NAT: Learning to Attack Neurons for Enhanced Adversarial Transferability

Krishna Kanth Nakka
VITA Lab, EPFL, Switzerland
krishkanth.92@gmail.com

Alexandre Alahi
VITA Lab, EPFL, Switzerland
alexandre.alahi@epfl.ch

1. Implementation

We provide below additional details for the reproducibility of our experiments.

Libraries. We conducted our experiments on NVIDIA GeForce RTX 3090 using PyTorch 2.1.1 [28], CUDA 11.8, Timm 1.0.9 [44], and Torchvision 0.16.1 [25].

Generator Modifications. Due to the non-deterministic behavior[1] of the `ReflectionPad2d` operation, which was used inside the generator architecture in prior works [27, 33, 50], we removed it in the `block1` layer and set padding to 1 in the subsequent `conv2d` operation within the same layer. We also removed `ReflectionPad2d` in the `blockf` layer and set the padding to 5 in the `conv2d` operation. Furthermore, the `residualblock` implementation contains two instances of `ReflectionPad2d`, both of which were removed, with padding set to 1 in the subsequent `conv2d` operations. These changes allowed us to eliminate `ReflectionPad2d` from the generator while compensating for changes in feature size with additional padding in the subsequent convolution operations.

Top-k Neurons After Initial Lightweight Training. In Section 3.1, we outlined that all 512 generators, corresponding to the 512 neurons in layer 18 of VGG16, undergo an initial lightweight training phase using 3.125% of the training set. After this phase, we selected the top-$k = 40$ generators based on their performance, either on the source VGG16 [35] or by using a held-out model such as DenseNet121 [13] or ResNet [11] in our experiments. We provide the top-ranked neurons from this step across different held-out models in Table 2. Additionally, for full transparency, we present the performance of all 512 neurons after the lightweight training phase on two heldout models in Figures 1, and 2 for DenseNet121 [13] and VGG16 [35] heldout models.

Target Models. We used 41 target models, publicly available in Torchvision [25] and Timm libraries [44]. We provide in Table 1 the exact version details used for each model and also report the clean accuracy on the evaluation set of 5K images [33], sampled from the ImageNet validation set [31]. We used 9 models for cross-domain provided by the authors of BIA [50]

[1] https://github.com/pytorch/pytorch/issues/98925

2. Additional Quantitative Results

Performance of each neuron-specific generator. In Tables 3 and 4, we provide the performance of all fully trained top-40 generators selected using DenseNet121 [13] as the held-out model. We observed that our neuron-specific generators outperform the baselines in 37 out of 40 cases in the single-query setting. Moreover, the best-performing generator G_{391}, attacking neuron 391, outperforms the highest-ranked generator G_{250} obtained at the end of lightweight training by more than 4%. We believe that with better neuron ranking schemes, the fooling rate in the single-query setting can be further enhanced.

Choice of Heldout model for top-k neuron selection. In Figure 3, we show the performance of top-40 generators trained with different models for neuron selection versus the number of queries. We observe that our top-k neuron position is robust to the choice of heldout model selection. And in all cases, top-1 generator consisted outperformed baselines in the single-query setting.

3. Additional Qualitative Results

We present the visualizations of adversarial images generated by different neuron-specific generators in Figures 4, 5, 6, and 7. Next to each adversarial image, we show the synthesized images that is optimized to have large activation magnitudes for the attacked neuron using activation maximization algorithm [34]. A clear visual correlation can be observed between the synthesized images and the adversarial patterns in the generated images, indicating that the perturbations are effectively disrupting neurons associated with specific concepts.

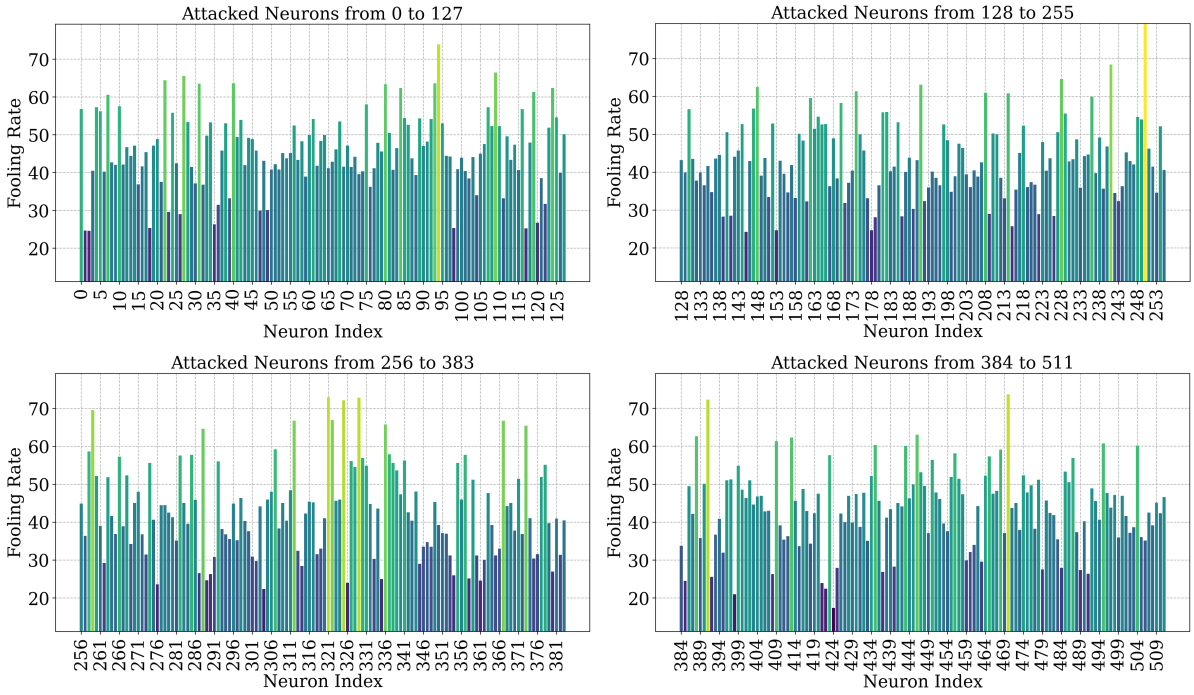


Figure 1. **Evaluation of 512 Neuron specific generators on DenseNet121 [13] obtained after end of lightweight training on VGG16 [35].** We observe that some filters are more transferable than others and we choose top-k generators that has highest transferability.

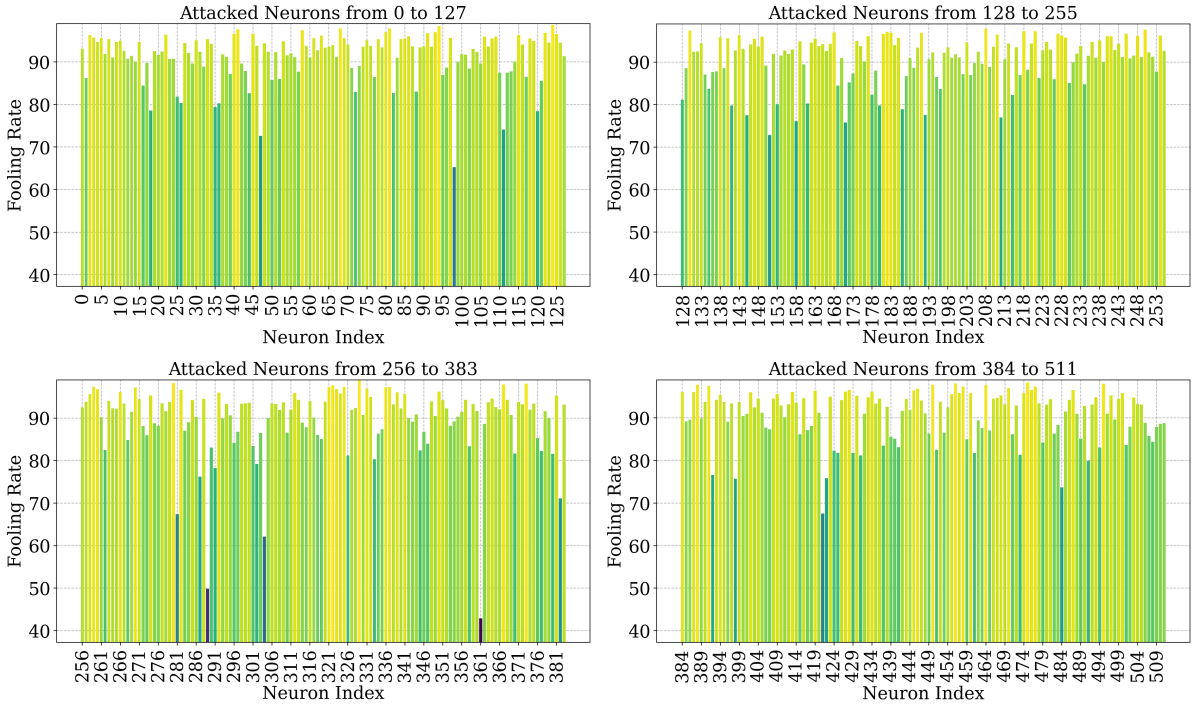
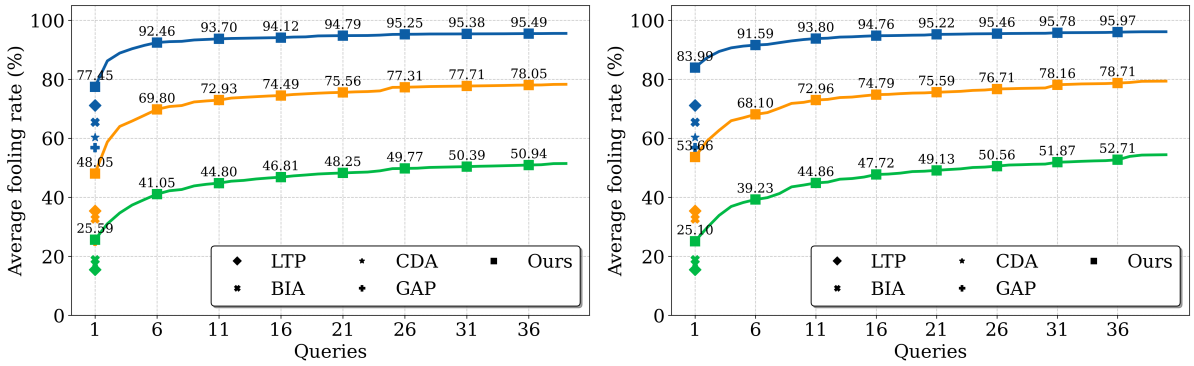


Figure 2. **Evaluation of 512 Neuron specific generators on VGG16 [35] obtained after end of lightweight training on the same VGG16 [35].** We observe that all generators have high fooling rates suggesting that overfitting to the source VGG16 [35] model.



(a) VGG16 [35]

(b) ResNet152 [11]

Figure 3. **Impact of heldout model for top-k neuron selection after lightweight training** We use the same source model VGG16 [35] as the heldout model for top-k neuron selection after lightweight training in the left and similarily using ResNet152 [11] on the right. We observe that for both the cases, the fooling rate with k-queries differ marginally and perform along the same lines as reported in the main paper with DenseNet121 [13] model.

Model	Version	Accuracy	Source
VGG [35]	vgg16	71.07%	TorchVision [25]
Resnet [11]	resnet152	78.05%	TorchVision [25]
Densenet [13]	densenet121	74.72%	TorchVision [25]
Squeezenet1 [14]	squeezenet1_1	58.58%	TorchVision [25]
Shufflenet [23]	shufflenetv2	69.52%	TorchVision [25]
Mobilenet [12]	mobilenet_v3	74.84%	TorchVision [25]
Mnasnet [36]	mnasnet1_0	73.84%	TorchVision [25]
Wide Resnet [49]	wide_resnet50_2	78.74%	TorchVision [25]
Convnext [22]	convnext_base	84.05%	TorchVision [25]
Efficientnet [16]	efficientnet_v2_m	80.00%	TorchVision [25]
Regnet [30]	regnet_x_32gf	83.62%	TorchVision [25]
Alexnet [17]	alexnet	56.88%	TorchVision [25]
ViT [1]	vit_b_16	81.06%	TorchVision [25]
ViT [1]	vit_l_16	80.01%	TorchVision [25]
Swin [21]	swin_base_patch4_window7_224	85.62%	timm [44]
BeiT [2]	beit_base_patch16_224	85.68%	timm [44]
DeiT [38]	deit3_base_patch16_224	84.01%	timm [44]
Mixer [37]	mixer_b16_224	77.02%	timm [44]
Convmixer [40]	convmixer_768_32	80.78%	timm [44]
Efficientvit [20]	efficientvit_m3	73.54%	timm [44]
Xception [6]	xception	75.54%	timm [44]
Crossvit [4]	crossvit_base_240	82.56%	timm [44]
Csp [42]	cspresnet50	78.66%	timm [44]
Davit [7]	davit	84.86%	timm [44]
Edgenext [24]	edgenext	83.07%	timm [44]
Gcvit [10]	gcvit_base	85.32%	timm [44]
Ghostnet [9]	ghostnetv2_100	75.92%	timm [44]
Visformer [5]	visformer_small	82.22%	timm [44]
Focalnet [46]	focalnet_base_lrf	84.54%	timm [44]
Hiera [32]	hiera_base_224	85.02%	timm [44]
Hrnet [43]	hrnet_w32	79.03%	timm [44]
Maxvit [41]	maxvit_base_tf_224	85.52%	timm [44]
Convformer [48]	convformer_s36	84.56%	timm [44]
Mobilevit [26]	mobilevitv2_150	79.08%	timm [44]
Pnasnet [19]	pnasnet5large	78.52%	timm [44]
Rdnet [15]	rdnet_base	85.14%	timm [44]
Levit [8]	levit_conv_256	82.18%	timm [44]
Mvit [18]	mvitv2_base	85.08%	timm [44]
Nf [3]	nf_resnet50	79.86%	timm [44]
Coat [45]	coat_lite_small	82.76%	timm [44]
Cait [39]	cait_s24_224	84.03%	timm [44]
Dla [47]	dla102	78.54%	timm [44]

Table 1. Details of all 41 target models used in our evaluation, along with their clean accuracy on 5K evaluation ImageNet dataset provided by LTP [33].

Heldout model	Top-40 Neuron locations
DenseNet121	250, 94, 470, 321, 329, 391, 325, 259, 241, 322, 312, 367, 109, 336, 27, 373, 288, 228, 22, 93, 40, 31, 80, 191, 446, 388, 148, 124, 84, 413, 409, 174, 119, 208, 495, 214, 7, 435, 504, 443
ResNet152	250, 94, 22, 259, 93, 321, 367, 325, 391, 27, 470, 329, 228, 119, 504, 322, 288, 214, 307, 292, 130, 40, 109, 413, 312, 435, 208, 336, 5, 495, 373, 236, 446, 84, 80, 124, 75, 266, 241, 218
VGG16	329, 124, 94, 475, 280, 456, 373, 495, 367, 68, 81, 450, 208, 388, 464, 322, 41, 250, 391, 130, 58, 458, 321, 477, 259, 337, 325, 336, 221, 218, 270, 182, 470, 93, 168, 80, 331, 191, 446, 183

Table 2. **Attacked Neuron Positions.** We provide the list of the top-40 neurons selected based on their transferability performance on the held-out model. Note that all 512 generators were initially trained on just 3.12% of the dataset, with each generator specifically targeting an individual neuron in layer 18 of the source model VGG16.

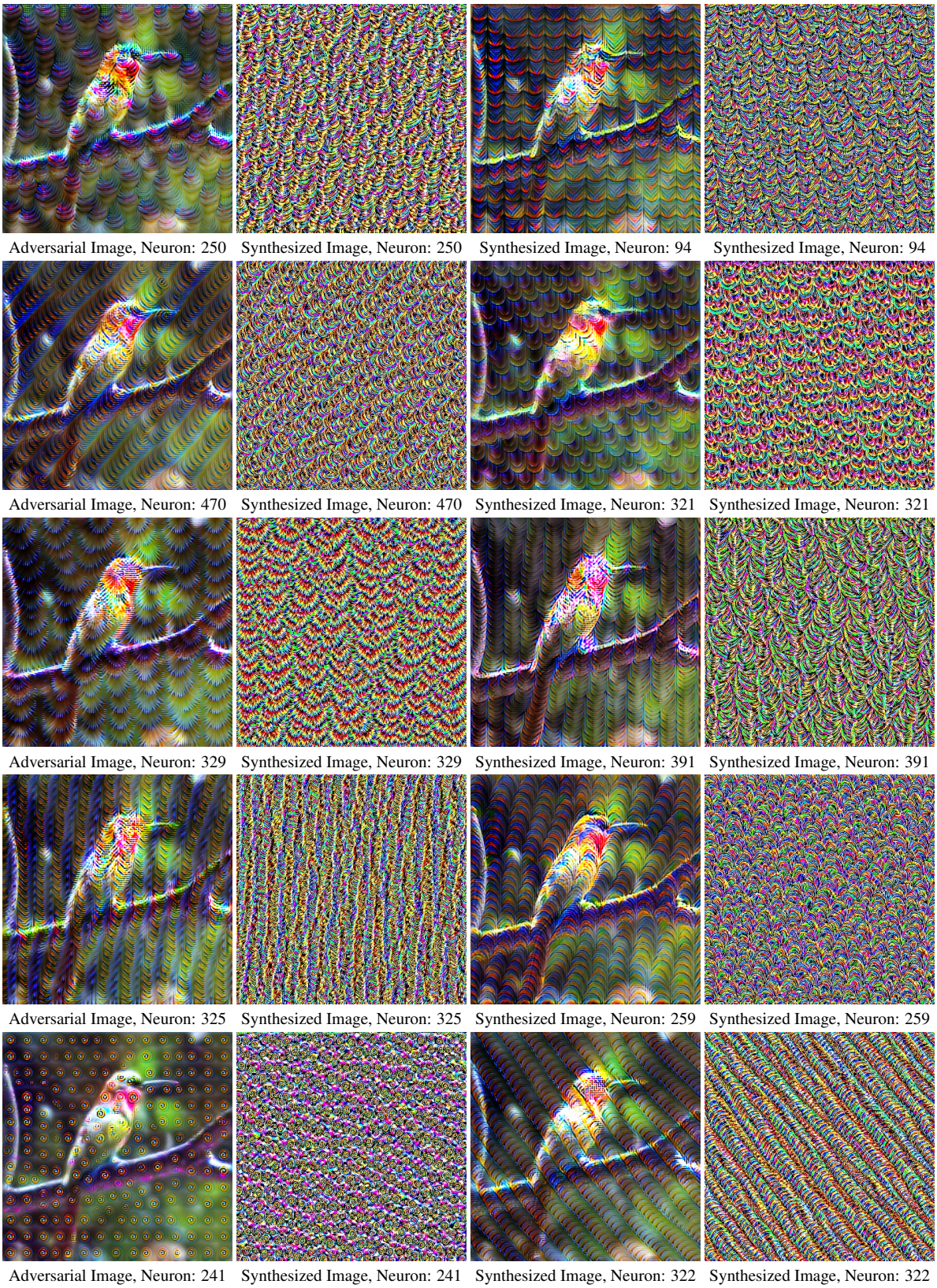


Figure 4. Generated unbounded adversarial images along with synthesized neuron visualizations for the top 10 attacked neurons. The positions of the neurons are listed below each image. Note that, the top-k neurons were selected based on the transferability to the DenseNet121 [13] as the held-out model after initial lightweight training.

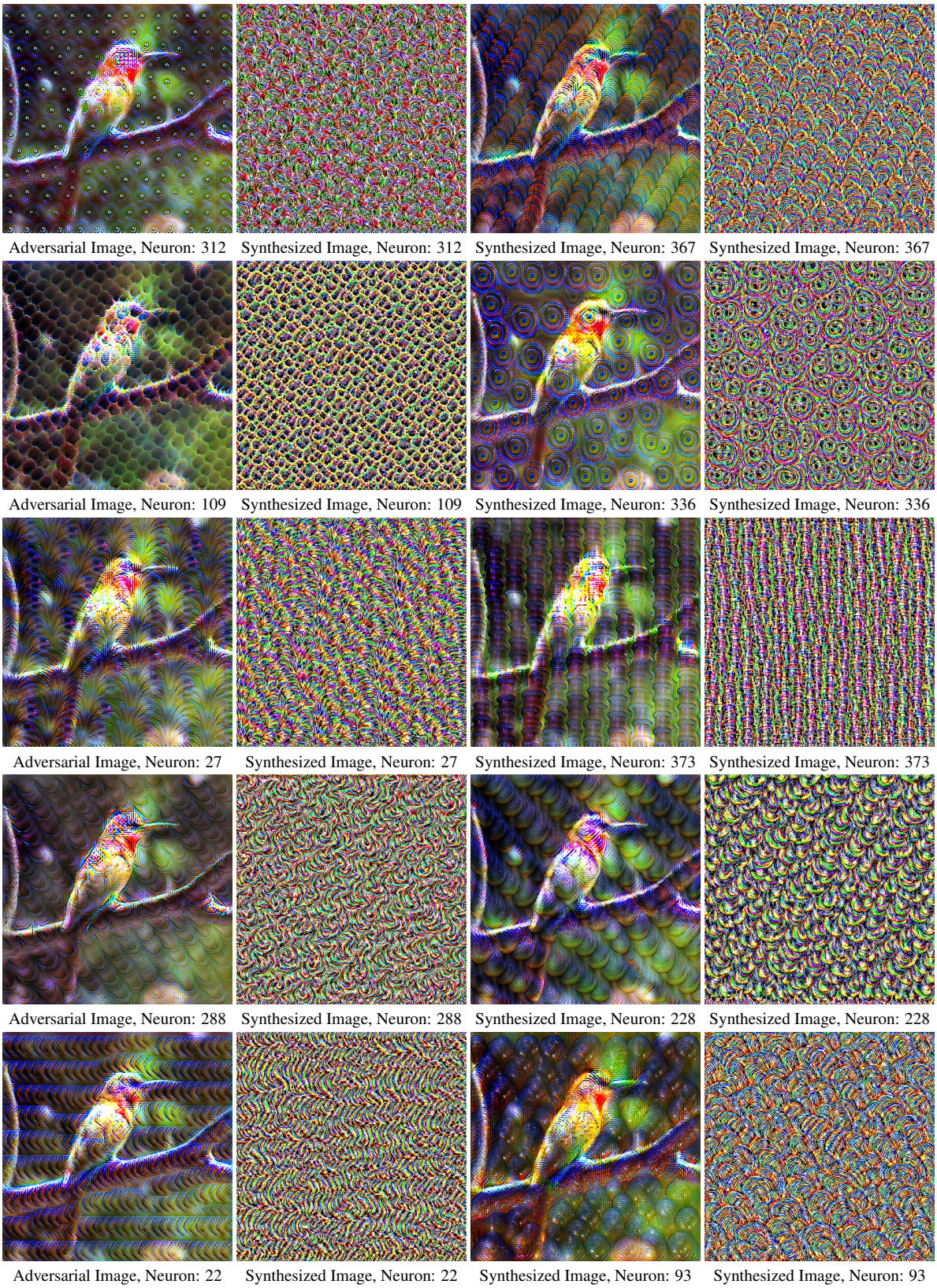


Figure 5. Generated unbounded adversarial images along with synthesized neuron visualizations for the top 10 to 20 attacked neurons. The positions of the neurons are listed below each image. Note that, the top-k neurons were selected based on the transferability to the DenseNet121 [13] as the held-out model after initial lightweight training

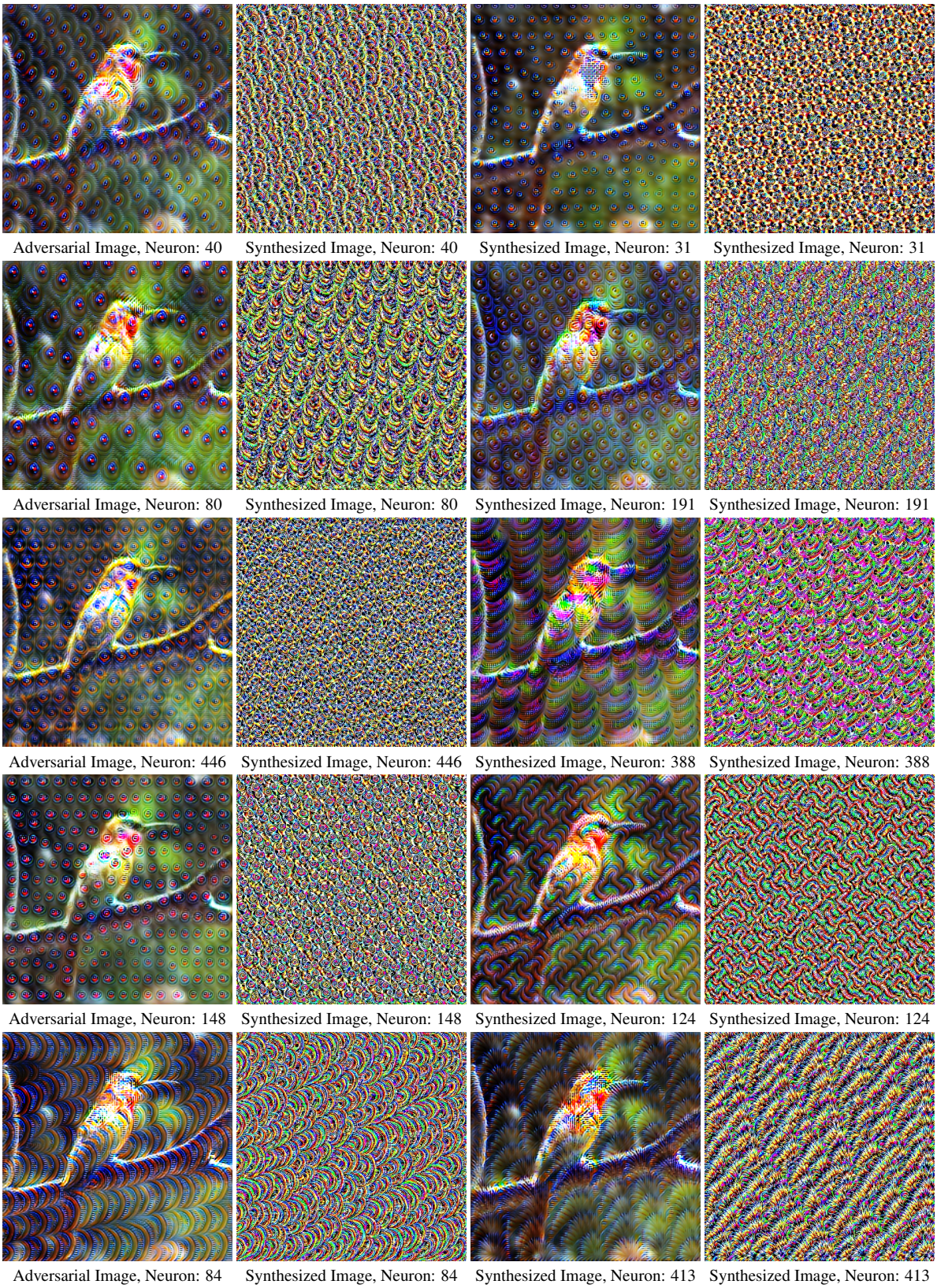


Figure 6. Generated unbounded adversarial images along with synthesized neuron visualizations for the top 20 to 30 attacked neurons. The positions of the neurons are listed below each image. Note that, the top-k neurons were selected based on the transferability to the DenseNet121 [13] as the held-out model after initial lightweight training

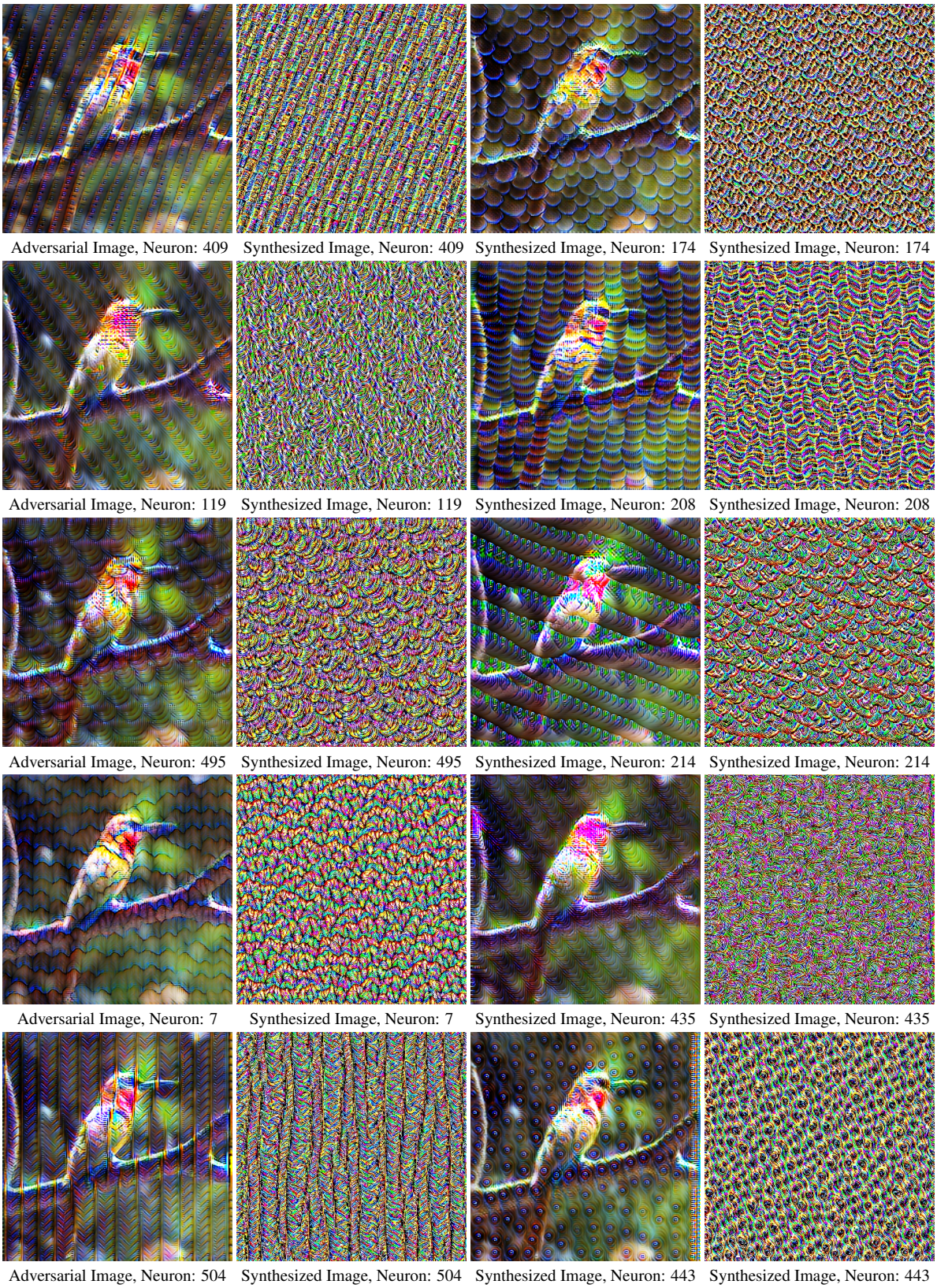


Figure 7. Generated unbounded adversarial images along with synthesized neuron visualizations for the top 30 to 40 neurons. . The positions of the neurons are listed below each image. Note that, the top-k neurons were selected based on the transferability to the DenseNet121 [13] as the held-out model after initial lightweight training

Model	250	94	470	321	329	391	325	259	241	322	312	367	109	336	27	373	288	228	22	93
ResNet152 [11]	89.10	90.66	84.34	88.26	80.44	**94.28**	79.14	90.06	66.70	80.22	73.14	85.76	77.84	75.66	79.68	65.40	77.28	79.16	81.54	78.00
DenseNet121 [13]	91.02	89.56	91.44	90.70	89.44	**94.68**	82.22	91.94	85.70	84.52	88.28	86.84	84.74	89.40	81.32	75.64	85.98	83.94	80.62	77.86
SqueezeNet1 [14]	89.46	83.04	88.24	85.22	89.56	86.94	82.46	81.16	87.04	72.98	90.44	78.38	79.50	**95.20**	88.36	74.78	83.28	87.34	87.06	84.90
ShuffleNet [23]	89.02	82.86	**91.64**	80.36	84.56	91.10	83.48	90.50	75.28	80.48	86.70	84.34	70.12	84.88	87.02	65.64	84.58	84.62	82.04	87.52
MNasNet [36]	93.32	86.32	**95.58**	87.28	87.72	93.12	85.76	88.54	85.28	90.60	93.16	88.58	84.48	92.30	84.86	79.54	89.20	94.16	88.90	89.22
PNasNet [19]	63.64	36.88	58.72	52.46	56.18	57.08	43.82	52.04	30.94	66.60	43.68	45.38	52.10	65.52	55.24	36.56	62.66	**69.38**	57.30	54.72
MobileNet [12]	84.26	66.86	**86.66**	71.22	80.98	86.22	75.94	76.26	57.00	71.88	71.58	72.46	54.34	80.42	82.08	55.66	78.64	79.08	79.72	79.48
WideResNet [49]	92.20	90.86	91.82	93.14	87.60	**93.32**	85.16	92.12	81.22	90.94	83.68	93.26	80.00	85.94	84.56	76.22	82.68	85.48	84.60	85.48
RegNet [30]	91.20	94.30	93.78	**96.78**	82.66	96.10	83.62	86.86	72.66	91.88	81.74	88.72	72.50	82.98	80.08	76.28	83.04	82.72	79.32	78.28
AlexNet [17]	49.82	47.28	46.68	43.86	48.40	41.98	42.94	45.28	38.04	48.92	41.12	42.20	49.22	**56.68**	50.06	49.24	46.12	55.66	49.94	46.90
NF ResNet [3]	76.86	75.34	77.98	77.94	65.46	**82.92**	56.56	80.38	45.62	76.88	59.26	76.34	63.22	68.26	63.34	54.32	69.86	63.04	70.76	64.86
DLA	97.86	96.74	96.72	97.10	95.16	**98.64**	94.70	98.00	92.26	96.46	94.98	97.96	89.98	91.98	95.16	94.28	93.28	93.88	93.66	94.64
GhostNetV2 [9]	82.68	75.54	**89.84**	84.04	76.50	86.80	75.06	83.50	58.90	83.82	75.10	83.08	65.74	79.78	79.52	65.00	80.42	82.24	80.54	77.70
CSPResNet [42]	83.38	78.54	83.10	78.60	67.74	**88.98**	64.08	82.80	64.46	75.36	72.68	78.08	61.24	77.24	70.92	52.32	73.28	71.20	74.10	72.32
HRNet [43]	97.10	95.20	94.14	**97.46**	90.32	97.34	88.60	89.08	90.66	93.78	94.30	92.36	89.48	95.48	85.06	92.10	90.82	92.52	91.22	87.44
Xception [6]	73.82	62.84	66.22	**74.48**	66.10	71.66	51.88	70.46	50.14	71.38	57.42	57.56	70.32	68.82	60.60	49.60	64.40	70.86	64.68	56.52
ViT_B_16 [1]	17.96	14.60	14.30	14.88	18.30	16.32	14.22	15.10	11.48	16.98	13.40	13.92	**18.38**	15.48	17.32	15.30	16.40	18.06	17.20	15.56
ViT_L_16 [1]	18.08	14.90	17.18	16.10	**19.14**	17.88	14.22	17.82	12.40	17.58	14.76	15.30	17.78	17.96	18.14	15.36	18.24	18.74	17.28	17.06
Swin [21]	28.78	24.52	30.94	28.86	27.26	**33.76**	21.34	28.20	17.62	21.58	22.16	25.46	18.32	21.66	25.46	19.32	25.80	25.52	25.02	24.54
BeIT [2]	**30.48**	16.22	25.60	22.04	23.46	27.94	18.36	24.70	12.00	24.02	16.46	17.50	21.40	19.64	29.20	18.30	21.44	28.32	25.38	21.94
DeiT [38]	20.54	17.42	19.02	19.64	20.32	22.14	17.48	19.84	11.90	20.38	14.86	17.08	18.44	17.84	**24.08**	16.92	17.50	18.68	18.60	15.76
Cait [39]	27.62	20.42	24.52	22.40	**31.72**	29.08	18.22	24.24	13.48	22.52	18.46	20.54	27.36	20.46	25.22	18.24	23.92	22.66	22.92	22.14
Davit [7]	31.62	31.70	37.70	36.50	**41.20**	40.84	24.46	34.48	23.46	31.14	27.16	37.10	19.66	19.94	24.90	21.22	32.60	27.98	27.72	29.50
ConvNext [22]	70.90	72.64	75.02	71.64	63.38	**83.56**	73.76	75.00	47.54	74.94	51.76	66.60	73.76	57.08	63.24	69.12	49.94	60.04	55.50	54.38
Mixer [37]	32.38	28.32	35.92	31.82	**42.82**	35.18	28.66	30.82	24.72	33.50	29.48	29.34	29.74	30.98	35.88	25.40	36.88	31.44	32.92	30.48
ConvMixer [40]	76.12	52.04	75.24	54.44	72.02	**84.60**	66.78	67.82	41.76	71.50	55.02	73.14	61.04	55.74	67.46	51.90	58.08	70.04	60.02	66.34
CrossViT [4]	19.18	16.50	17.92	17.44	17.70	**21.40**	15.36	17.36	12.28	16.16	16.00	14.50	18.26	17.02	17.78	15.94	19.82	17.24	17.68	15.62
Edgenext [24]	73.26	64.30	69.48	74.38	64.24	**80.80**	58.90	70.72	44.98	74.54	49.44	68.72	58.90	52.40	65.96	54.38	60.10	62.02	55.52	59.90
GCViT [10]	44.54	38.94	44.48	40.98	35.30	**49.38**	30.38	39.60	24.32	34.20	31.36	45.92	22.00	24.22	27.88	22.92	37.62	33.92	30.86	36.12
Visformer [5]	71.10	62.96	68.78	59.32	79.26	**80.60**	55.74	67.10	39.36	66.38	47.78	69.04	40.48	40.36	58.78	41.48	53.32	57.72	59.92	56.92
FocalNet [46]	45.66	36.48	47.98	44.82	37.76	**54.32**	35.60	38.32	31.02	40.14	35.12	39.82	27.88	30.84	37.34	25.24	42.38	38.36	33.40	36.38
Hiera [32]	43.16	33.72	43.56	36.40	38.92	**44.78**	31.18	41.14	19.82	34.50	24.22	36.46	21.30	24.92	37.04	24.58	31.56	38.72	32.80	34.06
MaxVit [41]	22.94	17.98	21.00	20.96	17.40	**25.60**	17.54	19.22	15.38	17.42	19.30	19.86	16.46	15.56	16.34	15.30	20.04	18.32	17.18	18.70
Conformer [48]	47.72	38.44	45.98	42.24	**53.34**	51.72	33.54	37.30	27.68	43.56	32.64	42.04	29.26	29.14	39.54	28.68	38.18	46.40	42.96	42.60
MobileViT [26]	92.06	91.96	88.10	91.12	68.42	**94.42**	81.82	90.68	72.34	87.40	79.12	88.36	61.04	78.04	76.06	85.48	80.62	84.84	80.42	79.52
RDNet [15]	50.72	42.44	44.28	47.88	36.58	**58.78**	33.00	41.88	25.68	43.76	31.04	43.74	21.82	24.98	28.28	28.96	36.62	37.52	34.72	37.34
LeViT [8]	67.52	58.20	66.52	62.00	63.98	**78.00**	61.14	75.12	37.40	62.40	46.94	66.24	43.46	46.66	62.48	39.78	55.86	58.26	60.10	57.04
MViT [18]	35.38	30.04	37.50	34.04	36.40	**43.42**	24.24	36.56	19.64	30.90	27.14	38.20	20.28	23.00	25.56	21.78	32.66	28.44	29.96	31.32
Coat [45]	52.14	55.50	51.10	52.22	36.20	**70.70**	42.32	45.70	32.08	39.02	34.52	46.00	30.94	30.28	34.56	30.14	45.92	38.98	34.86	39.74
EfficientViT [20]	60.04	44.78	66.24	49.00	62.44	68.62	51.62	65.74	41.48	52.02	54.46	59.30	41.32	61.32	**70.62**	43.32	58.10	57.48	62.30	57.30
EfficientNet [16]	60.64	48.00	**64.40**	49.48	55.08	63.26	43.48	55.00	36.96	51.92	44.70	47.54	35.40	54.68	47.72	36.64	50.08	54.62	54.70	48.40
Average	60.62	54.29	60.24	57.06	56.38	64.35	50.95	58.01	43.38	56.22	49.87	55.93	47.30	51.73	53.77	45.08	53.88	55.36	53.56	52.79

Table 3. **Fooling rate of top-20 generators on the 41 target models.** The attack neuron position of each generator is listed along with the columns. Note that the ranking here is obtained by using DenseNet121 [13] as the held-out model after lightweight training. The best performing generator for each target model is highlighted in bold, and the second-best is underlined for each target model. For comparison, baselines LTP [33], BIA [50], CDA [27], GAP [29] achieves average fooling rate of 44.6%, 42.0% 40.7% and 34.7% over 41 architectures. All the top-20 generators obtained with our method outperform the baselines in all cases.

Model	40	31	80	191	446	388	148	124	84	413	409	174	119	208	495	214	7	435	504	443
ResNet152 [11]	77.40	59.46	71.60	69.08	70.58	77.58	70.36	77.84	73.18	74.28	50.26	74.76	**81.86**	63.72	79.66	60.22	65.26	74.66	70.74	67.00
DenseNet121 [13]	82.98	76.14	81.00	83.74	81.30	84.80	**86.12**	84.82	78.62	76.44	62.36	84.96	82.52	70.08	83.20	67.54	74.46	77.94	73.08	80.94
SqueezeNet1 [14]	88.50	81.90	88.58	85.72	85.04	85.54	**89.34**	79.86	82.06	83.12	79.26	80.60	84.52	63.52	86.14	72.84	73.24	79.94	68.16	85.42
ShuffleNet [23]	77.08	63.02	78.12	76.46	75.28	72.70	75.00	69.42	72.40	83.74	50.74	79.88	**87.36**	51.74	84.36	69.26	64.04	83.84	67.92	70.68
MNasNet [36]	91.56	81.46	84.56	90.60	91.68	**92.90**	90.50	86.96	90.48	85.34	87.82	90.96	89.30	71.72	91.72	81.34	76.40	88.00	71.80	87.54
PNasNet [19]	54.46	31.46	53.46	38.80	37.78	55.42	33.68	48.72	56.86	54.60	35.44	63.28	62.34	43.00	**64.08**	50.70	40.84	56.48	30.78	37.92
MobileNet [12]	75.44	52.08	66.92	70.42	62.28	76.12	65.18	69.04	**81.80**	77.00	63.28	62.96	78.28	58.12	75.44	63.72	64.02	75.72	50.84	66.16
WideResNet [49]	**86.34**	78.88	82.54	84.20	85.98	85.34	85.70	84.86	83.78	80.12	67.64	84.36	85.02	70.24	82.16	71.56	75.94	82.86	74.38	82.96
RegNet [30]	82.54	75.20	79.58	82.94	80.68	83.40	83.82	**86.92**	83.06	74.70	63.60	75.72	82.68	69.62	85.48	76.34	71.98	77.58	63.52	79.02
AlexNet [17]	52.66	48.66	51.64	47.66	50.34	50.14	41.00	49.56	**56.54**	51.62	54.10	52.02	45.16	55.16	46.28	42.10	47.56	39.38	38.08	49.78
NF ResNet [3]	63.94	45.60	56.58	50.26	50.12	63.12	54.76	66.86	71.12	65.46	34.94	**71.50**	61.98	57.66	64.08	46.52	52.96	55.92	49.28	53.46
DLA	93.68	89.12	89.34	93.98	93.48	95.44	95.76	94.66	95.54	92.10	90.70	94.74	95.66	86.42	**96.20**	90.16	87.84	94.22	84.40	92.92
GhostNetV2 [9]	80.74	51.24	67.92	**86.42**	74.28	82.60	79.88	77.52	79.28	75.90	69.52	79.06	81.02	64.12	83.26	69.44	66.20	75.02	60.60	69.36
CSPResNet [42]	**74.20**	57.64	68.72	65.66	63.90	70.02	72.34	69.26	71.24	65.36	38.90	64.00	72.14	61.56	70.34	53.18	59.28	65.50	56.26	66.30
HRNet [43]	94.78	92.24	92.46	93.18	91.74	95.14	**96.28**	95.74	90.54	83.20	88.68	88.74	86.74	89.82	92.76	88.22	79.24	86.14	77.78	91.36
Xception [6]	65.28	46.96	54.94	55.02	54.84	61.12	55.12	66.74	65.86	57.42	49.12	**74.86**	60.72	59.42	61.84	55.32	63.30	52.18	44.64	55.66
ViT_B_16 [1]	14.42	13.62	15.98	14.60	14.24	18.00	13.28	15.82	18.26	18.68	16.36	**22.56**	14.56	17.48	17.04	15.34	17.30	12.64	13.14	14.56
ViT_L_16 [1]	15.88	15.46	17.20	14.92	15.94	18.44	13.24	17.96	20.64	19.90	15.42	**20.96**	16.74	18.68	17.88	16.64	17.12	15.06	13.72	15.08
Swin [21]	20.38	20.22	21.72	17.98	16.70	23.06	18.36	22.40	26.12	22.82	14.26	20.06	**26.58**	19.58	25.76	21.64	22.78	20.50	16.08	19.22
BeIT [2]	17.14	13.94	19.42	15.12	14.50	23.46	13.96	18.24	24.10	**30.00**	16.38	22.90	20.02	21.18	23.72	18.56	21.12	14.24	14.92	16.76
DeiT [38]	15.44	12.52	14.78	14.24	13.68	16.88	12.74	14.96	**19.84**	17.92	12.54	17.94	17.68	16.46	18.50	15.76	16.40	14.90	14.32	13.96
Cait [39]	20.34	16.28	23.60	17.60	15.50	22.38	15.14	**27.60**	26.20	24.70	15.70	26.48	21.14	22.68	24.32	18.34	20.32	17.26	14.78	16.98
Davit [7]	24.18	21.12	20.62	23.78	22.42	24.06	25.76	28.62	26.72	25.62	17.18	22.66	28.00	21.12	27.22	21.48	**28.96**	25.82	18.82	23.68
ConvNext [22]	63.86	45.16	57.24	49.64	44.74	60.28	57.66	**84.10**	58.86	55.10	43.84	59.22	58.58	57.36	59.12	54.98	58.36	48.92	52.90	59.32
Mixer [37]	29.56	26.30	28.46	26.54	26.02	31.26	25.26	29.34	32.88	**34.92**	24.80	30.44	33.84	32.36	33.92	27.86	31.28	29.94	25.64	27.36
ConvMixer [40]	64.28	45.70	52.48	50.88	56.56	55.92	48.00	**74.74**	57.68	66.12	36.38	73.74	69.60	49.58	68.88	42.38	56.20	54.64	36.88	57.82
CrossViT [4]	15.88	14.30	15.70	13.94	14.78	16.26	13.04	17.32	18.46	17.60	13.50	**18.64**	17.70	17.66	18.08	14.76	17.62	14.20	13.88	14.52
Edgenext [24]	56.52	43.66	56.44	56.36	46.36	65.96	52.34	65.06	**67.22**	57.56	42.84	61.60	56.40	55.18	65.56	50.64	52.26	48.26	45.22	50.08
GCViT [10]	30.22	23.96	23.32	26.48	24.84	26.70	27.74	**39.34**	34.08	26.20	18.26	26.54	39.18	26.06	34.58	24.92	30.92	29.02	19.52	28.52
Visformer [5]	49.90	40.92	47.00	43.70	41.90	61.82	47.02	54.00	63.68	56.44	35.62	46.24	**63.82**	43.54	62.18	41.98	44.16	52.28	35.38	43.32
FocalNet [46]	32.56	29.00	30.86	36.44	29.72	33.62	32.00	**44.56**	35.80	30.12	23.12	29.68	44.22	25.50	40.78	28.74	32.26	33.34	22.04	36.48
Hiera [32]	26.24	21.48	22.20	22.76	21.76	39.68	21.48	26.50	37.98	36.04	21.32	27.62	39.58	23.90	**39.86**	25.38	25.02	31.40	21.10	22.86
MaxVit [41]	18.60	16.58	14.28	16.06	14.72	15.94	17.04	19.00	15.86	17.18	12.72	16.68	20.30	17.94	17.92	17.04	**20.44**	16.08	13.54	16.64
Conformer [48]	34.94	31.72	37.32	31.64	28.08	43.10	31.36	**46.12**	42.92	42.20	25.64	33.84	40.84	38.68	40.34	36.18	38.38	31.88	24.66	30.18
MobileViT [26]	77.08	69.44	76.70	82.36	78.18	85.88	80.64	**90.80**	84.96	66.94	68.46	83.98	76.54	79.54	84.96	72.88	82.84	67.14	69.42	76.64
RDNet [15]	33.00	24.82	27.26	27.80	26.88	33.38	27.88	34.66	33.14	29.86	21.88	27.40	**37.90**	32.58	36.16	26.62	33.52	34.90	23.30	27.58
LeViT [8]	47.30	35.56	40.02	48.36	41.72	60.04	42.58	51.94	60.72	57.92	31.44	51.76	54.94	37.72	**64.62**	35.32	42.70	46.82	36.16	40.10
MViT [18]	26.34	19.56	22.88	22.02	21.46	25.84	21.04	31.20	**35.40**	25.24	16.46	26.62	31.76	27.02	28.84	21.90	28.86	26.18	19.00	21.46
Coat [45]	36.18	32.02	30.48	33.52	31.24	39.94	33.44	**43.46**	42.04	36.14	26.74	36.70	41.34	37.52	42.10	30.22	33.70	32.60	29.18	32.58
EfficientViT [20]	52.70	42.14	47.56	48.86	45.26	57.40	44.20	47.20	66.62	**67.20**	40.26	48.32	63.62	44.14	61.38	42.54	41.06	55.48	38.68	51.48
EfficientNet [16]	51.92	35.96	47.58	48.26	42.92	48.76	42.32	48.94	**60.02**	45.60	27.78	45.90	46.42	41.68	47.92	40.70	37.86	41.88	31.56	41.76
Average	51.62	42.50	48.27	48.24	46.33	53.16	47.62	53.72	54.70	51.42	39.64	51.73	54.11	45.39	54.84	44.42	46.68	48.31	40.15	47.21

Table 4. **Fooling rate of next 20 generators on the 41 target models.** The attack neuron position of each generator is listed along with the columns. Note that the ranking here is obtained by using DenseNet121 [13] as the heldout model after lightweight training.The best performing generator is highlighted in bold and second best is underlined for each target model.For comparison, baselines LTP [33], BIA [50], CDA [27], GAP [29] achieves average fooling rate of 44.6%, 42.0% 40.7% and 34.7% over 41 architectures. Among the 20 generators listed here, our method outperform the baselines in 17 out of 20 cases.

References

[1] Dosovitskiy Alexey. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929*, 2020. 4, 9, 10

[2] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 4, 9, 10

[3] Andrew Brock, Soham De, and Samuel L Smith. Characterizing signal propagation to close the performance gap in unnormalized resnets. *arXiv preprint arXiv:2101.08692*, 2021. 4, 9, 10

[4] MDN Chandrasiri and Priyanga Dilini Talagala. Cross-vit: Cross-attention vision transformer for image duplicate detection. In *2023 8th International Conference on Information Technology Research (ICITR)*, pages 1–6. IEEE, 2023. 4, 9, 10

[5] Zhengsu Chen, Lingxi Xie, Jianwei Niu, Xuefeng Liu, Longhui Wei, and Qi Tian. Visformer: The vision-friendly transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 589–598, 2021. 4, 9, 10

[6] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017. 4, 9, 10

[7] Mingyu Ding, Bin Xiao, Noel Codella, Ping Luo, Jingdong Wang, and Lu Yuan. Davit: Dual attention vision transformers. In *European conference on computer vision*, pages 74–92. Springer, 2022. 4, 9, 10

[8] Benjamin Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Hervé Jégou, and Matthijs Douze. Levit: a vision transformer in convnet's clothing for faster inference. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12259–12269, 2021. 4, 9, 10

[9] Kai Han, Yunhe Wang, Qi Tian, Jianyuan Guo, Chunjing Xu, and Chang Xu. Ghostnet: More features from cheap operations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1580–1589, 2020. 4, 9, 10

[10] Ali Hatamizadeh, Hongxu Yin, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. Global context vision transformers. In *International Conference on Machine Learning*, pages 12633–12646. PMLR, 2023. 4, 9, 10

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 3, 4, 9, 10

[12] Andrew G Howard. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 4, 9, 10

[13] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10

[14] Forrest N Iandola. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016. 4, 9, 10

[15] Donghyun Kim, Byeongho Heo, and Dongyoon Han. Densenets reloaded: Paradigm shift beyond resnets and vits. *arXiv preprint arXiv:2403.19588*, 2024. 4, 9, 10

[16] Brett Koonce and Brett Koonce. Efficientnet. *Convolutional neural networks with swift for Tensorflow: image recognition and dataset categorization*, pages 109–123, 2021. 4, 9, 10

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 4, 9, 10

[18] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4804–4814, 2022. 4, 9, 10

[19] Chenxi Liu, Barret Zoph, Maxim Neumann, Jonathon Shlens, Wei Hua, Li-Jia Li, Li Fei-Fei, Alan Yuille, Jonathan Huang, and Kevin Murphy. Progressive neural architecture search. In *Proceedings of the European conference on computer vision (ECCV)*, pages 19–34, 2018. 4, 9, 10

[20] Xinyu Liu, Houwen Peng, Ningxin Zheng, Yuqing Yang, Han Hu, and Yixuan Yuan. Efficientvit: Memory efficient vision transformer with cascaded group attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14420–14430, 2023. 4, 9, 10

[21] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 4, 9, 10

[22] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 4, 9, 10

[23] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pages 116–131, 2018. 4, 9, 10

[24] Muhammad Maaz, Abdelrahman Shaker, Hisham Cholakkal, Salman Khan, Syed Waqas Zamir, Rao Muhammad Anwer, and Fahad Shahbaz Khan. Edgenext: efficiently amalgamated cnn-transformer architecture for mobile vision applications. In *European conference on computer vision*, pages 3–20. Springer, 2022. 4, 9, 10

[25] TorchVision maintainers and contributors. Torchvision: Pytorch's computer vision library. https://github.com/pytorch/vision, 2016. 1, 4

[26] Sachin Mehta and Mohammad Rastegari. Mobilevit: lightweight, general-purpose, and mobile-friendly vision transformer. *arXiv preprint arXiv:2110.02178*, 2021. 4, 9, 10

[27] Muhammad Muzammal Naseer, Salman H Khan, Muhammad Haris Khan, Fahad Shahbaz Khan, and Fatih Porikli.

Cross-domain transferability of adversarial perturbations. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 9, 10

[28] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 1

[29] Omid Poursaeed, Isay Katsman, Bicheng Gao, and Serge Belongie. Generative adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4422–4431, 2018. 9, 10

[30] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10428–10436, 2020. 4, 9, 10

[31] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 1

[32] Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, et al. Hiera: A hierarchical vision transformer without the bells-and-whistles. In *International Conference on Machine Learning*, pages 29441–29454. PMLR, 2023. 4, 9, 10

[33] Mathieu Salzmann et al. Learning transferable adversarial perturbations. *Advances in Neural Information Processing Systems*, 34:13950–13962, 2021. 1, 4, 9, 10

[34] Karen Simonyan. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 1

[35] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1, 2, 3, 4

[36] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V Le. Mnasnet: Platform-aware neural architecture search for mobile. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2820–2828, 2019. 4, 9, 10

[37] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, 34:24261–24272, 2021. 4, 9, 10

[38] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. 4, 9, 10

[39] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 32–42, 2021. 4, 9, 10

[40] Asher Trockman and J Zico Kolter. Patches are all you need? *arXiv preprint arXiv:2201.09792*, 2022. 4, 9, 10

[41] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. In *European conference on computer vision*, pages 459–479. Springer, 2022. 4, 9, 10

[42] Chien-Yao Wang, Hong-Yuan Mark Liao, Yueh-Hua Wu, Ping-Yang Chen, Jun-Wei Hsieh, and I-Hau Yeh. Cspnet: A new backbone that can enhance learning capability of cnn. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 390–391, 2020. 4, 9, 10

[43] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020. 4, 9, 10

[44] Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019. 1, 4

[45] Weijian Xu, Yifan Xu, Tyler Chang, and Zhuowen Tu. Co-scale conv-attentional image transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9981–9990, 2021. 4, 9, 10

[46] Jianwei Yang, Chunyuan Li, Xiyang Dai, and Jianfeng Gao. Focal modulation networks. *Advances in Neural Information Processing Systems*, 35:4203–4217, 2022. 4, 9, 10

[47] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2403–2412, 2018. 4

[48] Weihao Yu, Chenyang Si, Pan Zhou, Mi Luo, Yichen Zhou, Jiashi Feng, Shuicheng Yan, and Xinchao Wang. Metaformer baselines for vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 4, 9, 10

[49] Sergey Zagoruyko. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. 4, 9, 10

[50] Qilong Zhang, Xiaodan Li, Yuefeng Chen, Jingkuan Song, Lianli Gao, Yuan He, and Hui Xue. Beyond imagenet attack: Towards crafting adversarial examples for black-box domains. *arXiv preprint arXiv:2201.11528*, 2022. 1, 9, 10