Vax-a-Net: Training-time Defence Against Adversarial Patch Attacks
Supplementary Material

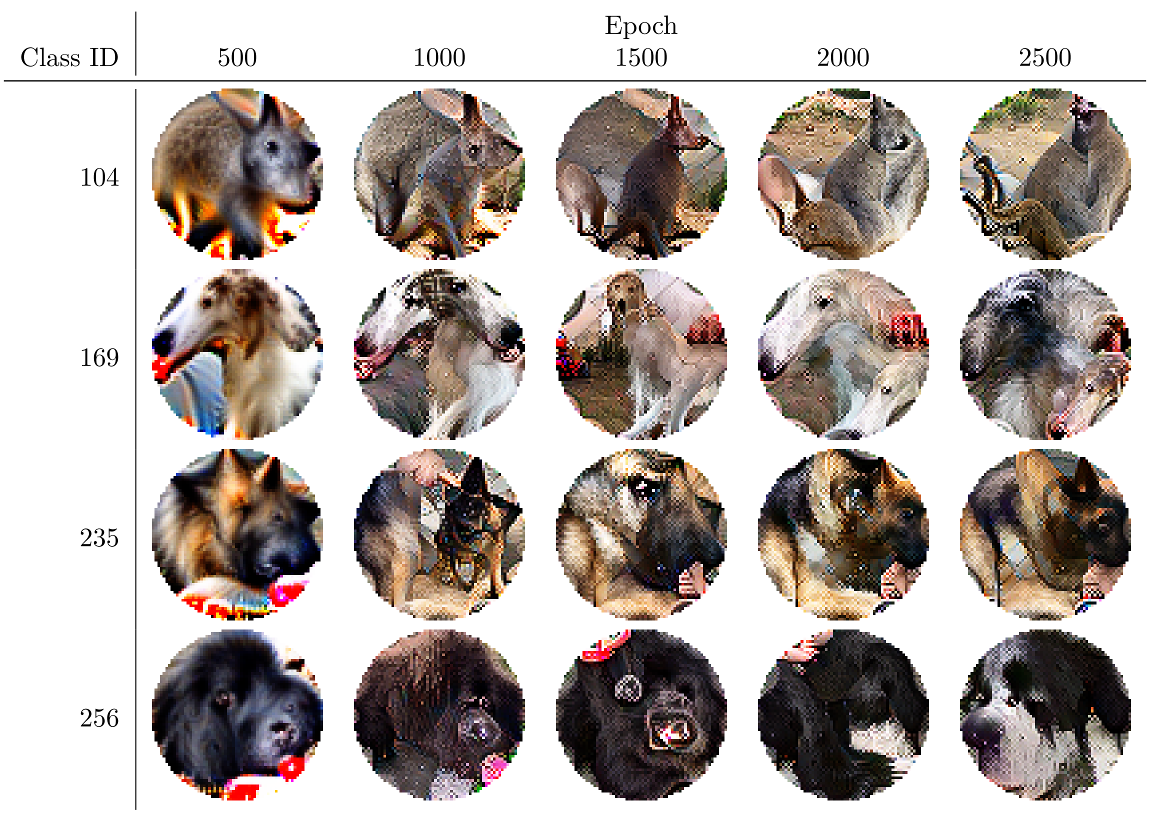
T. Gittings[1], S. Schneider[2] and J. Collomosse[1]

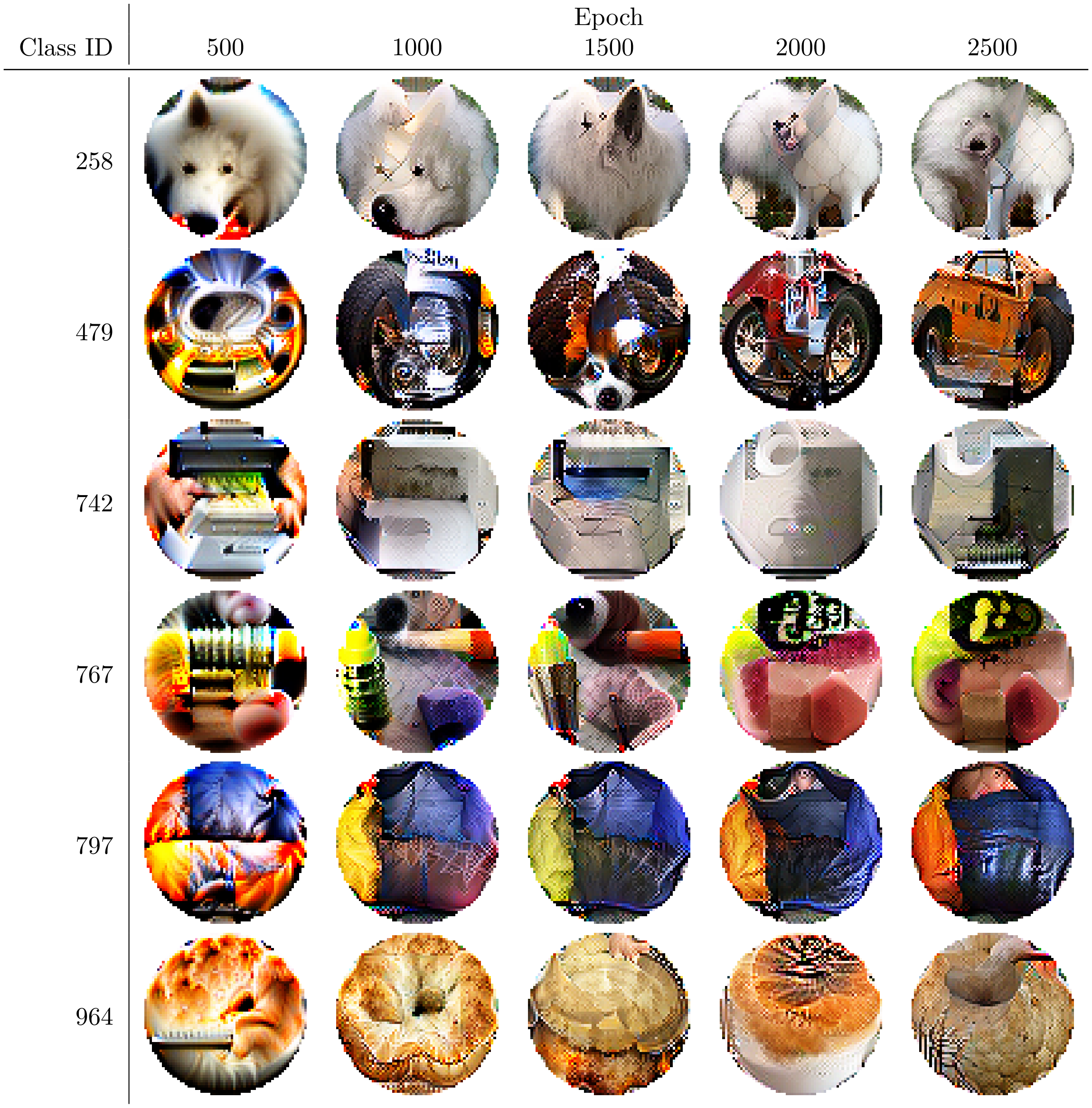
[1] Centre for Vision Speech and Signal Processing (CVSSP), University of Surrey, Guildford, UK.
[2] Surrey Centre for Cyber Security (SCCS), University of Surrey, Guildford, UK.
{t.gittings,s.schneider,j.collomosse}@surrey.ac.uk

1 Sampled Vax-a-Net Attack Patches

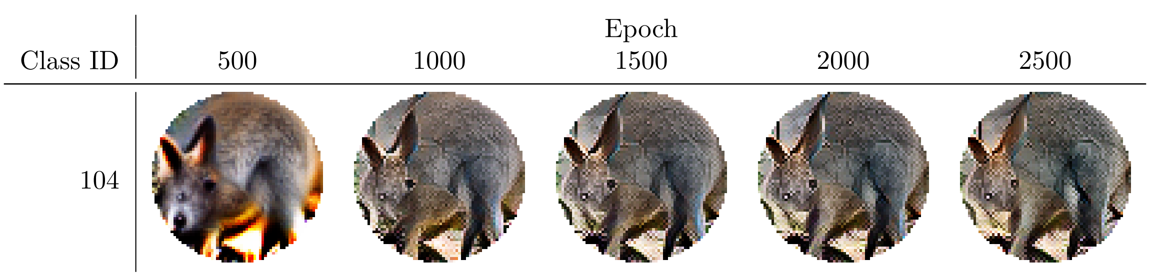
In this section we show representative patches sampled from our generator at training epochs 500-2500. We include patches for VGG-19, Inception-v3 and InceptionResNet-v2 networks, produced by both 10 and 50 class generators.

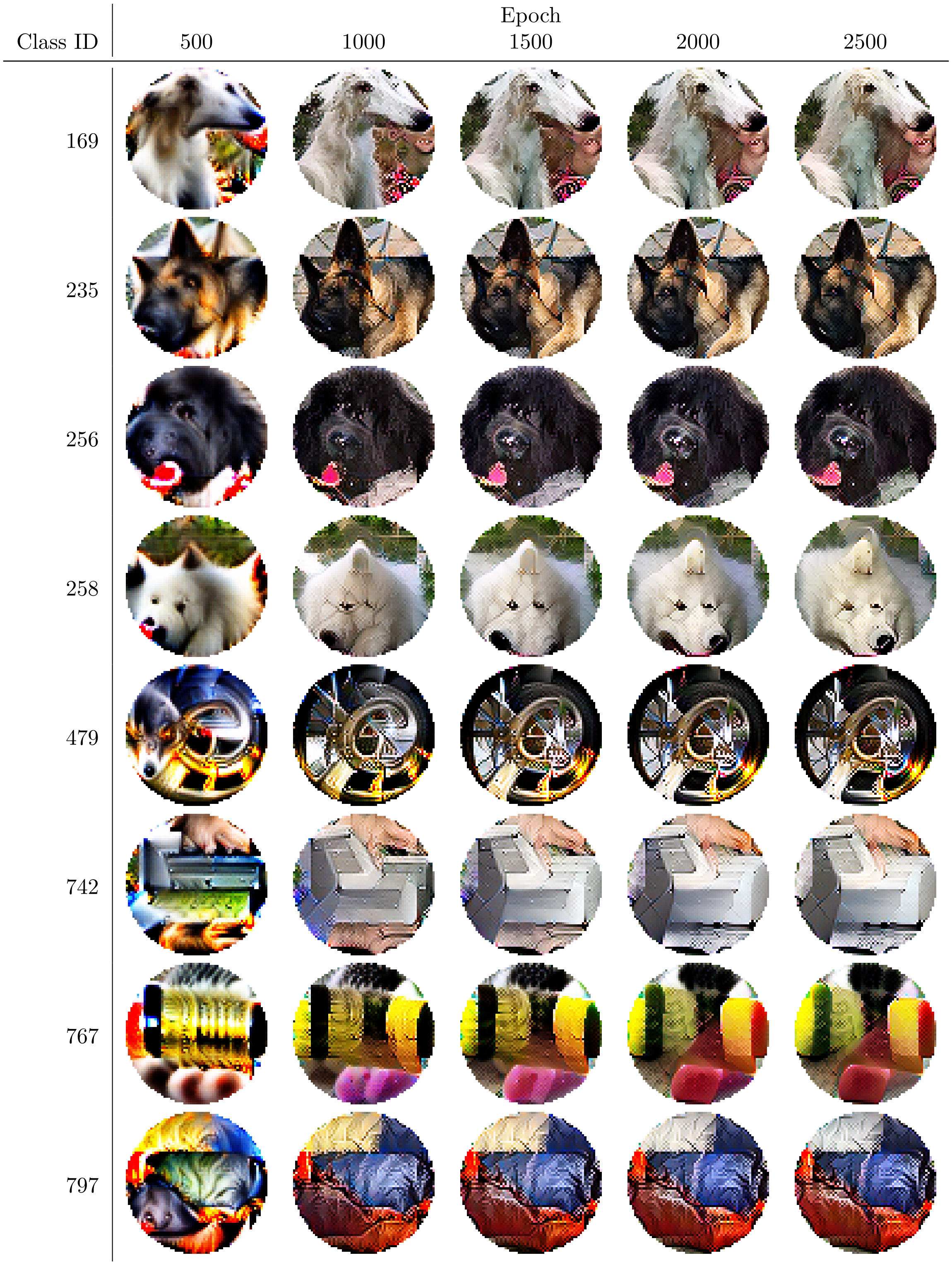
1.1 VGG-19 Model, 10 Class Generator

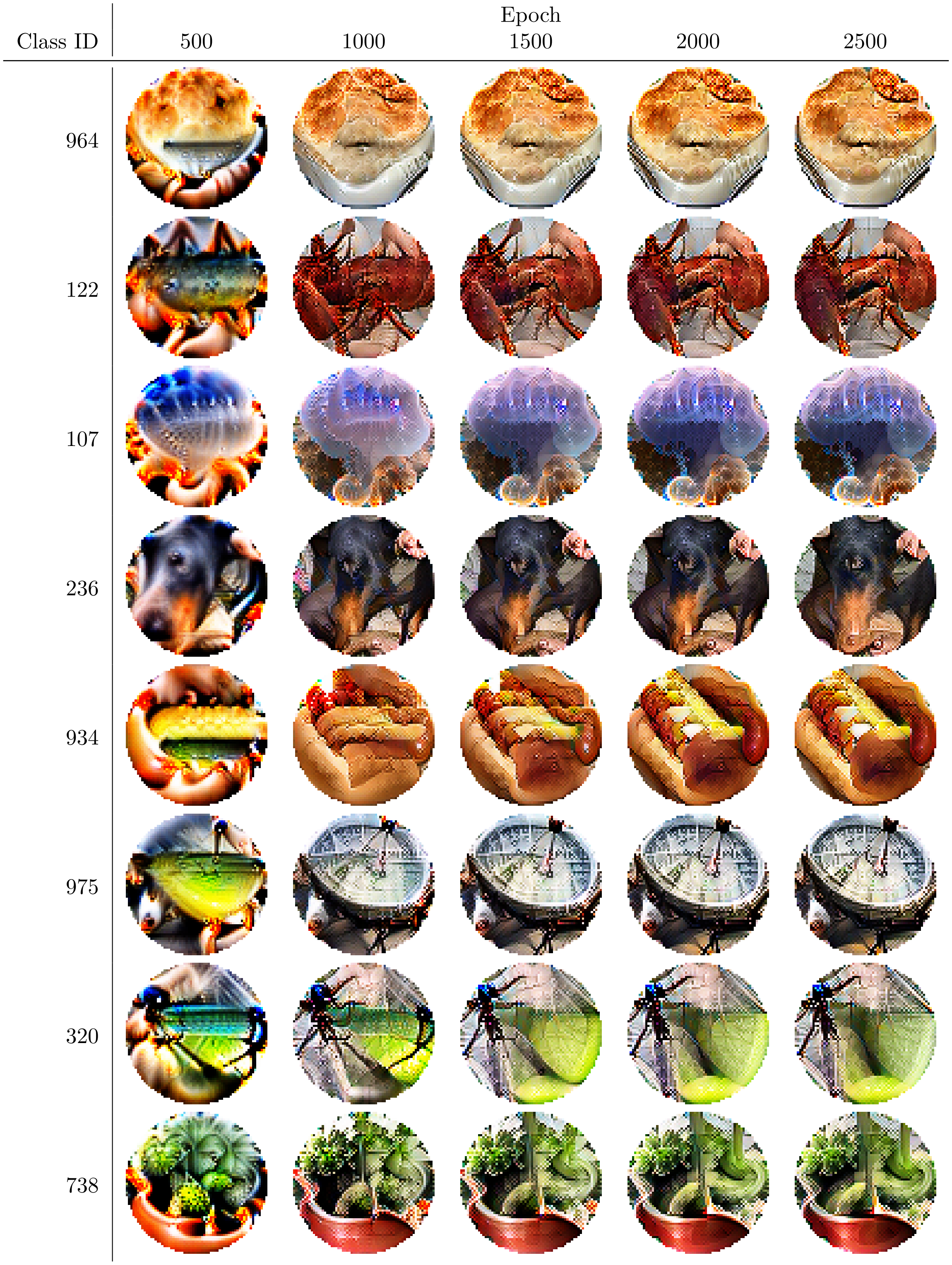


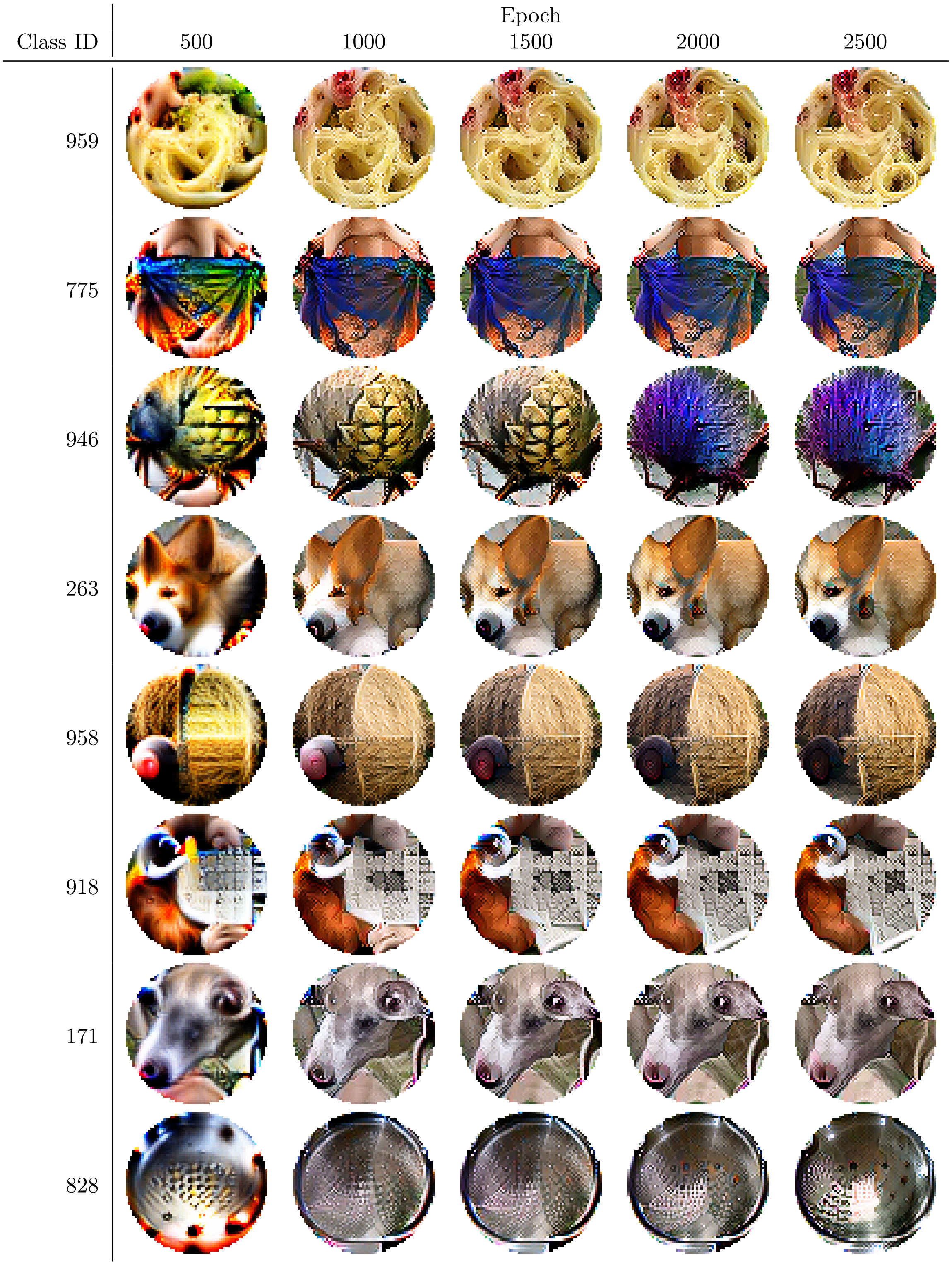


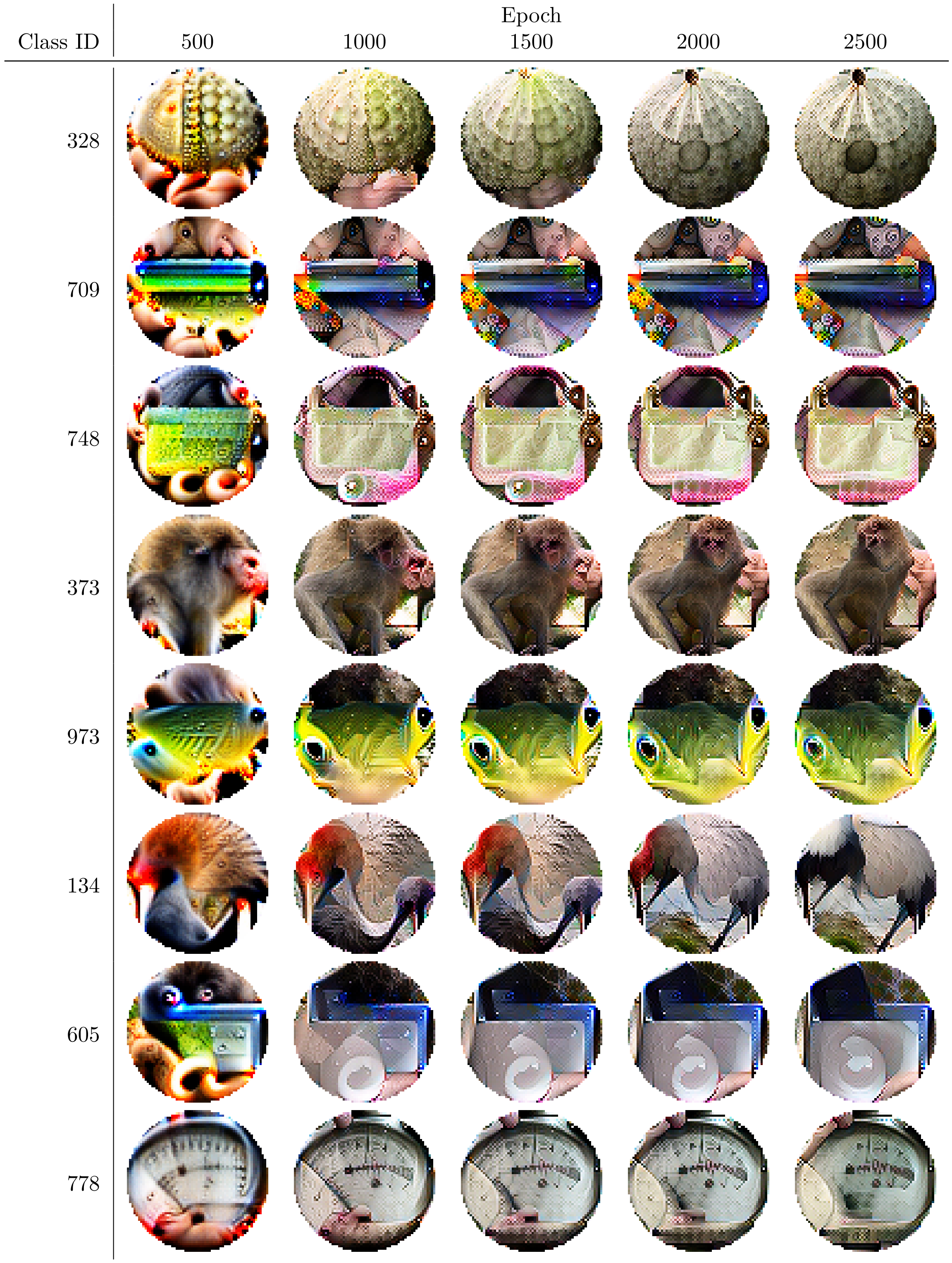
1.2 VGG-19 Model, 50 Class Generator

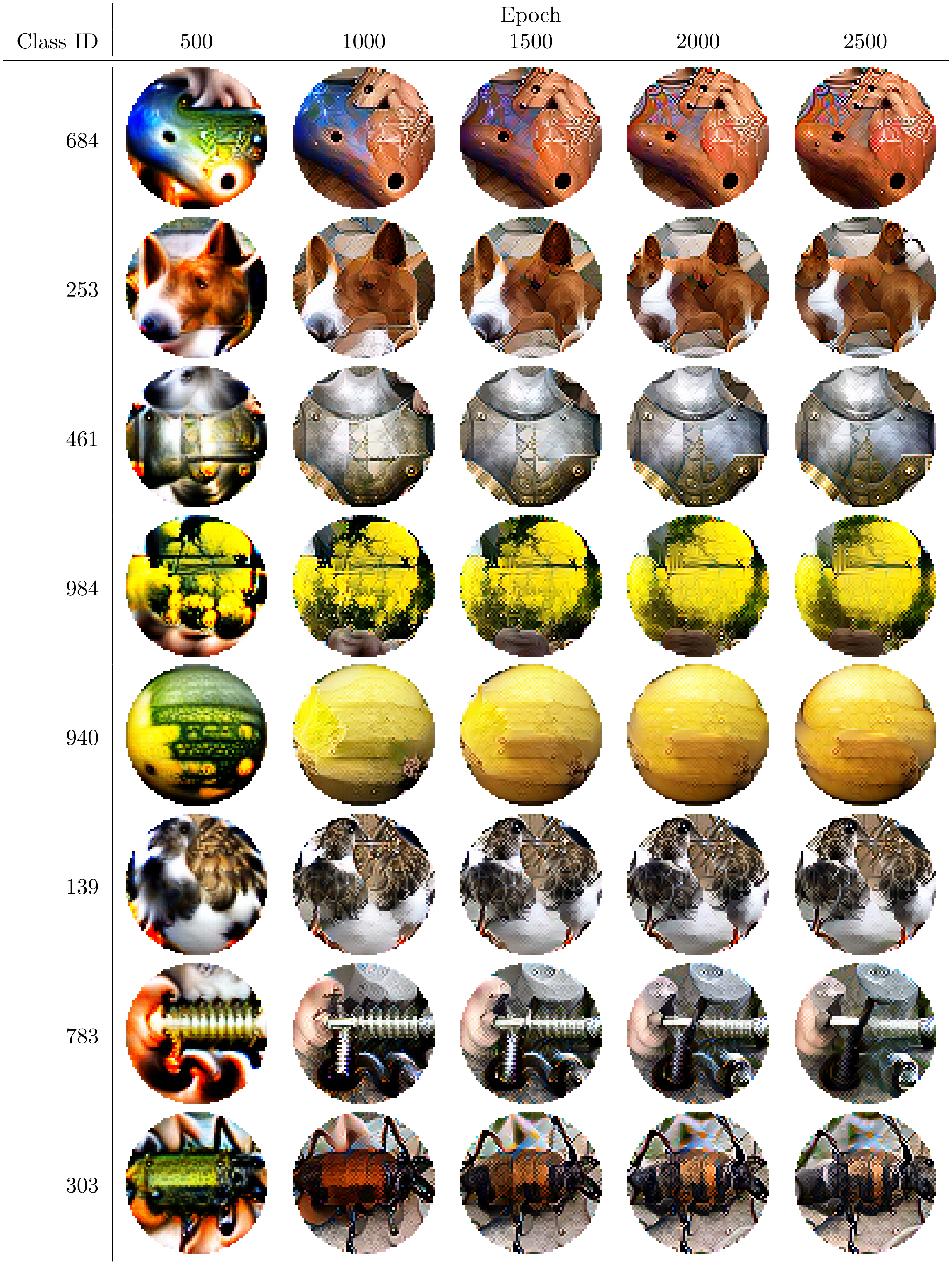



Class ID
Epoch
500
1000
1500
2000
2500
169
235
256
258
479
742
767
797

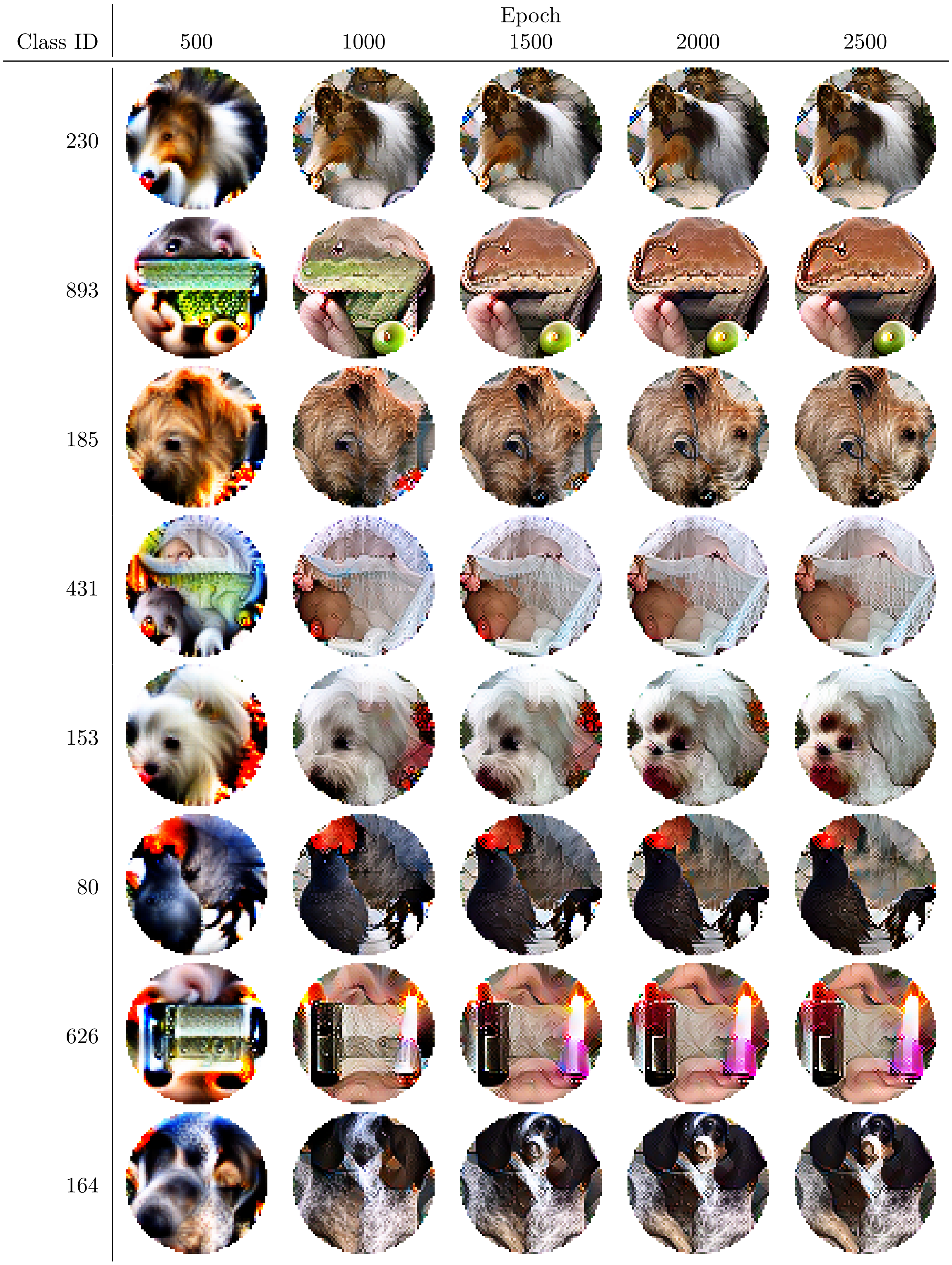

Epoch
Class ID
500
1000
1500
2000
2500
964
122
107
236
934
975
320
738


Class ID
Epoch
500
1000
1500
2000
2500
959
775
946
263
958
918
171
828


Epoch
Class ID
500
1000
1500
2000
2500
328
709
748
373
973
134
605
778

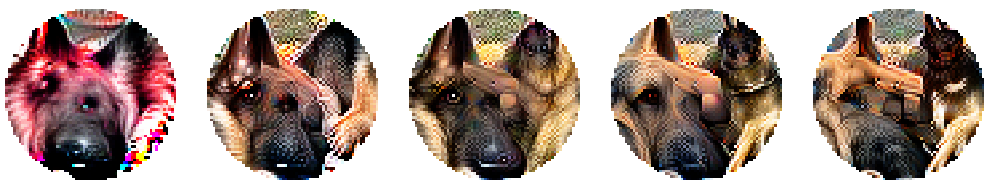
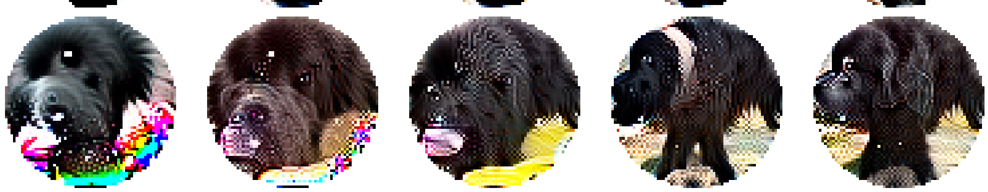


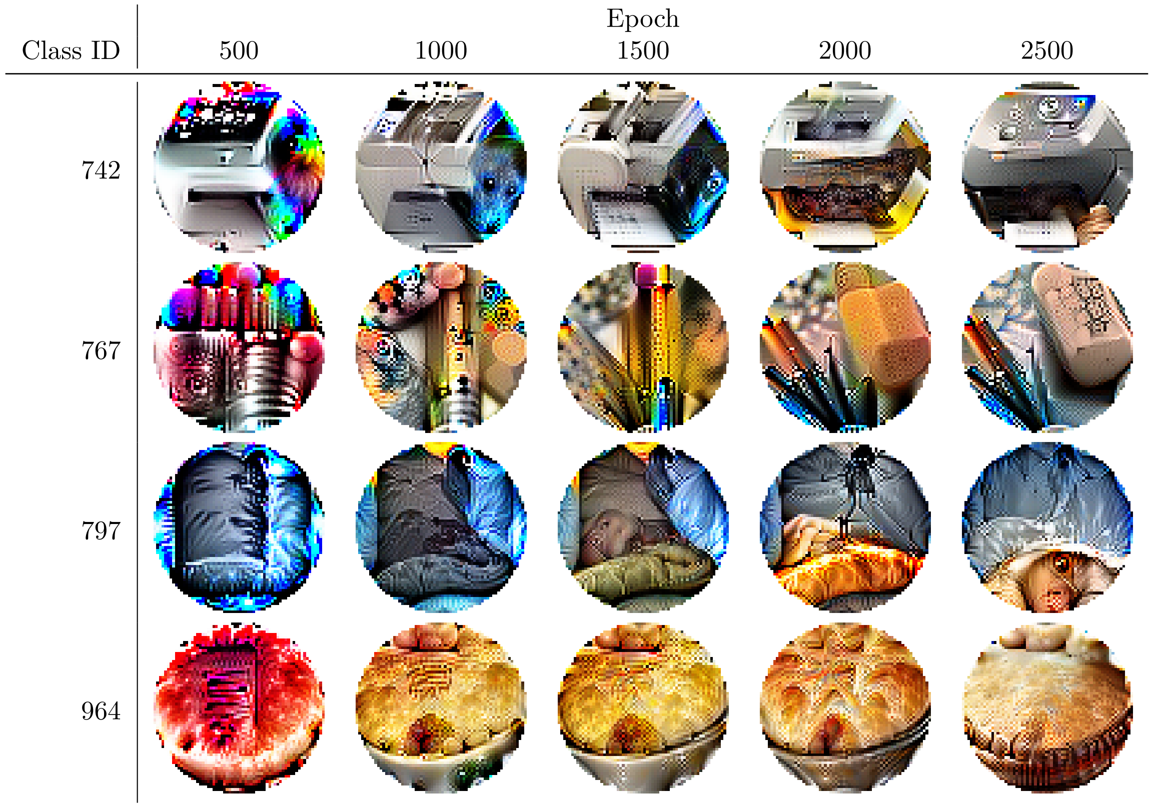
Class ID
Epoch
500
1000
1500
2000
2500
684
253
461
984
940
139
783
303


Epoch
Class ID
500
1000
1500
2000
2500
230
893
185
431
153
80
626
164

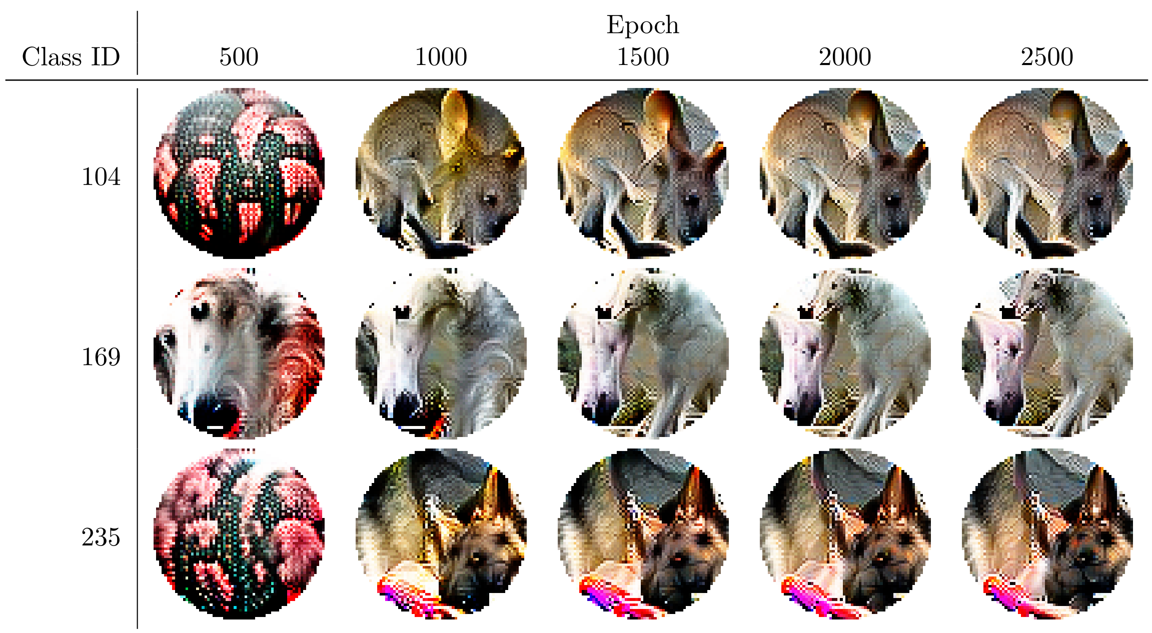
Class ID	Epoch 500	1000	1500	2000	2500
115					

1.3 Inception-v3 Model, 10 Class Generator

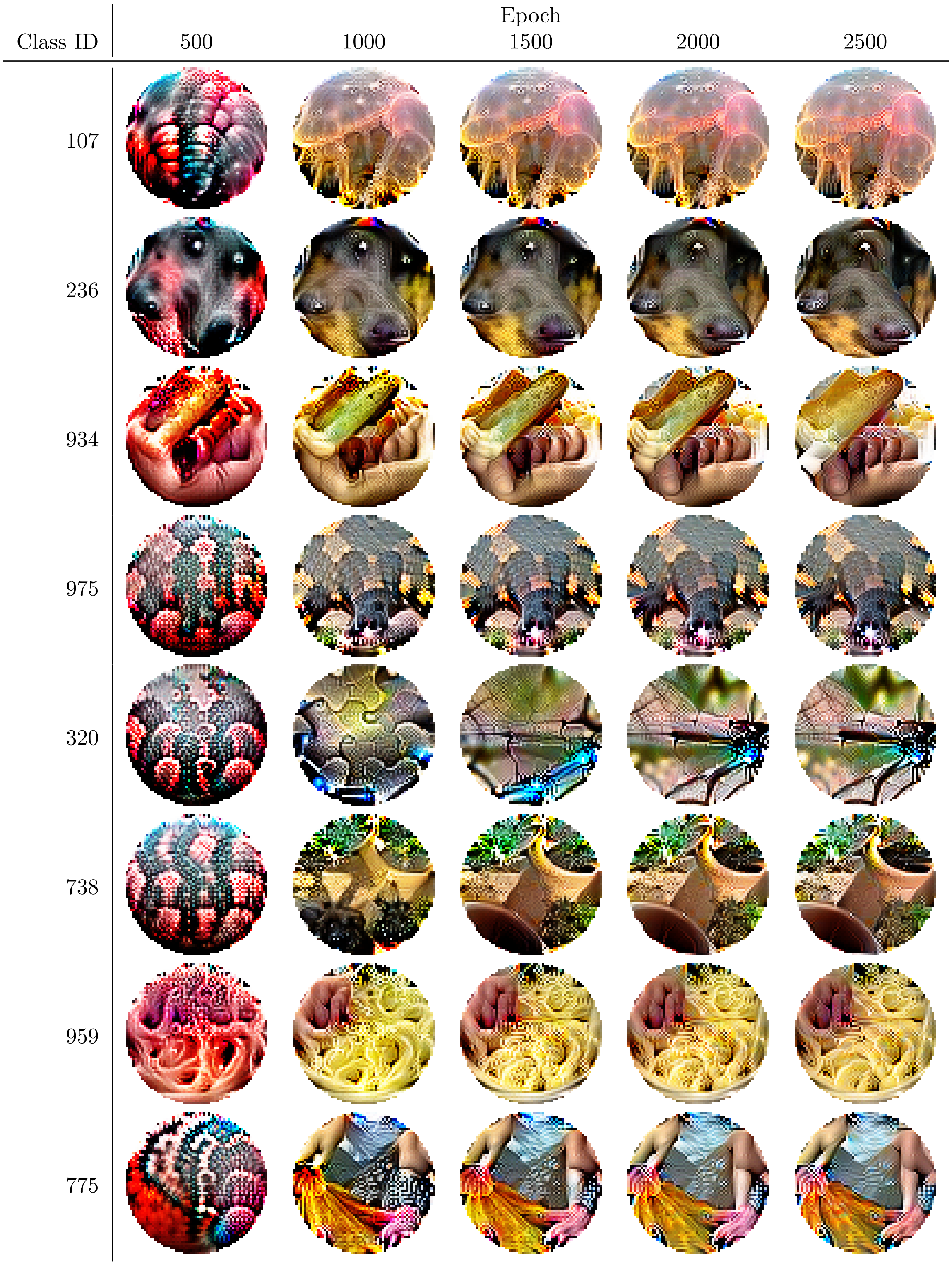
Class ID	Epoch 500	1000	1500	2000	2500
104					
169					
235					
256					
258					
479					

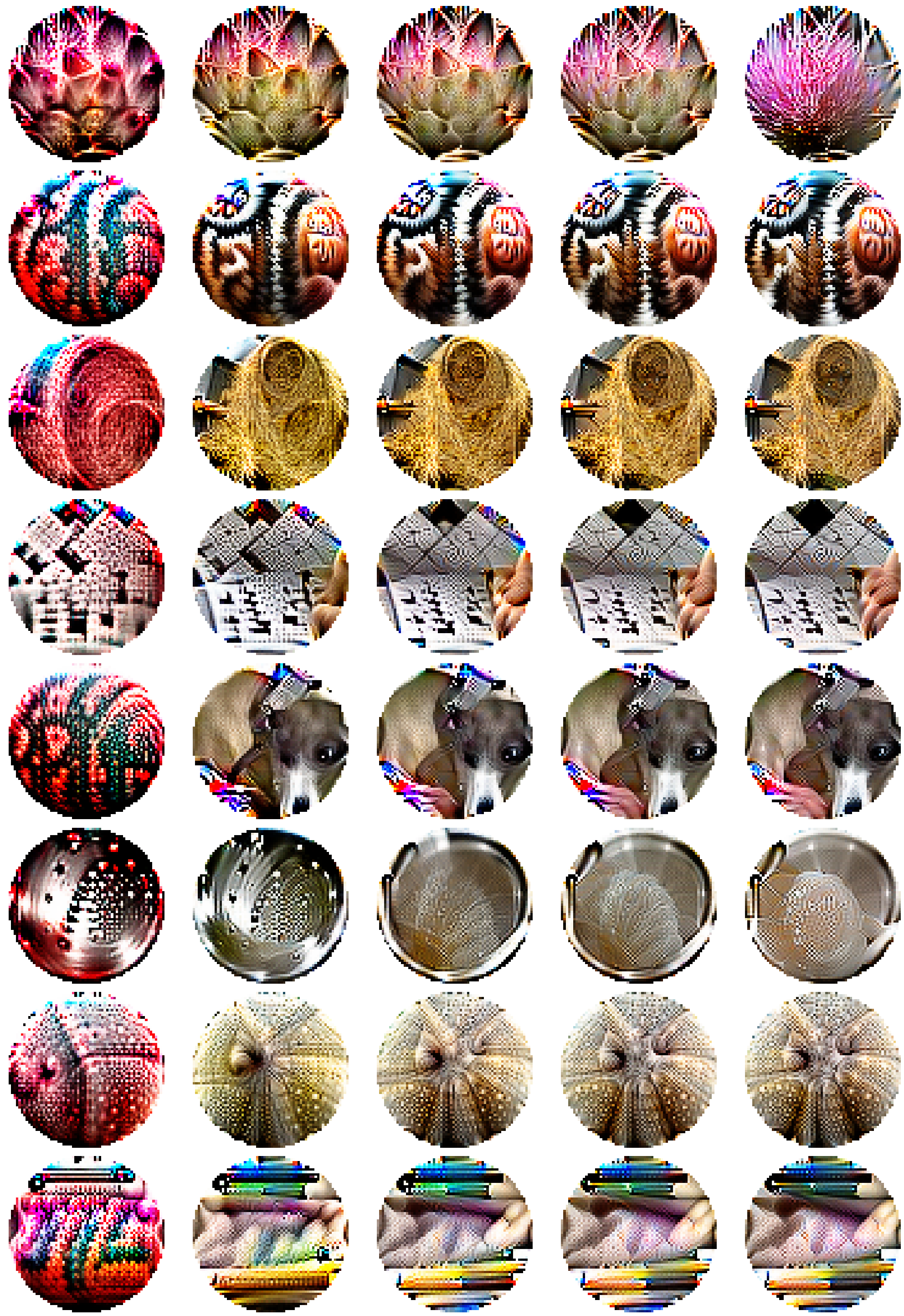


1.4 Inception-v3 Model, 50 Class Generator

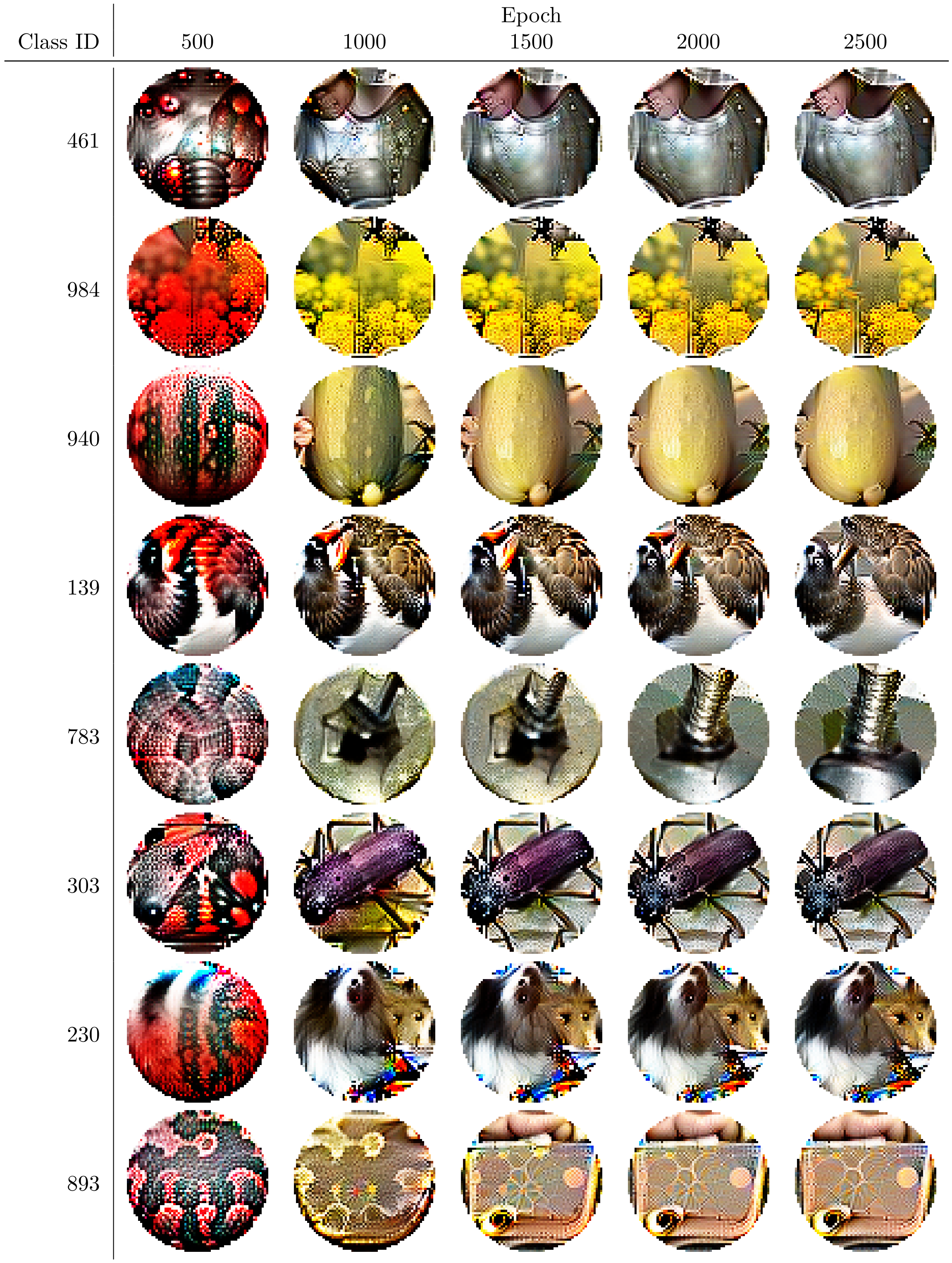


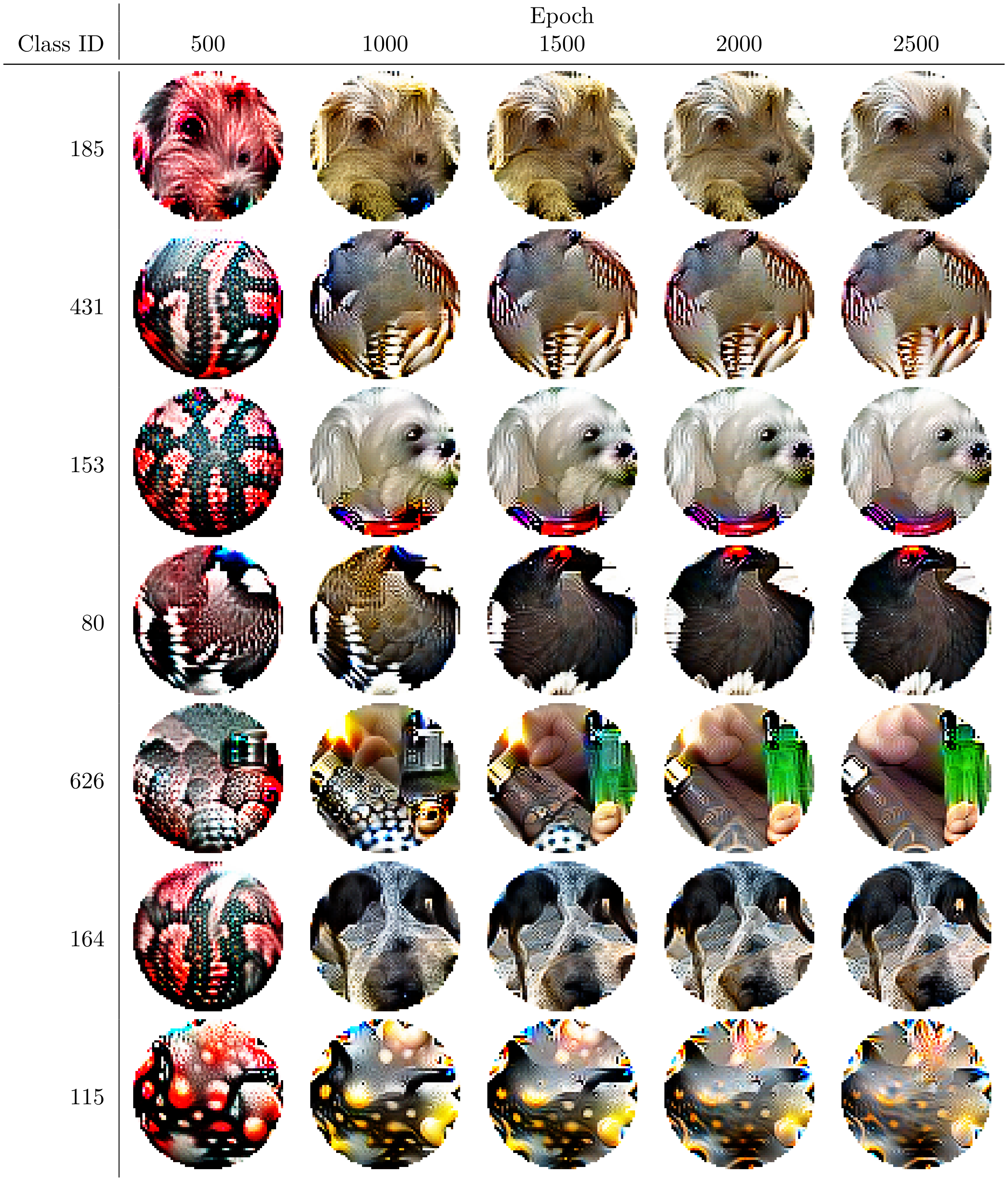

Class ID
Epoch
500
1000
1500
2000
2500
256
258
479
742
767
797
964
122


Epoch
Class ID
500
1000
1500
2000
2500
107
236
934
975
320
738
959
775


Class ID
Epoch
500
1000
1500
2000
2500
946
263
958
918
171
828
328
709

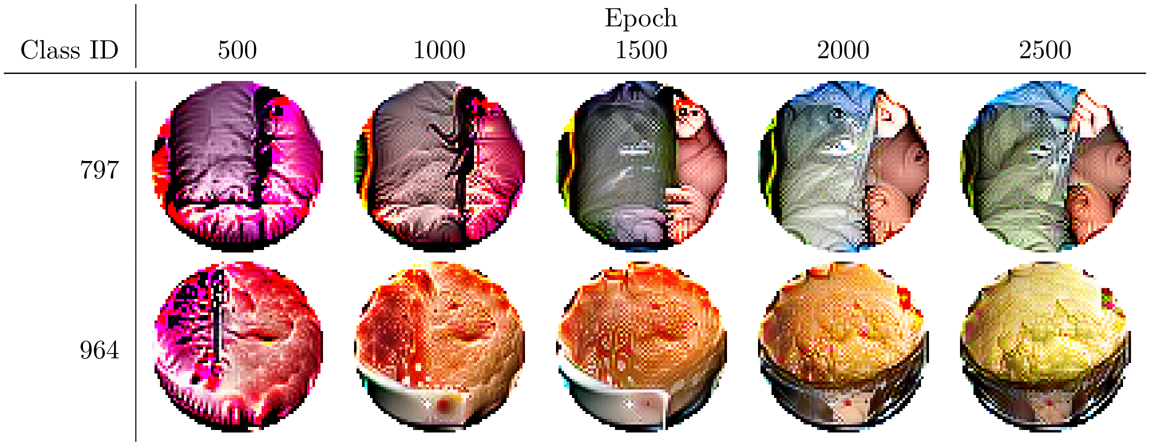

Epoch
Class ID
500
1000
1500
2000
2500
748
373
973
134
605
778
684
253


Class ID
Epoch
500
1000
1500
2000
2500
461
984
940
139
783
303
230
893

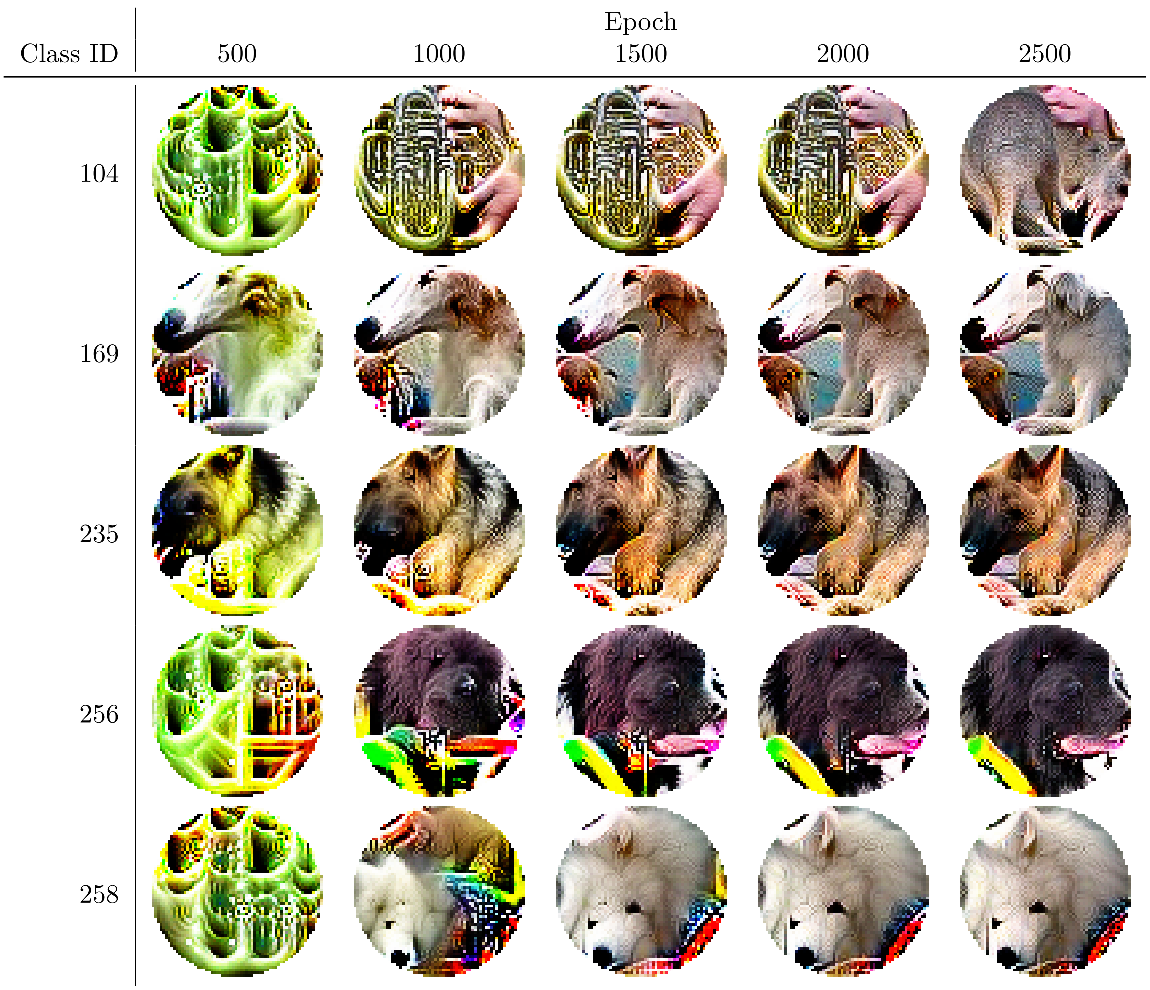


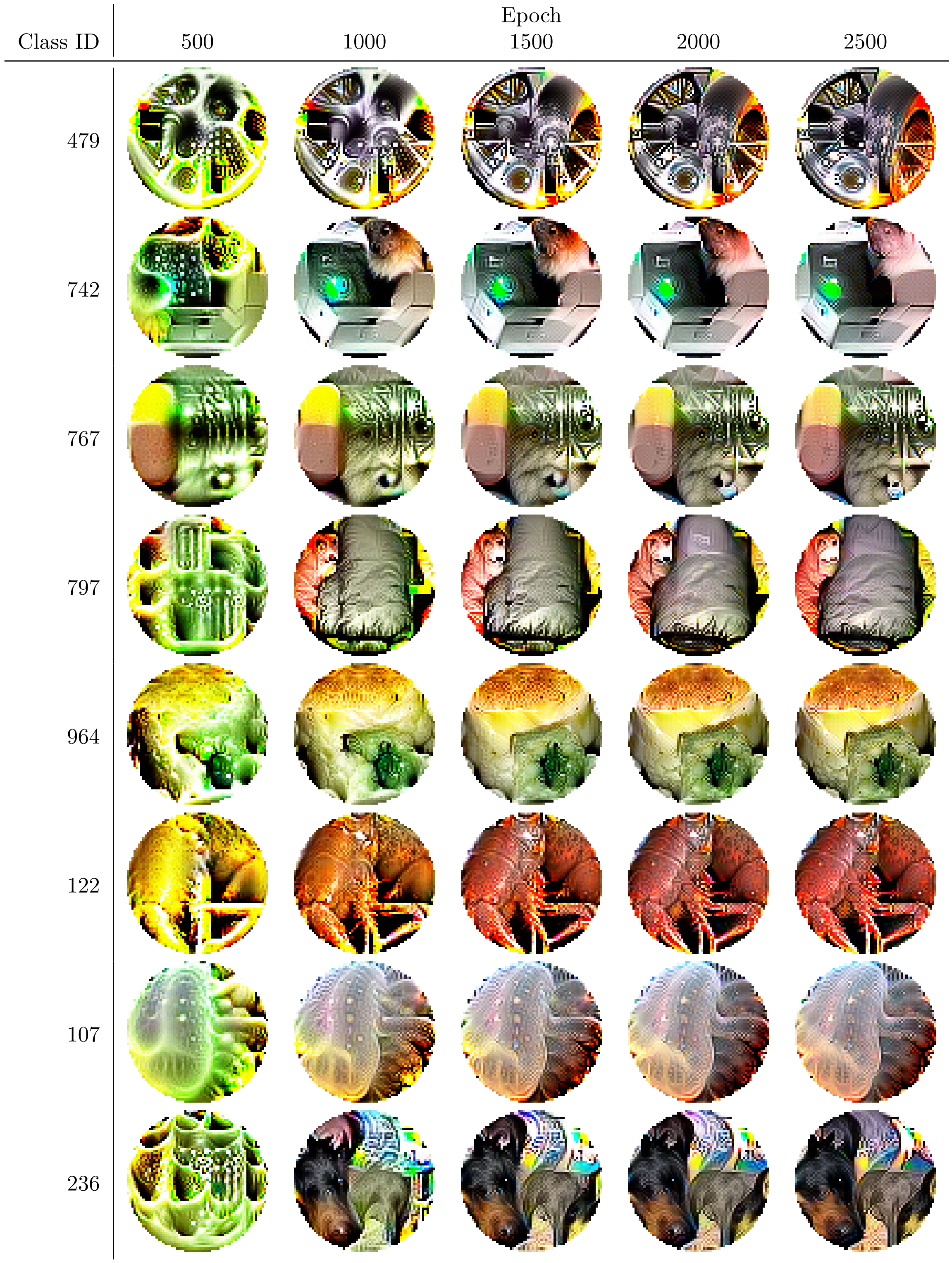
1.5 InceptionResNet-v2 Model, 10 Class Generator

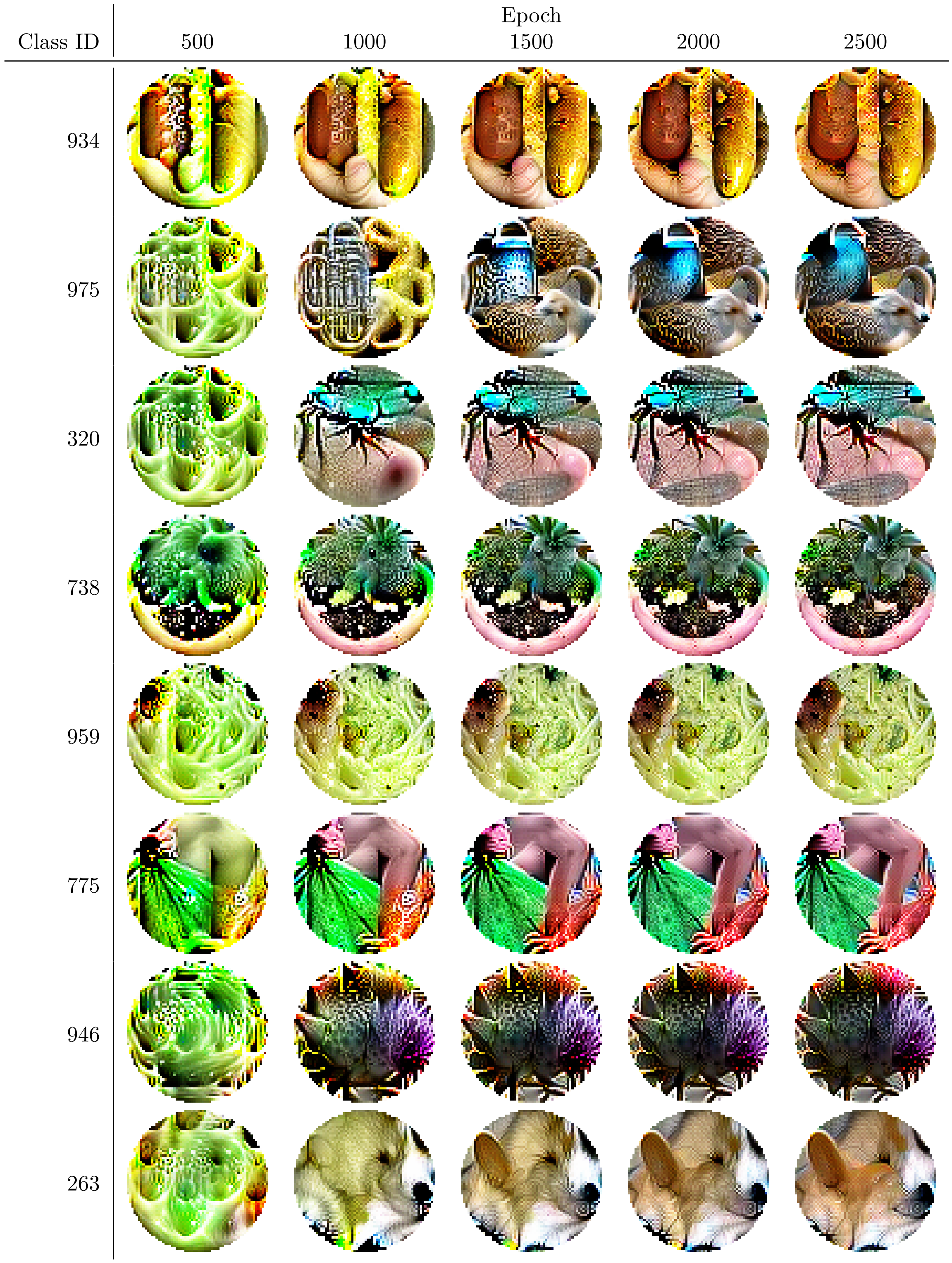

Epoch
Class ID
500
1000
1500
2000
2500
104
169
235
256
258
479
742
767

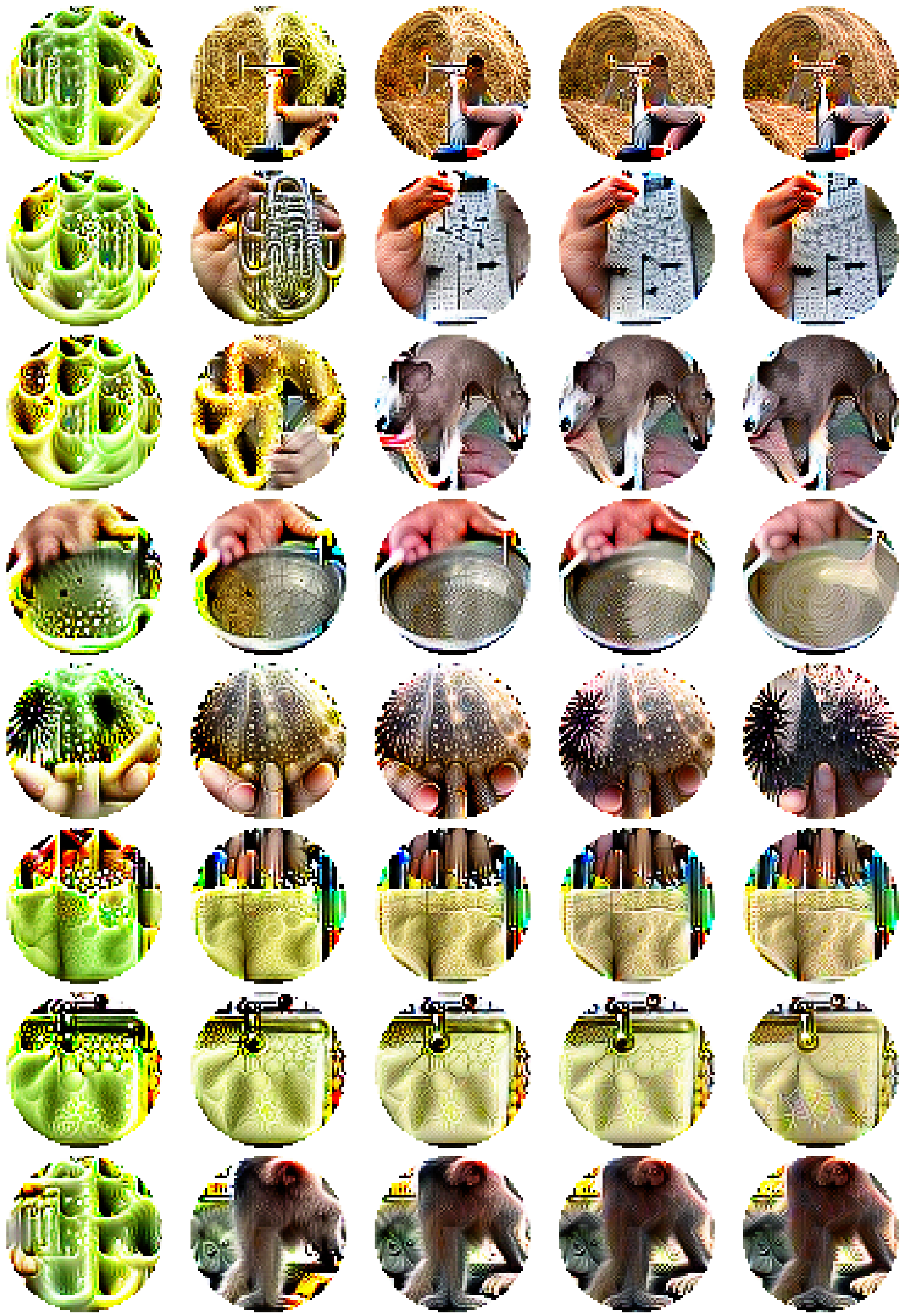


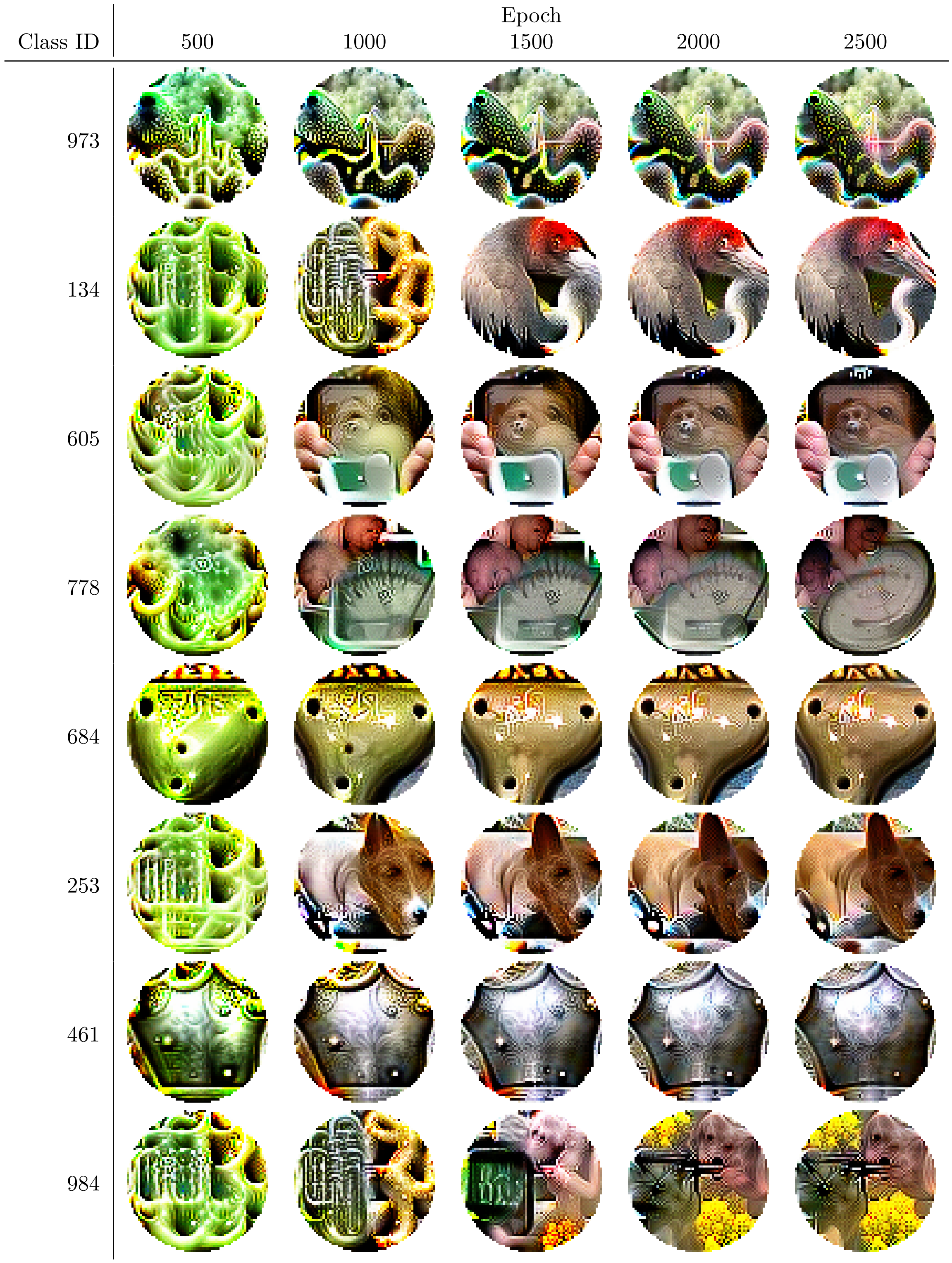
1.6 InceptionResNet-v2 Model, 50 Class Generator

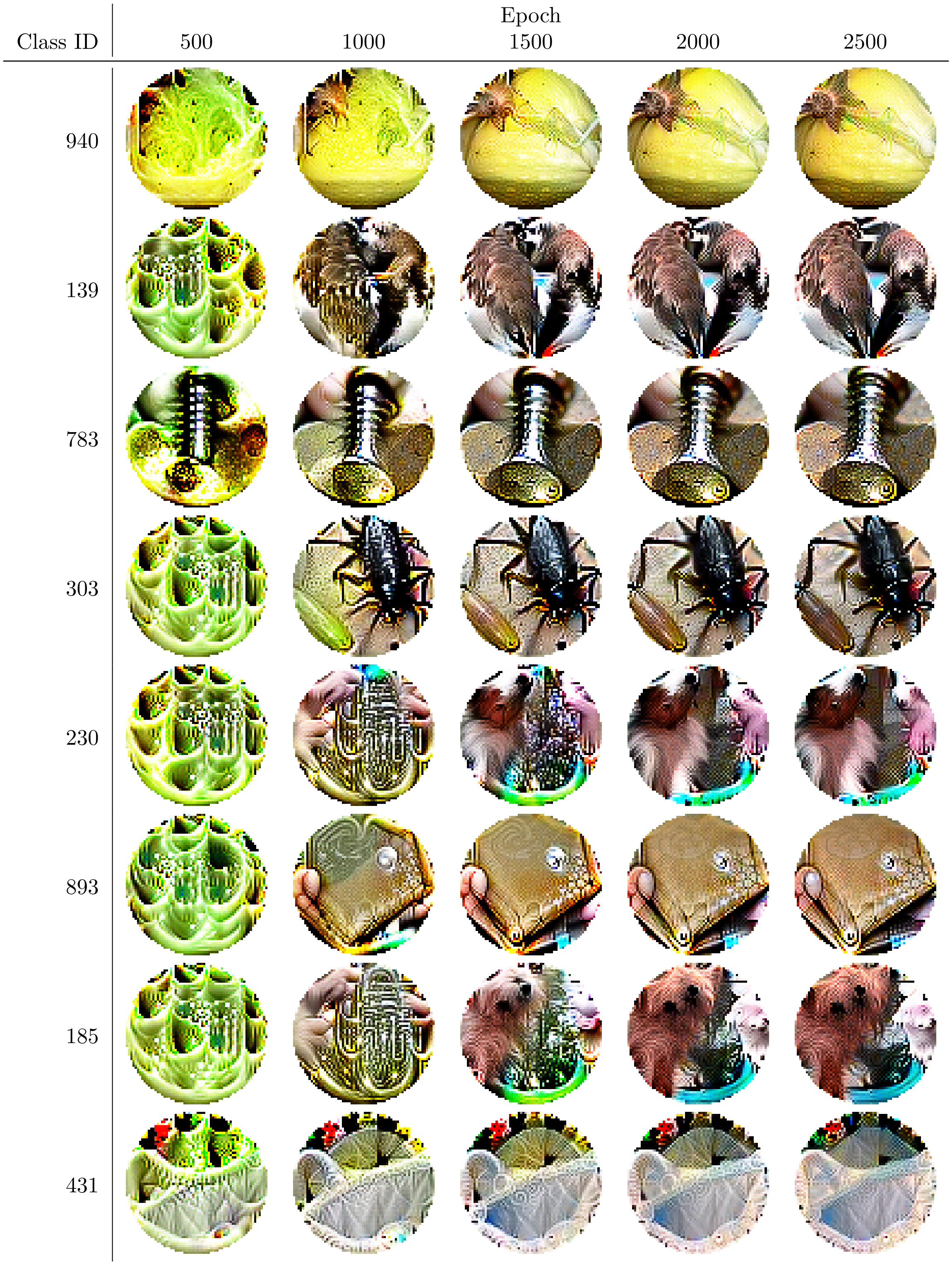


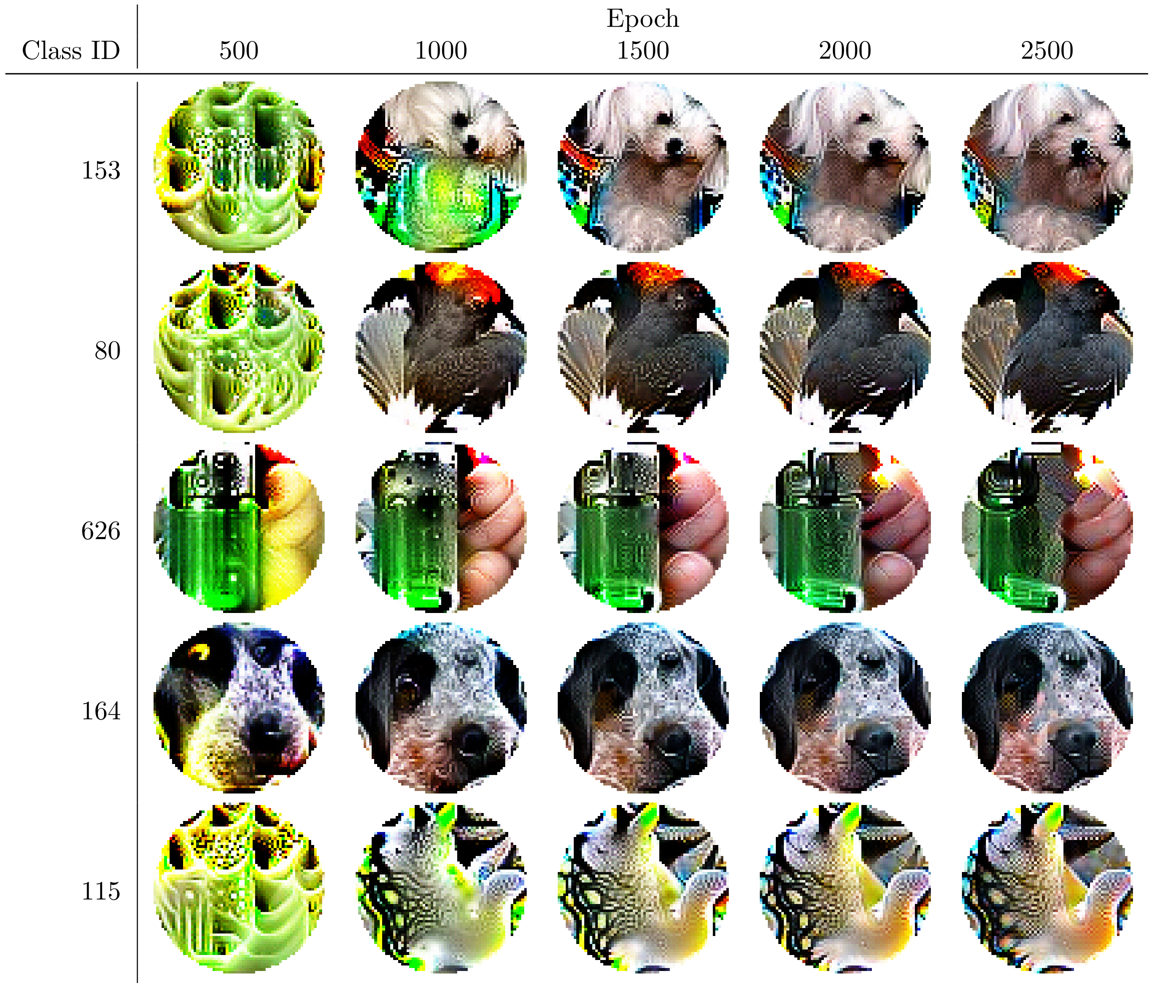

Class ID
Epoch
500
1000
1500
2000
2500
479
742
767
797
964
122
107
236


Epoch
Class ID
500
1000
1500
2000
2500
934
975
320
738
959
775
946
263


Epoch
Class ID
500
1000
1500
2000
2500
958
918
171
828
328
709
748
373


Epoch
Class ID
500
1000
1500
2000
2500
973
134
605
778
684
253
461
984

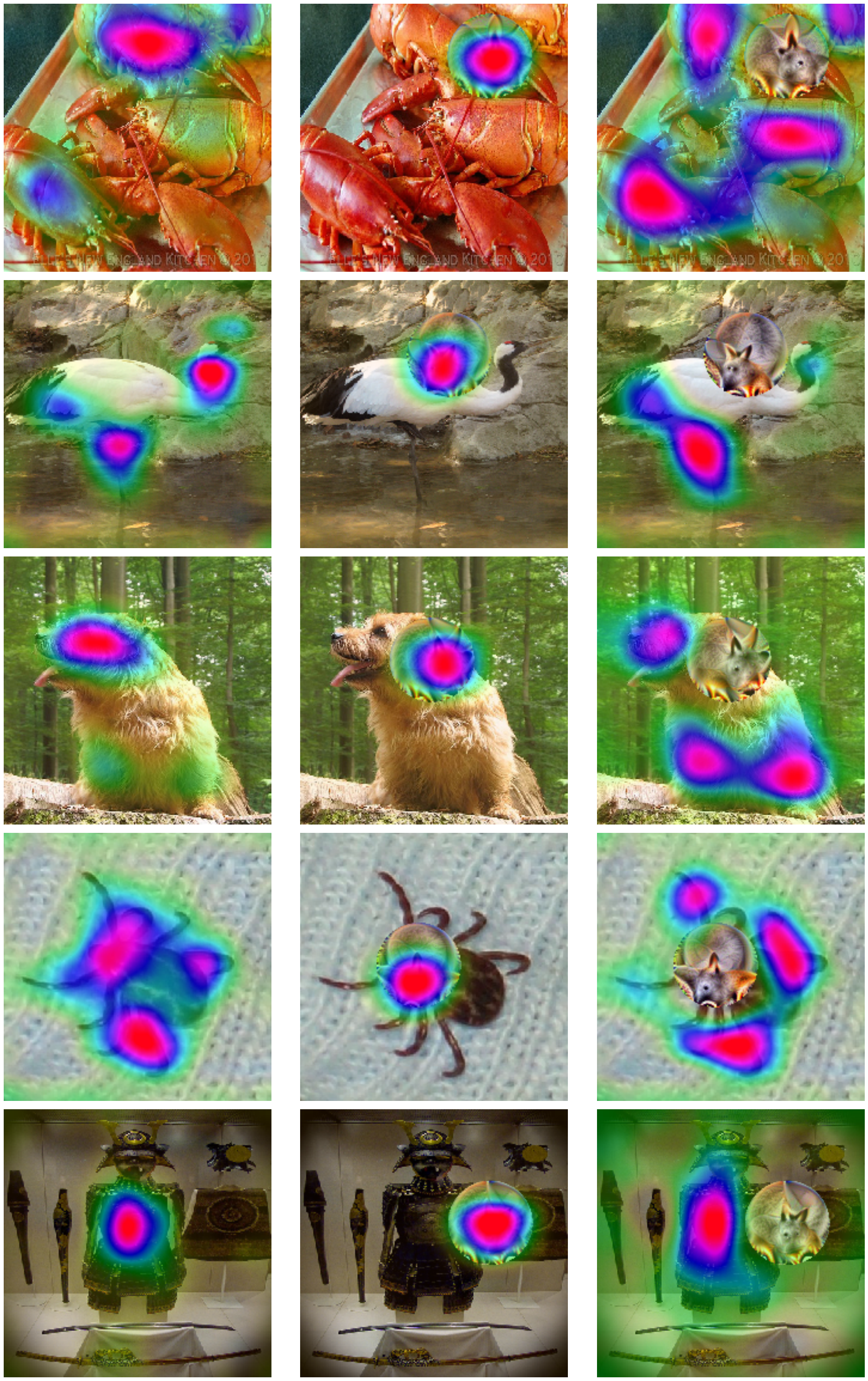

Class ID
Epoch
500
1000
1500
2000
2500
940
139
783
303
230
893
185
431



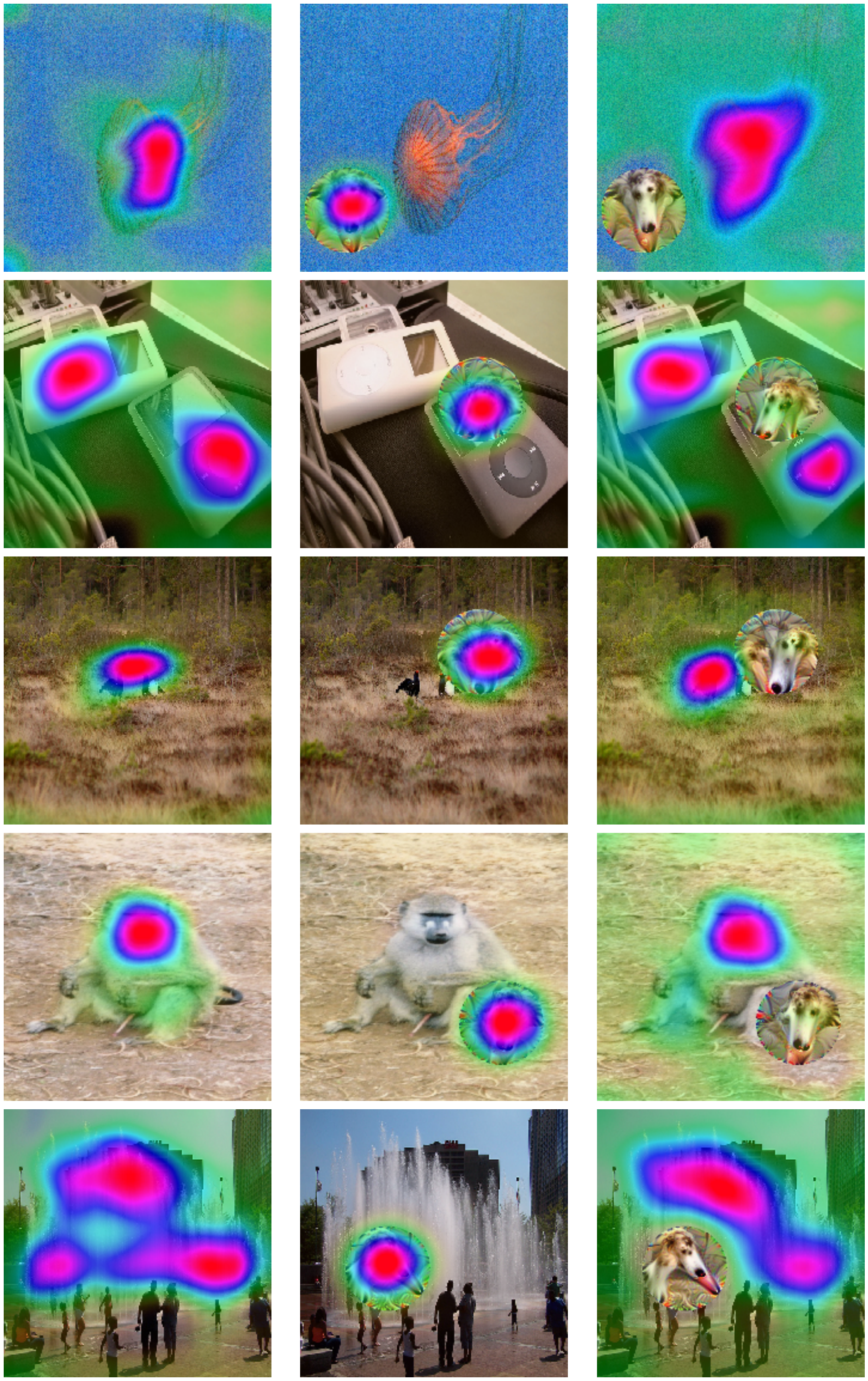
2 Grad-CAM Visualisations

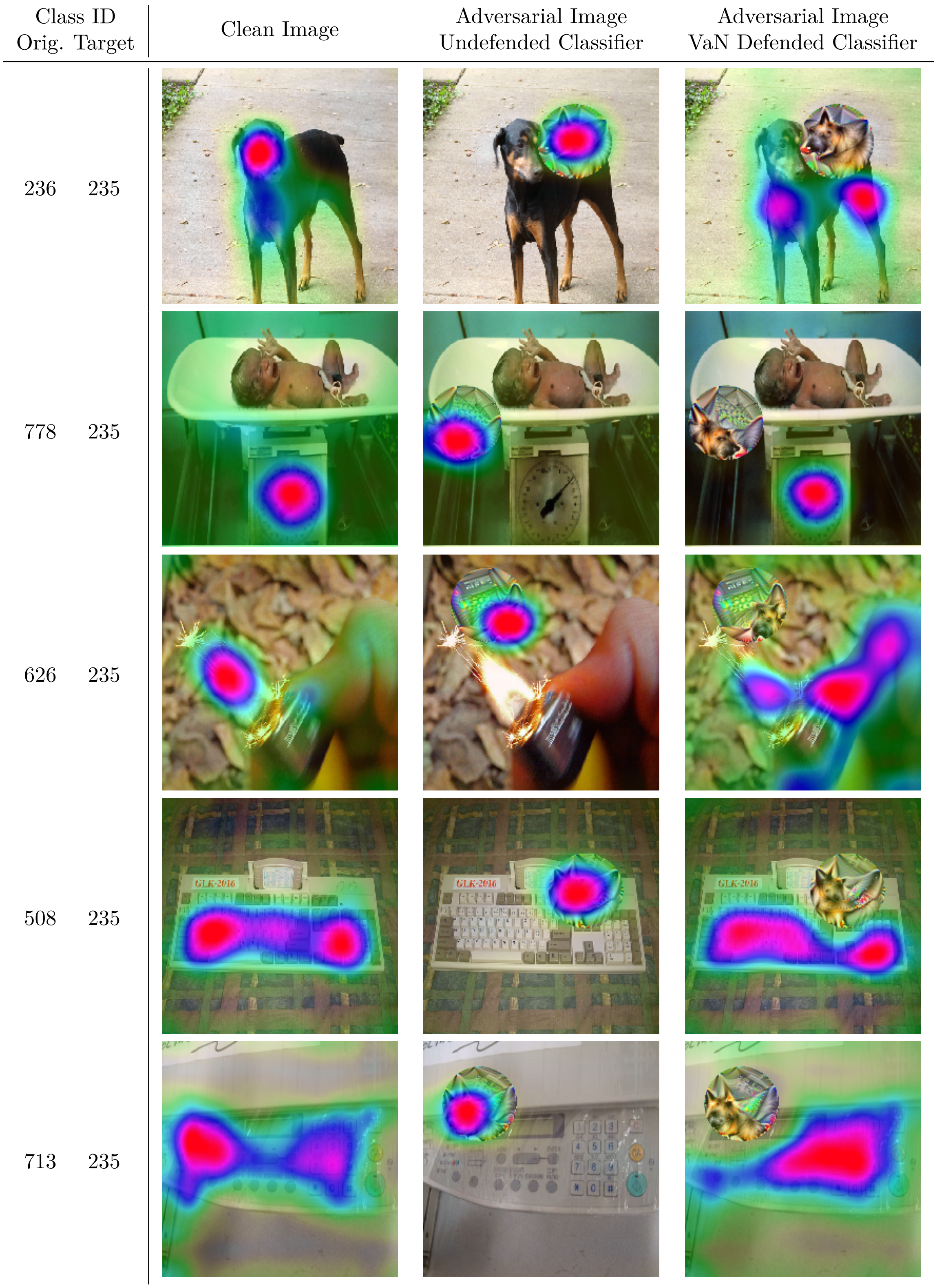
In this section we show visualisations of the network's attention via Grad-CAM (as explained in Section 4.3 of the main text). The Undefended Classifier is a VGG-19 network, pretrained on ImageNet, and the Defended Classifier is the same network defended by Vax-a-Net (VaN) with a 10 class generator. We show examples of 10 classes of patch (created with the Brown *et al.* method) applied to 50 different images. In all cases the clean image is classified correctly by both the original network and our network. When the patch is applied the original network classifies the image as the target class, but our defended network continues to classify it correctly.

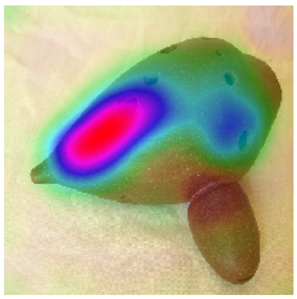
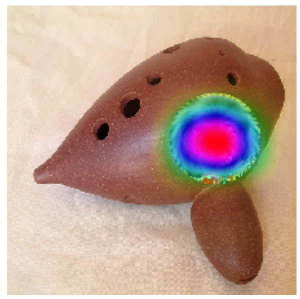
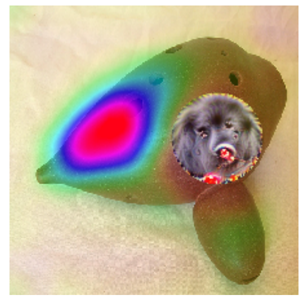
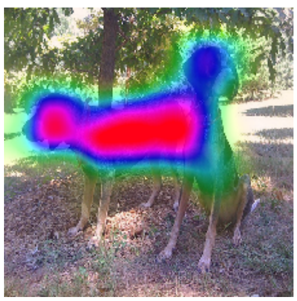
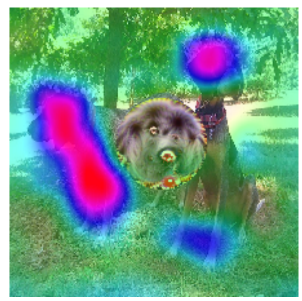
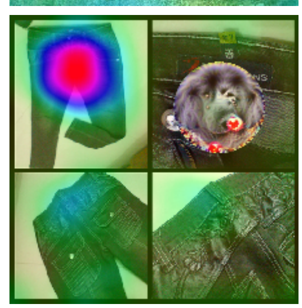
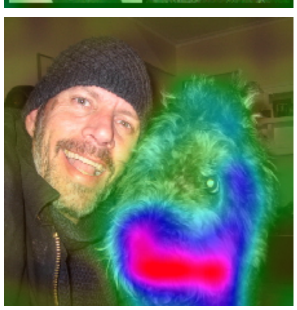
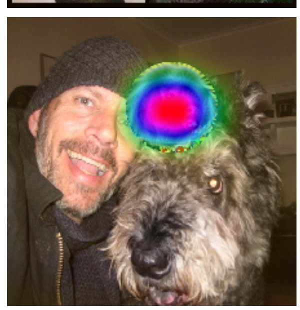
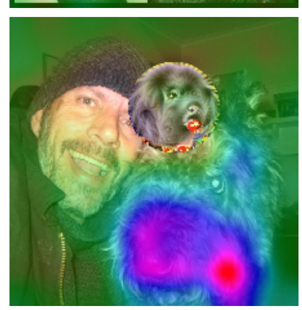
Class ID Orig.	Class ID Target	Clean Image	Adversarial Image Undefended Classifier	Adversarial Image VaN Defended Classifier
122	104			
134	104			
185	104			
78	104			
524	104			

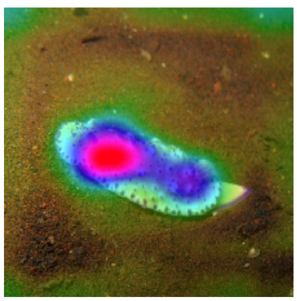
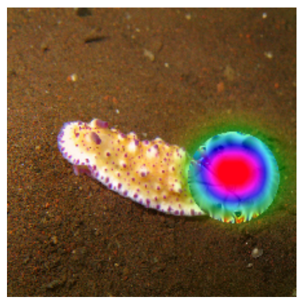
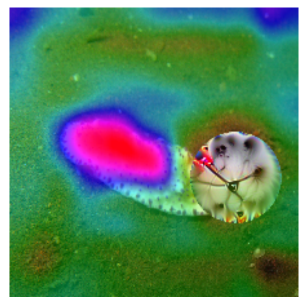
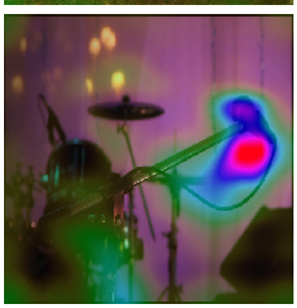
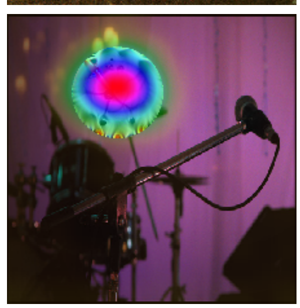
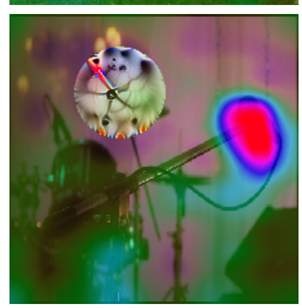
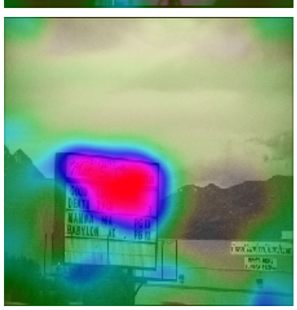
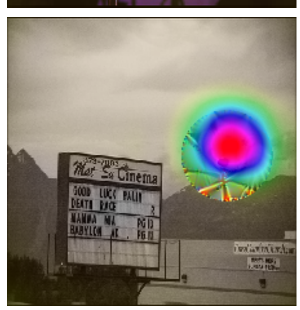


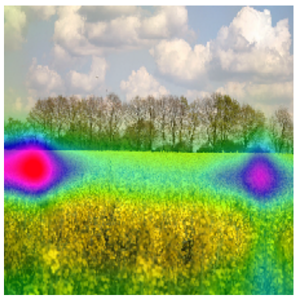
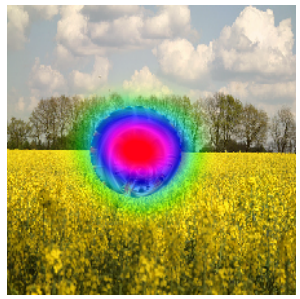
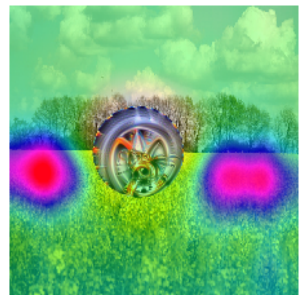
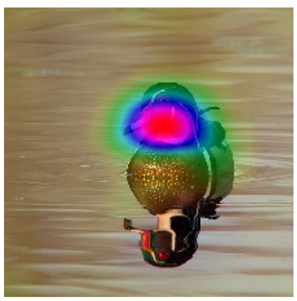
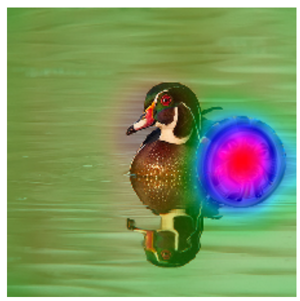
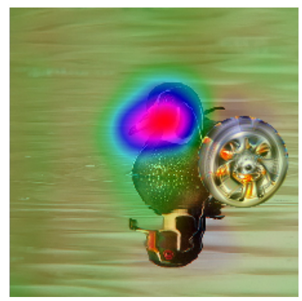
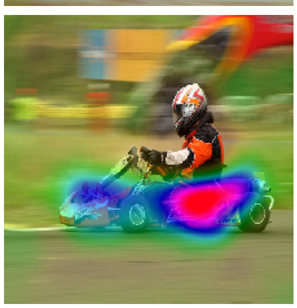
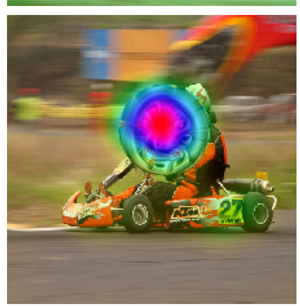
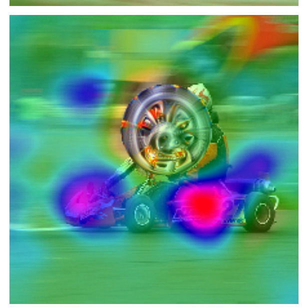
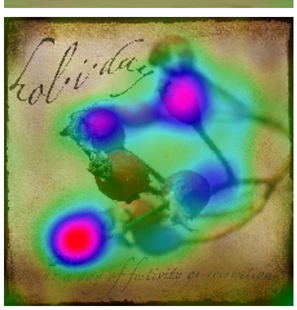
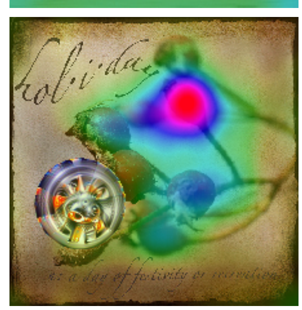
Class ID Orig.	Class ID Target	Clean Image	Adversarial Image Undefended Classifier	Adversarial Image VaN Defended Classifier
107	169			
605	169			
80	169			
372	169			
562	169			



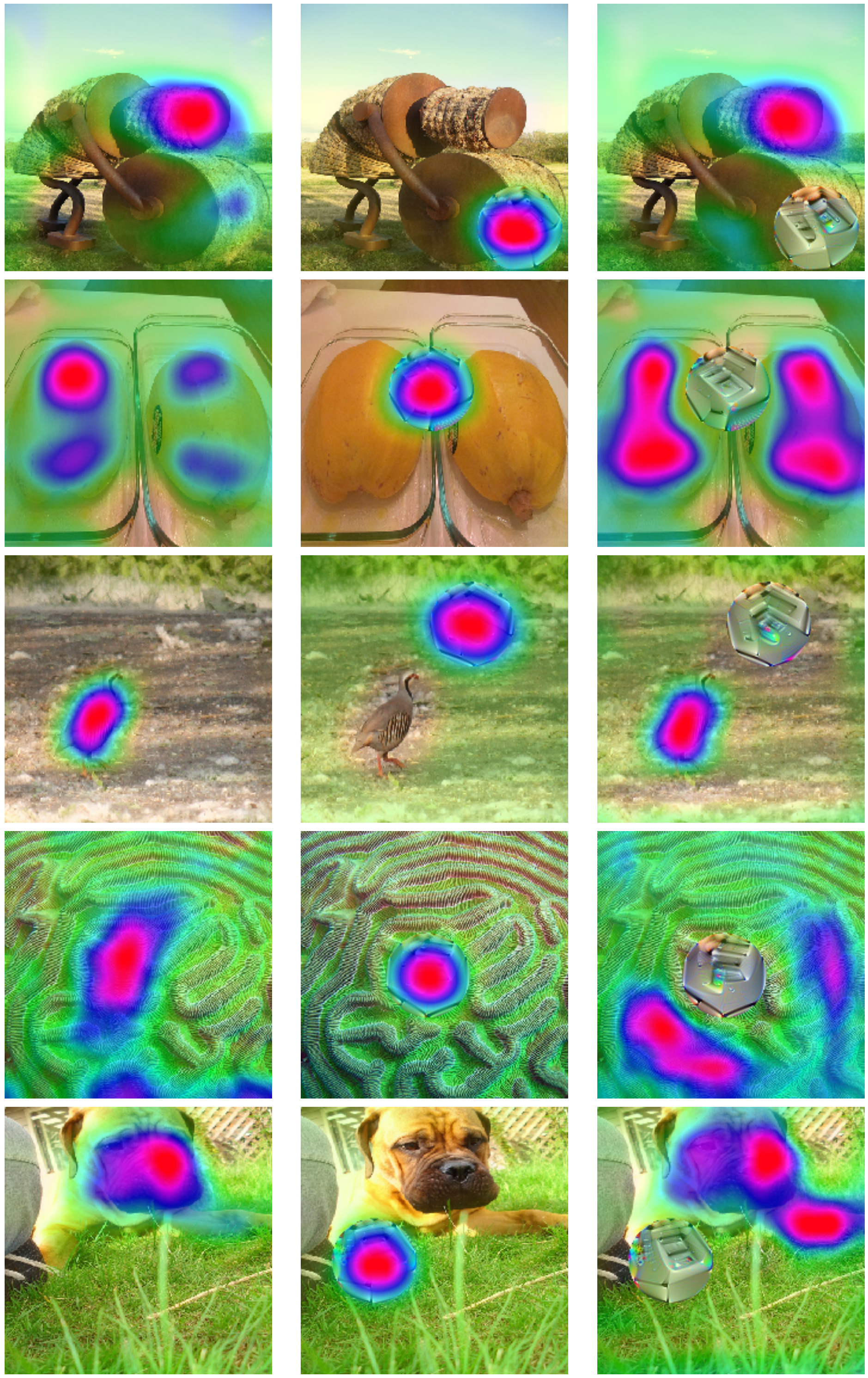

Class ID
Orig. Target
Clean Image
Adversarial Image
Undefended Classifier
Adversarial Image
VaN Defended Classifier
236 235
778 235
626 235
508 235
713 235

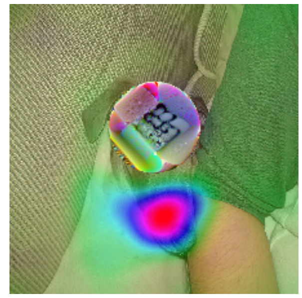
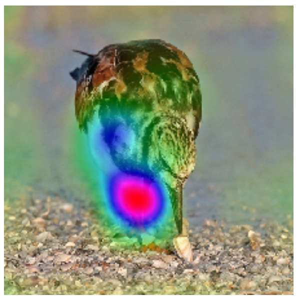
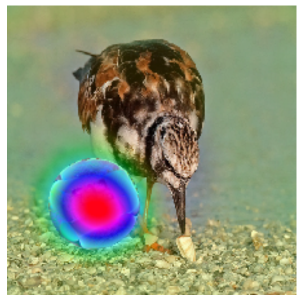
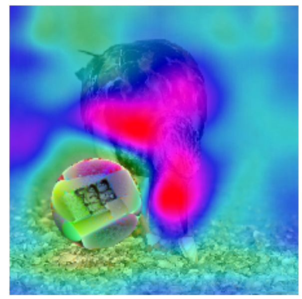
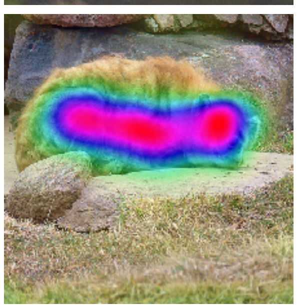
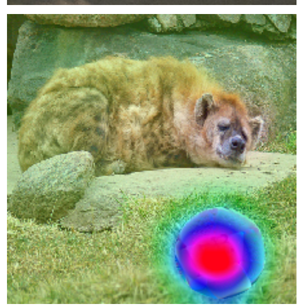
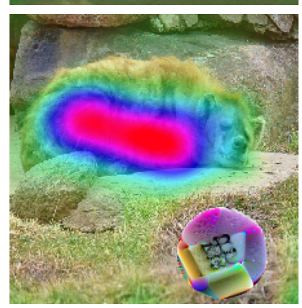
Class ID Orig. Target	Clean Image	Adversarial Image Undefended Classifier	Adversarial Image VaN Defended Classifier
738 256			
684 256			
164 256			
608 256			
233 256			

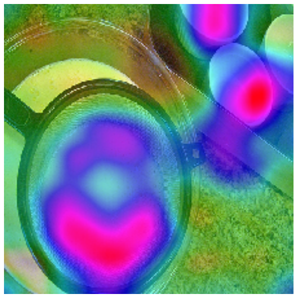
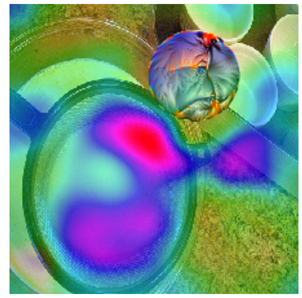
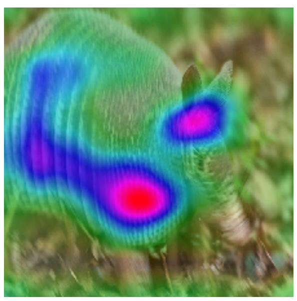
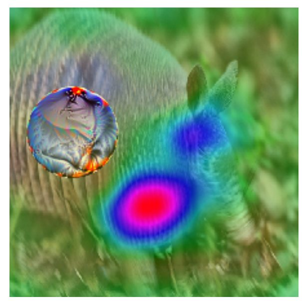
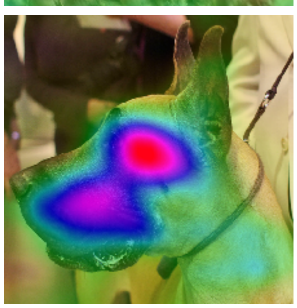
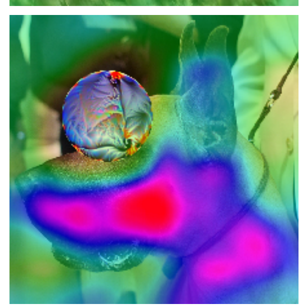
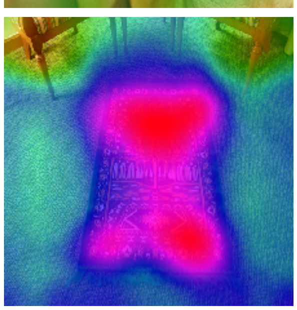
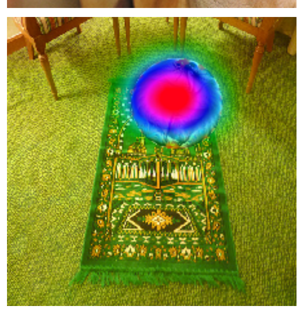
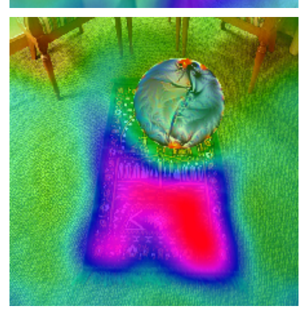
Class ID Orig.	Class ID Target	Clean Image	Adversarial Image Undefended Classifier	Adversarial Image VaN Defended Classifier
775	258			
253	258			
115	258			
650	258			
498	258			

Class ID Orig.	Target	Clean Image	Adversarial Image Undefended Classifier	Adversarial Image VaN Defended Classifier
946	479			
984	479			
97	479			
573	479			
989	479			

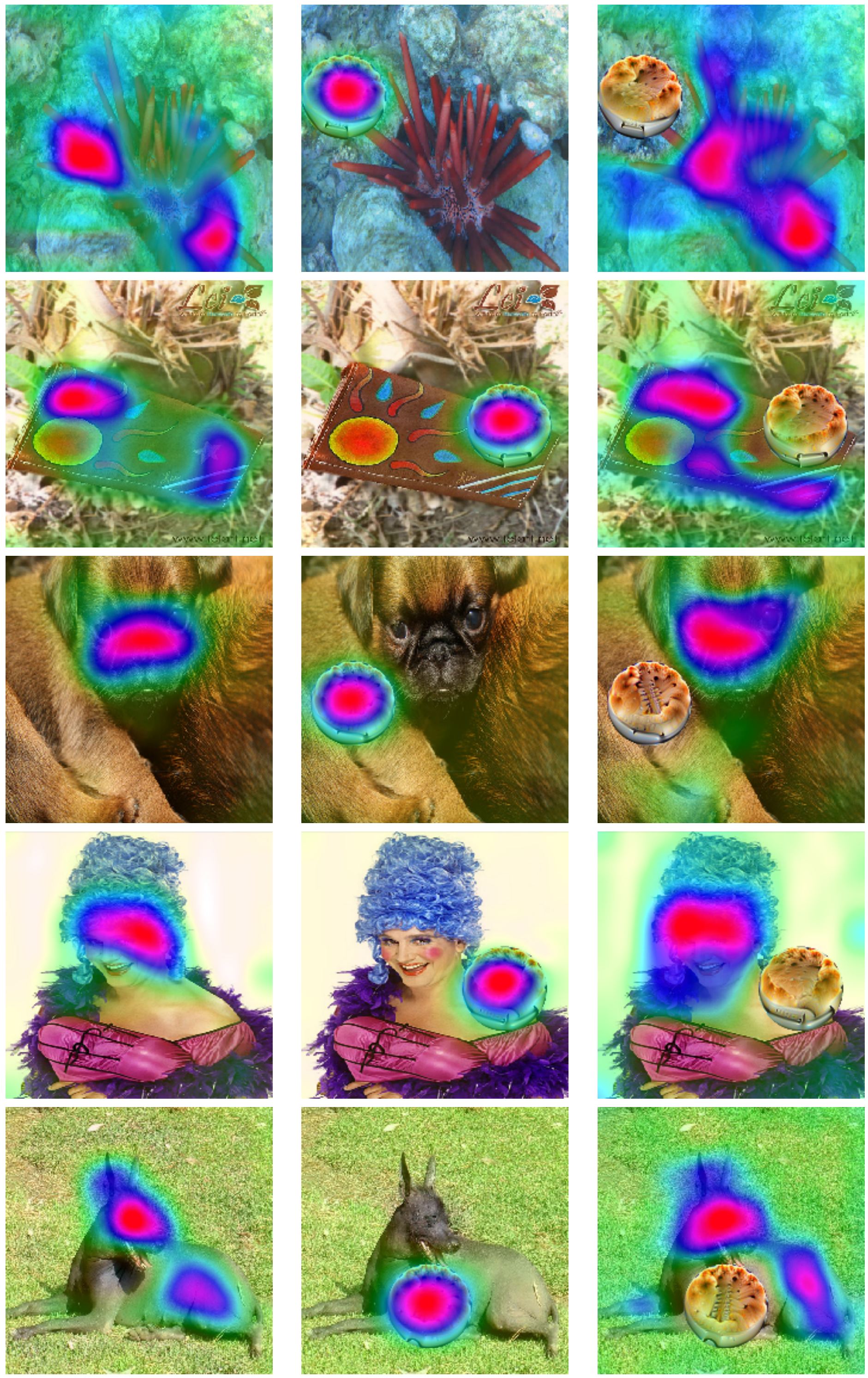
Class ID Orig.	Class ID Target	Clean Image	Adversarial Image Undefended Classifier	Adversarial Image VaN Defended Classifier
958	742			
940	742			
86	742			
109	742			
243	742			



Class ID Orig.	Class ID Target	Clean Image	Adversarial Image Undefended Classifier	Adversarial Image VaN Defended Classifier
171	767			
139	767			
675	767			
276	767			
291	767			

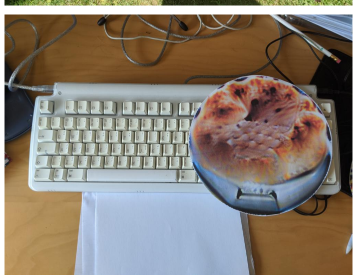
Class ID Orig.	Class ID Target	Clean Image	Adversarial Image Undefended Classifier	Adversarial Image VaN Defended Classifier
828	797			
783	797			
363	797			
246	797			
741	797			

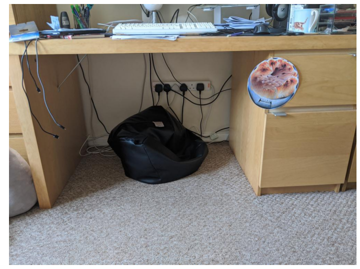
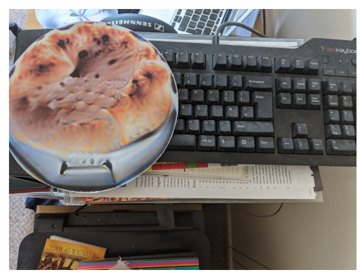
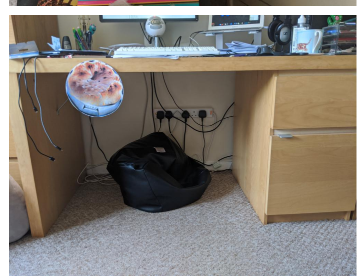
Class ID Orig.	Class ID Target	Clean Image	Adversarial Image Undefended Classifier	Adversarial Image VaN Defended Classifier
328	964			
893	964			
262	964			
903	964			
268	964			

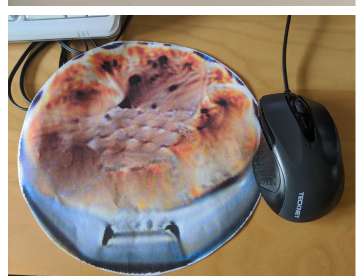
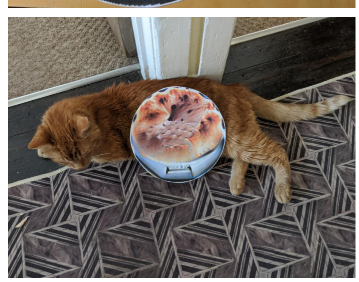


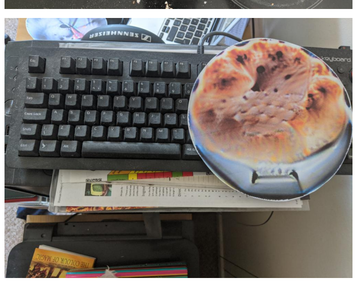
3 Physical Experiment Images

In this section we provide a sample of 50 images from our physical experiment, which was described in Section 4.6 of the main text.

Class ID	Test Image	Class ID	Test Image
721		518	
664		435	
428		742	
508		753	
611		611	

Class ID	Test Image	Class ID	Test Image
539		526	
435		636	
753		651	
544		508	
619		526	
659		636	

Class ID	Test Image	Class ID	Test Image
567		748	
539		508	
650		428	
742		409	
748		673	
673		281	

Class ID	Test Image	Class ID	Test Image
651		748	
650		664	
281		549	
761		549	
567		754	
508		748	

Class ID	Test Image	Class ID	Test Image
647		721	
409		647