.2	3.1/5.6	3.0/4.6	1.1/1.4	1.0/1.6	0.3/0.9	0.2/0.4	0.0/0.1	0.0/0.0
	DIM [15]	10.4/14.4	16.2/26.0	13.4/20.9	13.4/19.8	6.4/6.7	4.8/7.7	1.7/3.2	0.5/1.2	0.2/0.5	0.1/0.1
	Po-TRIP [9]	12.5/15.0	18.2/30.0	15.9/23.7	14.2/22.3	7.3/8.9	5.5/9.0	2.1/3.7	0.8/2.0	0.3/0.7	0.1/0.1
	FDA-fd [6]	16.0/25.3	21.0/33.1	19.7/32.9	17.1/28.4	12.0/18.7	15.3/19.3	3.1/6.3	1.2/3.0	0.1/1.9	0.1/0.3
	FDA-N [7]	32.1/38.6	48.3/52.3	37.5/39.0	35.5/40.7	19.0/28.3	20.3/30.3	5.0/16.6	3.0/10.7	0.6/4.7	0.2/0.8
	SGM [14]	19.2/26.3	25.9/40.6	19.7/31.1	21.6/30.4	13.5/13.7	10.5/15.9	2.6/6.1	1.3/2.8	0.5/1.2	0.1/0.3
	SGM [14] + LinBP [3]	22.0/27.1	34.5/40.0	30.5/32.9	25.1/21.0	14.8/15.0	17.3/25.3	4.6/14.3	2.4/8.0	0.3/2.9	0.1/0.3
	Ours (TTP)	**79.0/81.4**	**84.4/87.0**	**81.9/86.6**	**80.2/81.2**	**79.4/78.2**	**72.7/81.2**	**30.5/42.4**	**29.3/36.9**	**5.5/50.1**	**0.4/17.1**

Table 2: ***Target Transferability:*** {10-Targets (sub-source)} Top-1 target accuracy (%) averaged across 10 targets. Perturbation budget: $l_\infty \leq 16/32$. SIN [2] and Adv (l_∞=0.5), and Adv (l_∞=1.0) [13] are ResNet50 models trained using stylized and adversarial examples, respectively. Augs. represents augmentation based training [4] of ResNet50.

Model	Defense	Accuracy	Difference
VGG19$_{BN}$	–	74.24	0.0
	JPEG	67.34	-6.90
	Blur	53.86	-20.38
	NRP	72.00	-2.24
Dense121	–	74.65	0.0
	JPEG	68.92	-5.73
	Blur	61.27	-13.38
	NRP	72.01	-2.63
ResNet50	–	76.15	0.0
	JPEG	70.82	-5.33
	Blur	61.30	-14.85
	NRP	73.21	-2.94

Table 3: ***Effect of Input Processing on Clean Accuracy:*** Top-1 (%) accuracy on ImageNet val. set (50k images). Median Blur with window size 5×5 causes large drop in clean accuracy while NRP [11] has the least effect on the model's clean accuracy.

linearity while back-propagating gradients improve transferability. Here, we analyze target transferability of both of these techniques [14, 3] and present a holistic comparison of approach (TTP) against iterative instance-specific attacks in Table 2. Our approach sets new state-of-the-art in targeted adversarial transferability by notable large margins.

Appendix D. Clean Accuracy vs. Defenses

We evaluate the effect of different defenses on model's clean accuracy. We study the input processing methods including JPEG with quality 50% [12], Median Blur with kernel size 5×5 [12] and NRP [11] as well as different training mechanisms including Augmix [4], stylized [2] and adversarial training methods [10, 13]. Results are presented in Tables 3 & 4. We observe that Median Blur causes a significant drop in clean accuracy (Table 3) while among training methods, stylized training (SIN) [2] has the most negative effect on the clean accuracy.

Model	Training Type	Accuracy	Difference
ResNet50	IN	76.15	0.0
	SIN	60.18	-15.97
	SIN-IN	74.59	-1.56
	Augmix	77.53	+1.38
	Adv. ($l_\infty, \epsilon = .5$)	73.73	-2.42
	Adv. ($l_\infty, \epsilon = 1$)	72.05	-4.10
	Adv. ($l_2, \epsilon = .1$)	74.78	-1.37
	Adv. ($l_2, \epsilon = .5$)	73.16	-2.99
VGG16	IN	71.59	0.0
	SIN	52.26	-19.33

Table 4: ***Effect of Robust Training on Clean Accuracy:*** Top-1 (%) accuracy on ImageNet val. set (50k images). Every training mechanism with the exception of Augmix [4] reduces model's clean accuracy. Stylized training [2] causes significant drop in accuracy in comparison to other types of training methods.

Appendix E. 100 Targets Names

The performance of TTP is evaluated against the following randomly selected 100 targets (see Sec. 4.1 of the paper). We divide ImageNet classes into 100 mutually exclusive sets. Each set contains 10 classes. We randomly selected one target from each set.

`Tiger-Shark, Bulbul, Grey-Owl, Terrapin, Komodo-Dragon, Thunder-Snake, Trilobite, Scorpion, Quail, Goose, Jellyfish, Slug, Flamingo, Bustard, Dowitcher, Chihuahua, Beagle, Weimaraner, Lakeland-Terrier,` Australian-Terrier, Golden-Retriever, English-Setter, Komondor, Appenzeller, French-Bulldog, Chow, Keeshond, Hyaena, Egyptian-Cat, Lion, Bee, Leafhopper, Sea-Urchin, Zebra, Hippopotamus, Polecat, Gorilla, Langur, Eel, Anemone-Fish, Airliner, Banjo, Bassinet, Beaker, Bell-Cote, Bookcase, Buckle, Cannon, CD-Player, Chain-Saw, Coil, Cornet, Crutch, Dome, Electric-Guitar, Fire-Truck, Garbage-Truck, Greenhouse, Grocery-Store, Honeycomb, iPod, Jigsaw-Puzzle, Lipstick, Maillot, Maze, Military-Uniform, Model-T, Neck-Brace, Overskirt, Parachute, Pay-Phone, Pickup, Pirate-Ship, Poncho, Purse, Rain-Barrel, Rotisserie, School-Bus, Sewing-Machine, Shopping-Cart, Snowmobile, Spatula, Stove, Sunglass, Teapot, Toaster, Tractor, Umbrella, Velvet, Wallet, Whiskey-Jug, Street-Sign, Ice-Lolly, Pretzel, Cardoon, Hay, Pizza, Volcano, Rapeseed, Agaric

Appendix F. Visual Demos

Figures 3, 4, 5, 6, 7 and 8 show different targeted patterns produced by TTP trained against naturally trained ResNet50. We demonstrate how adversarial patterns evolve as TTP learns to model a certain target distribution from different networks of the same family in Figures 9 and 10.

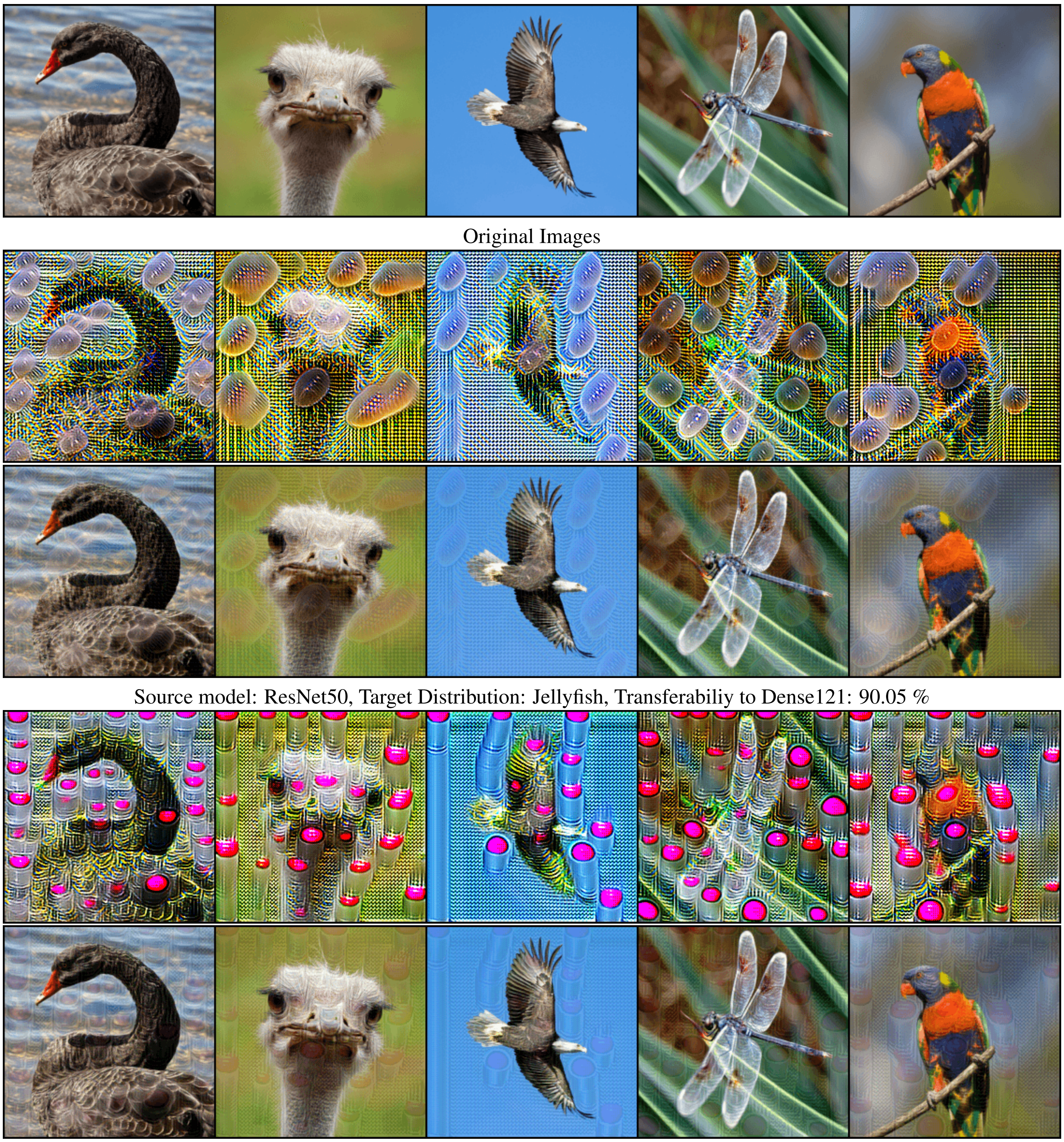


Figure 3: Targeted adversaries produced by TTP (before and after valid projection) trained against ResNet50. Observe that adversarial patterns are not constant rather TTP adapts to the input sample and adds different patterns to different samples to achieve maximum transferability. Transferability is measured as Top-1 target accuracy on the ImageNet val. set (49.95k samples excluding the target images).

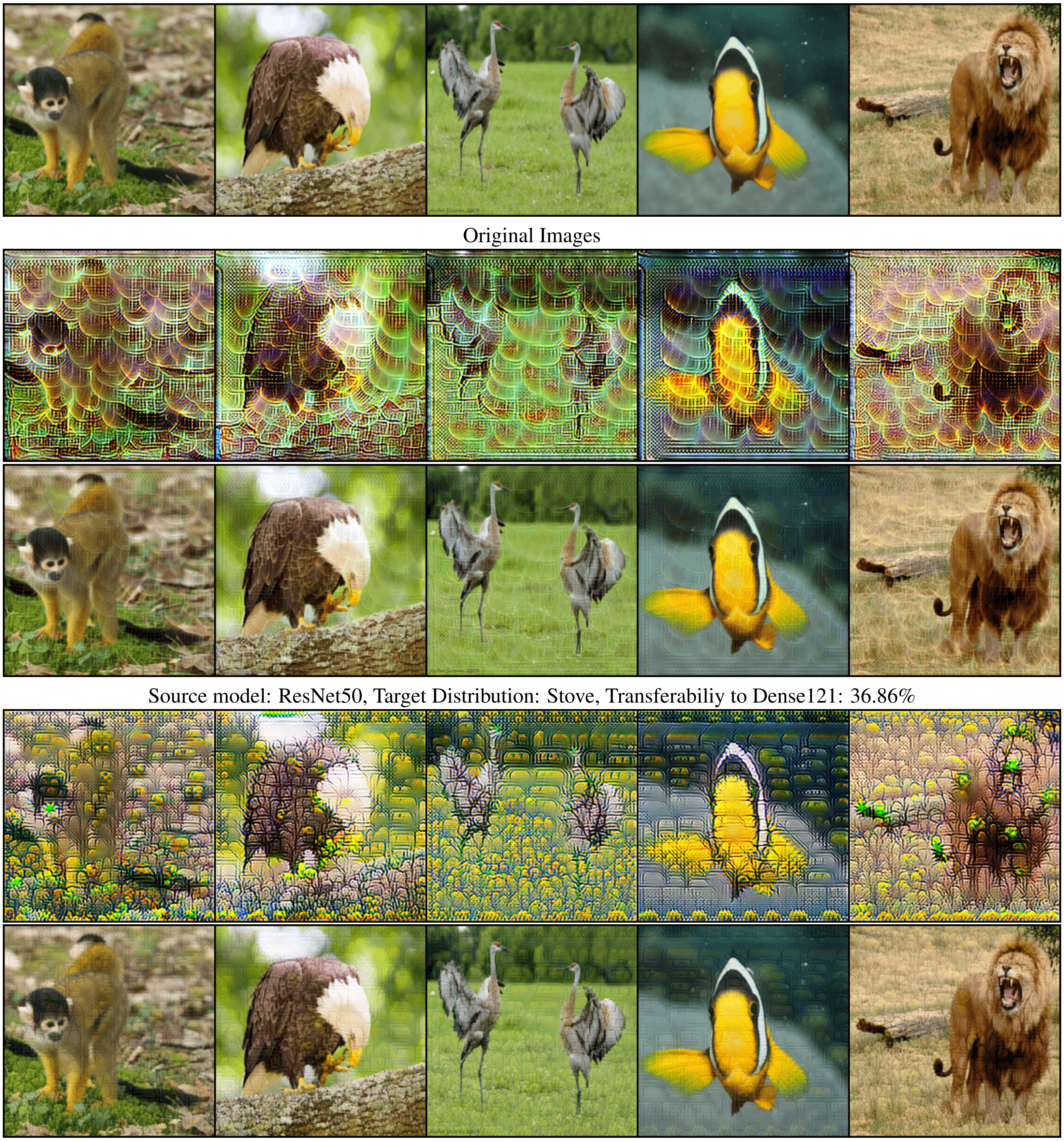
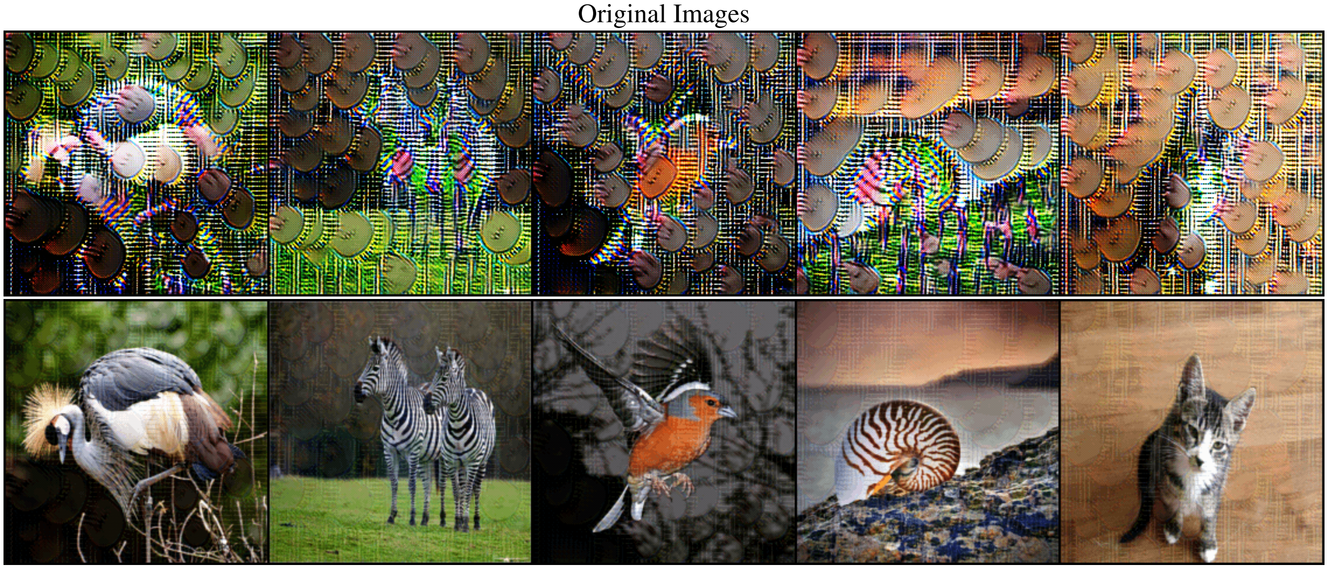
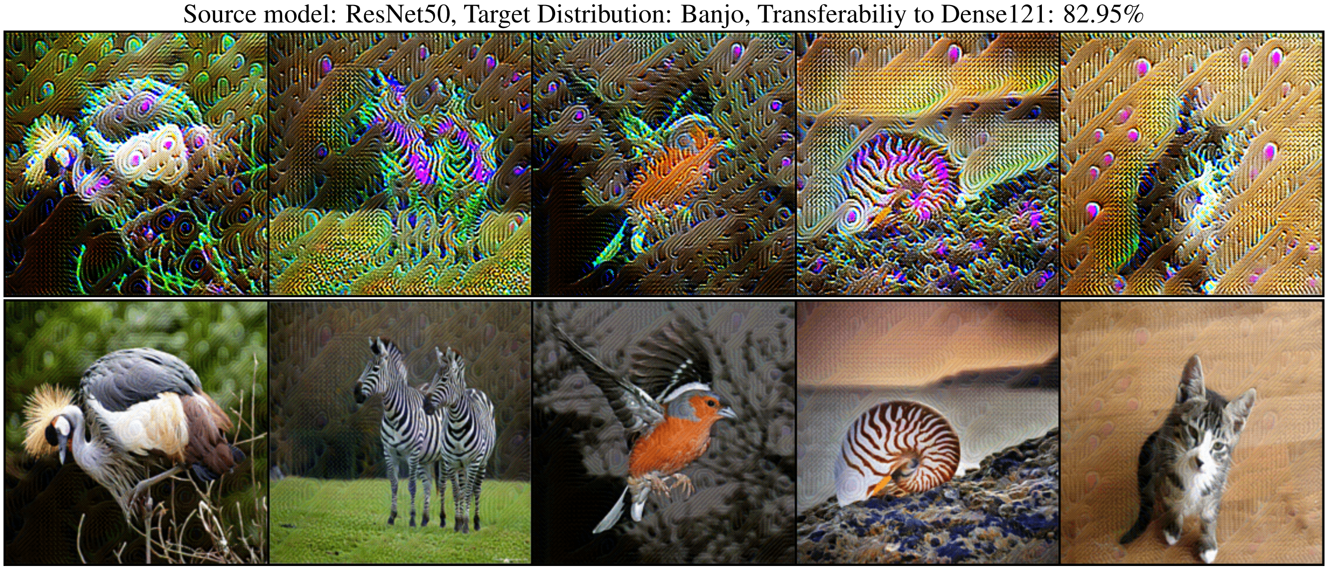


Figure 4: Targeted adversaries produced by TTP (before and after valid projection) trained against ResNet50. Observe that adversarial patterns are not constant rather TTP adapts to the input sample and adds different patterns to different samples to achieve maximum transferability. Transferability is measured as Top-1 target accuracy on the ImageNet val. set (49.95k samples excluding the target images).

Original Images



Source model: ResNet50, Target Distribution: Banjo, Transferabiliy to Dense121: 82.95%



Source model: ResNet50, Target Distribution: Anemone Fish, Transferabiliy to Dense121: 74.45%

Figure 5: Targeted adversaries produced by TTP (before and after valid projection) trained against ResNet50. Observe that adversarial patterns are not constant rather TTP adapts to the input sample and adds different patterns to different samples to achieve maximum transferability. Transferability is measured as Top-1 target accuracy on the ImageNet val. set (49.95k samples excluding the target images).

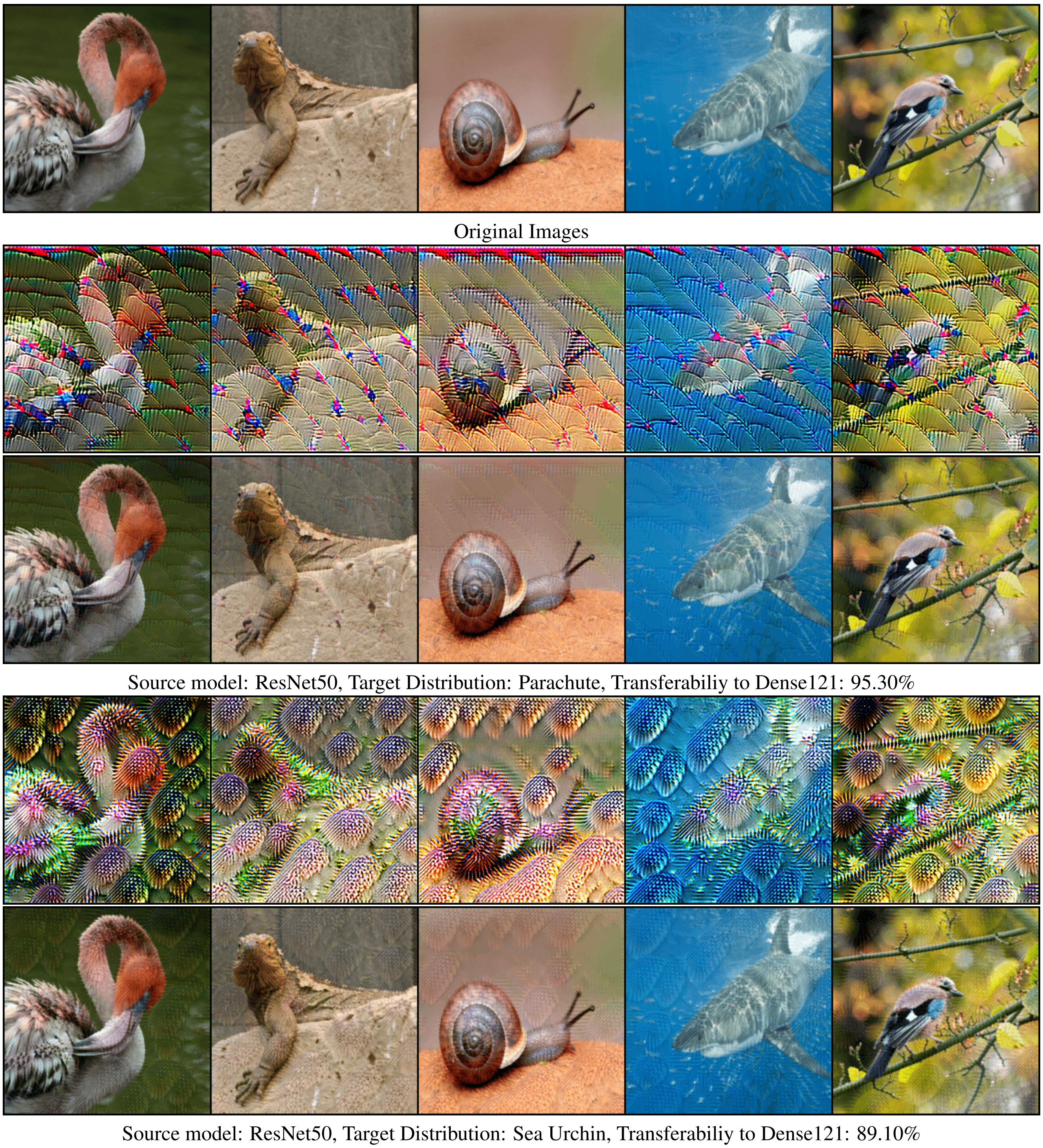
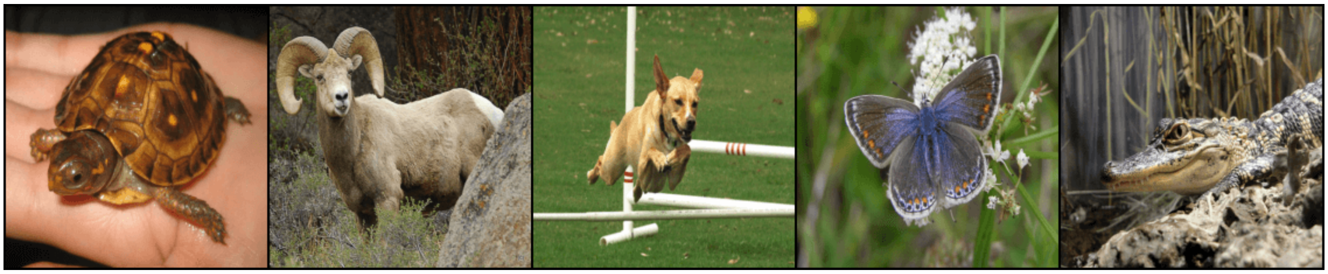
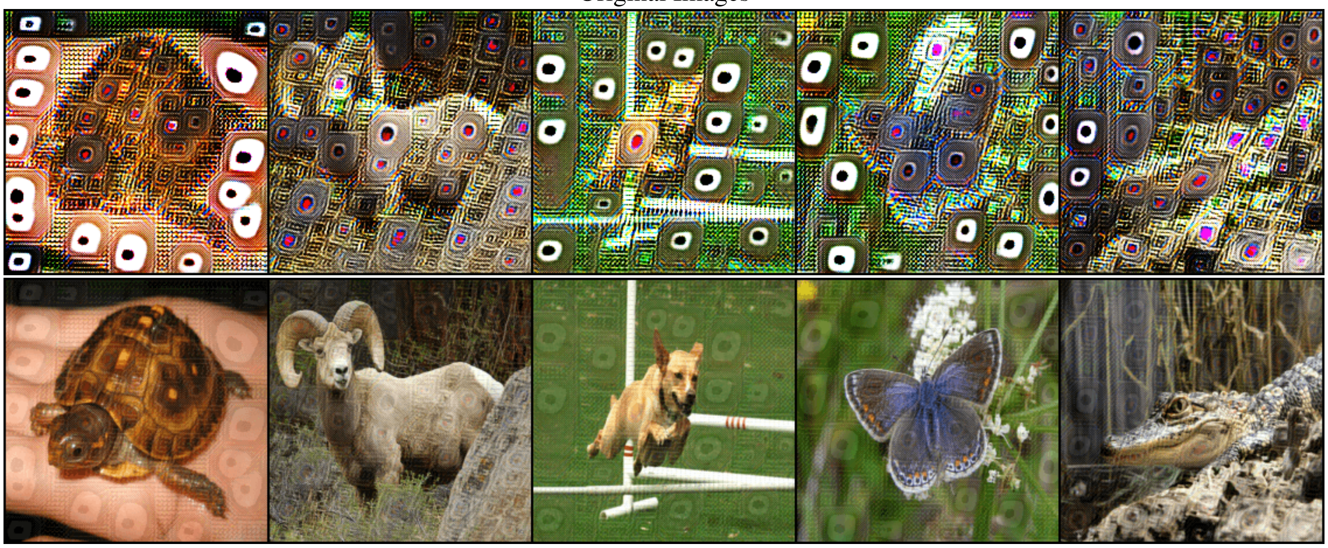


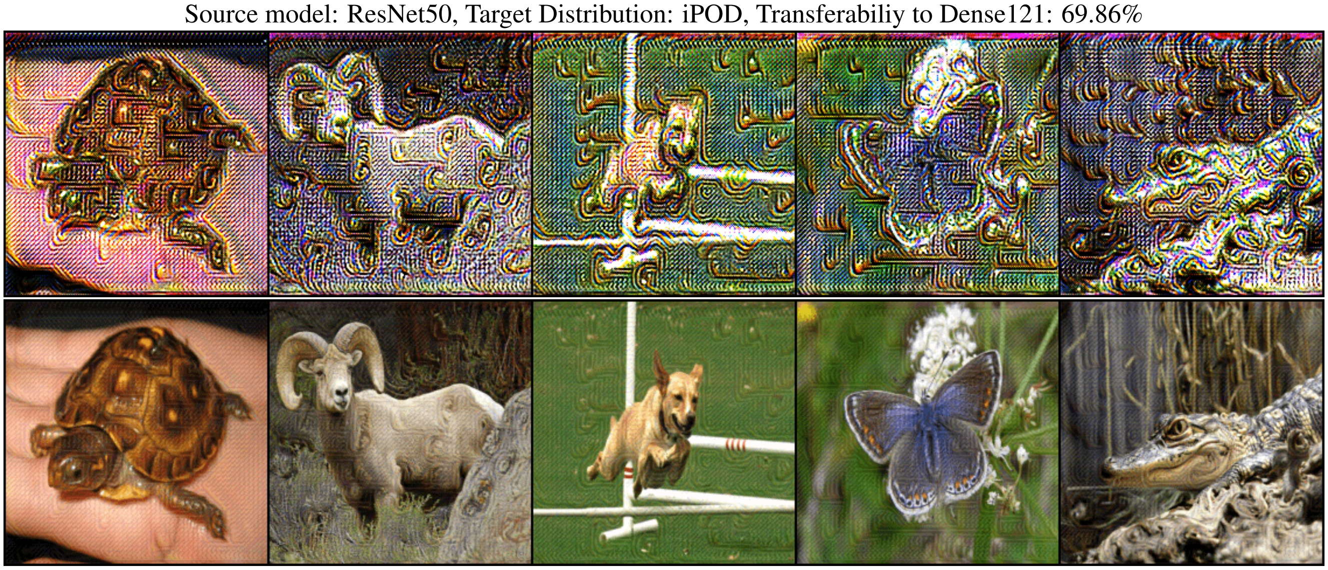
Figure 6: Targeted adversaries produced by TTP (before and after valid projection) trained against ResNet50. Observe that adversarial patterns are not constant rather TTP adapts to the input sample and adds different patterns to different samples to achieve maximum transferability. Transferability is measured as Top-1 target accuracy on the ImageNet val. set (49.95k samples excluding the target images).



Original Images



Source model: ResNet50, Target Distribution: iPOD, Transferabiliy to Dense121: 69.86%



Source model: ResNet50, Target Distribution: Buckle, Transferabiliy to Dense121: 77.06%

Figure 7: Targeted adversaries produced by TTP (before and after valid projection) trained against ResNet50. Observe that adversarial patterns are not constant rather TTP adapts to the input sample and adds different patterns to different samples to achieve maximum transferability. Transferability is measured as Top-1 target accuracy on the ImageNet val. set (49.95k samples excluding the target images).

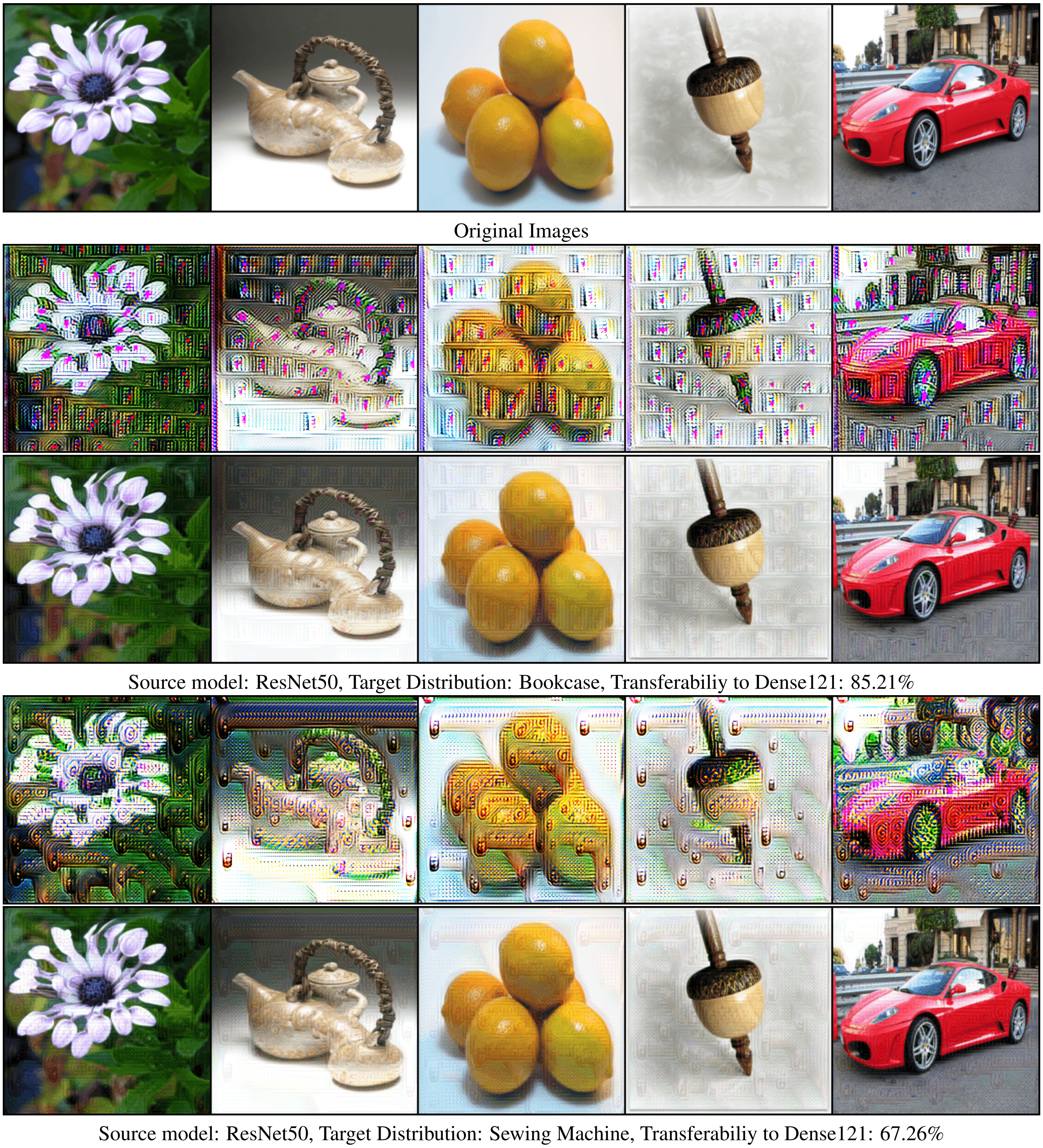


Figure 8: Targeted adversaries produced by TTP (before and after valid projection) trained against ResNet50. Observe that adversarial patterns are not constant rather TTP adapts to the input sample and adds different patterns to different samples to achieve maximum transferability. Transferability is measured as Top-1 target accuracy on the ImageNet val. set (49.95k samples excluding the target images).

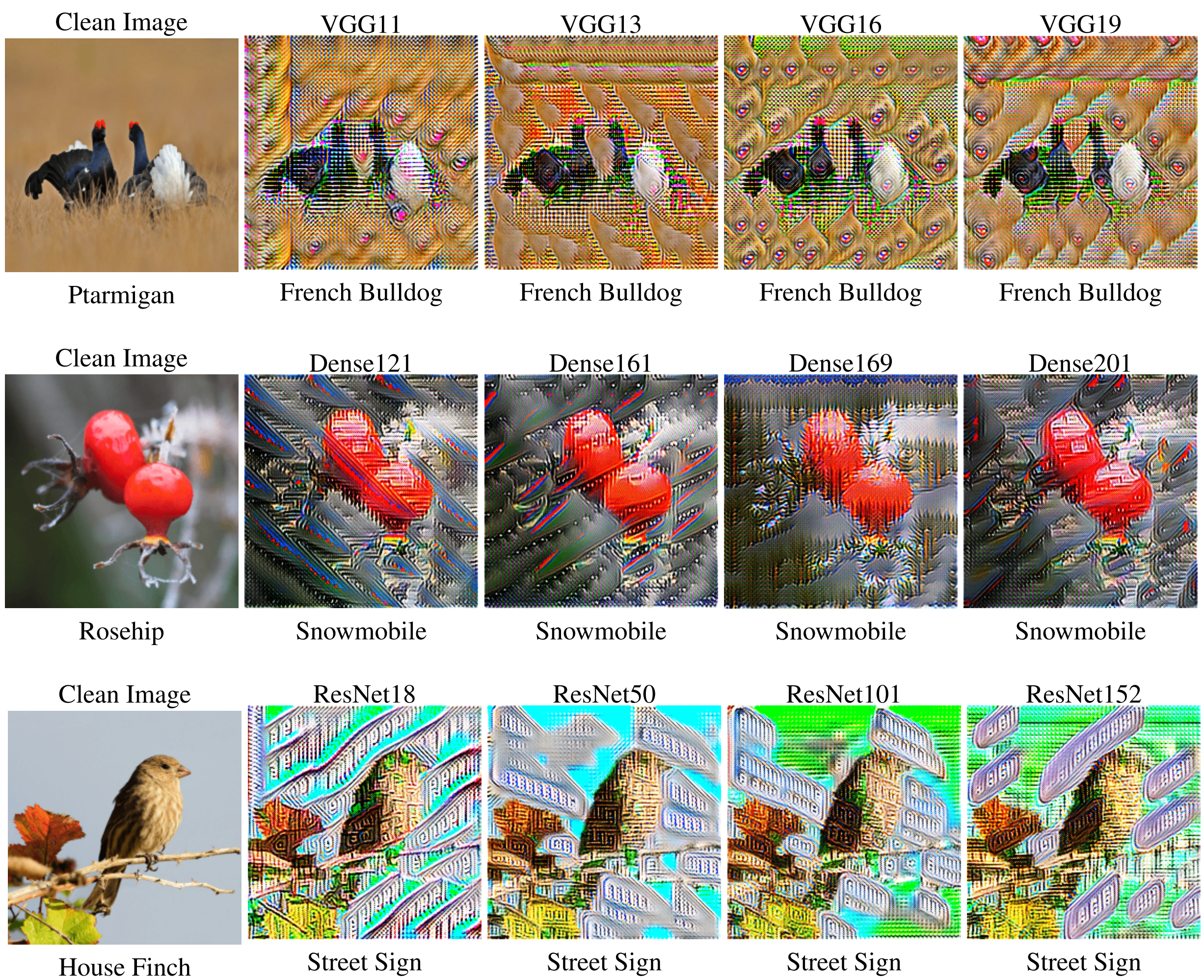


Figure 9: ***Evolution of* TTP*:*** Unconstrained targeted adversarial patterns generated by TTP are shown to demonstrate how TTP evolves as it learns perturbations from different source models of a certain family of networks.

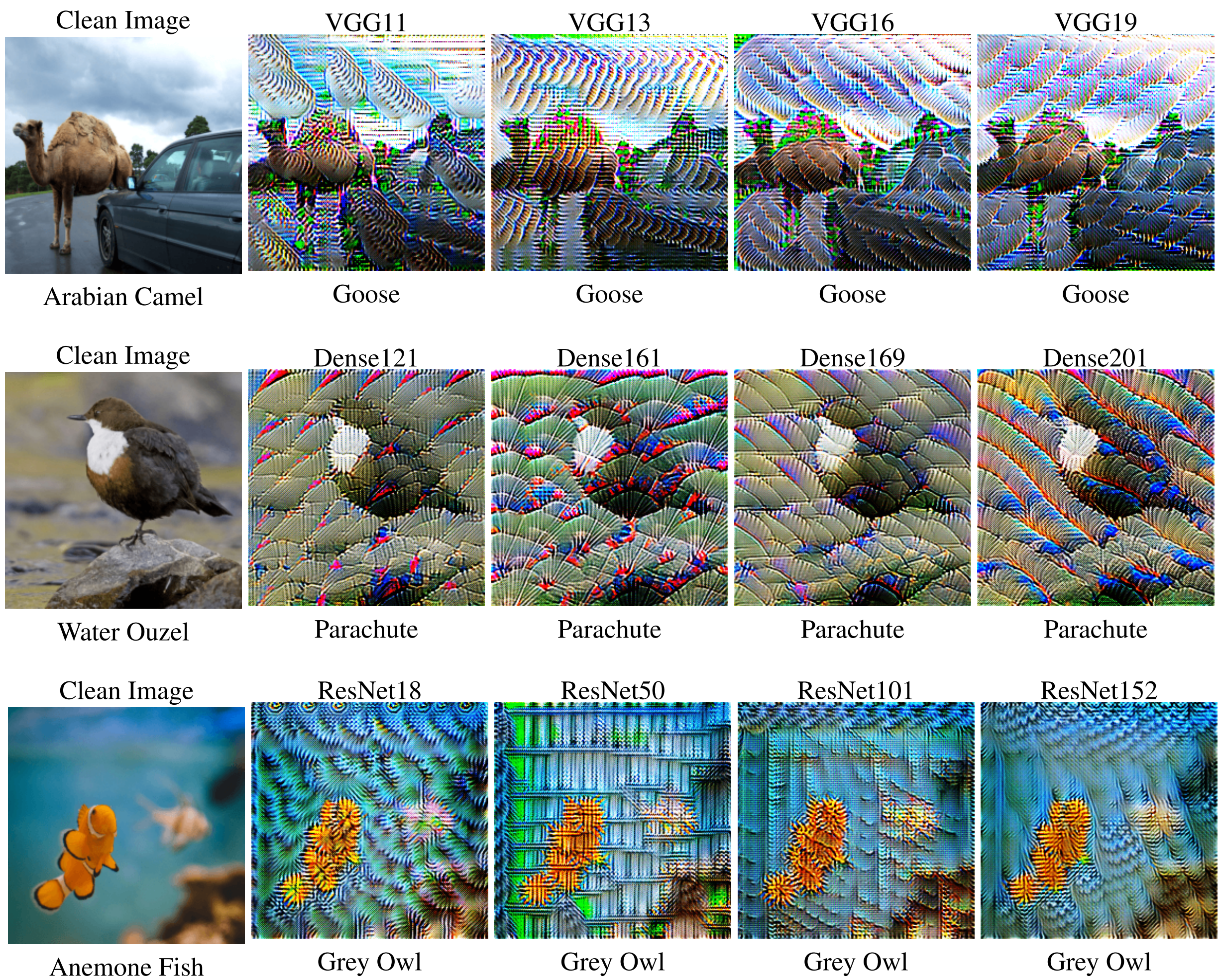


Figure 10: ***Evolution of* TTP*:*** Unconstrained targeted adversarial patterns generated by TTP are shown to demonstrate how TTP evolves as it learns perturbations from different source models of a certain family of networks.

References

[1] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 3

[2] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. 1, 2, 3

[3] Yiwen Guo, Qizhang Li, and Hao Chen. Backpropagating linearly improves transferability of adversarial examples. *arXiv preprint arXiv:2012.03528*, 2020. 1, 3

[4] Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. AugMix: A simple data processing method to improve robustness and uncertainty. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020. 3

[5] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 1

[6] Nathan Inkawhich, Kevin Liang, Lawrence Carin, and Yiran Chen. Transferable perturbations of deep feature distributions. In *International Conference on Learning Representations*, 2020. 3

[7] Nathan Inkawhich, Kevin J Liang, Binghui Wang, Matthew Inkawhich, Lawrence Carin, and Yiran Chen. Perturbing across the feature hierarchy to improve standard and strict blackbox attack transferability. *arXiv preprint arXiv:2004.14861*, 2020. 3

[8] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 1, 2

[9] Maosen Li, Cheng Deng, Tengjiao Li, Junchi Yan, Xinbo Gao, and Heng Huang. Towards transferable targeted attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 641–649, 2020. 3

[10] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. 3

[11] Muzammal Naseer, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Fatih Porikli. A self-supervised approach for adversarial robustness. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[12] Muzammal Naseer, Salman Khan, and Fatih Porikli. Local gradients smoothing: Defense against localized adversarial attacks. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1300–1307. IEEE, 2019. 3

[13] Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust imagenet models transfer better? In *ArXiv preprint arXiv:2007.08489*, 2020. 3

[14] Dongxian Wu, Yisen Wang, Shu-Tao Xia, James Bailey, and Xingjun Ma. Skip connections matter: On the transferability of adversarial examples generated with resnets. In *ICLR*, 2020. 1, 3

[15] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan Yuille. Improving transferability of adversarial examples with input diversity. In *Computer Vision and Pattern Recognition*. IEEE, 2019. 3