

## Fooling Network Interpretation in Image Classification (Supplementary Material)

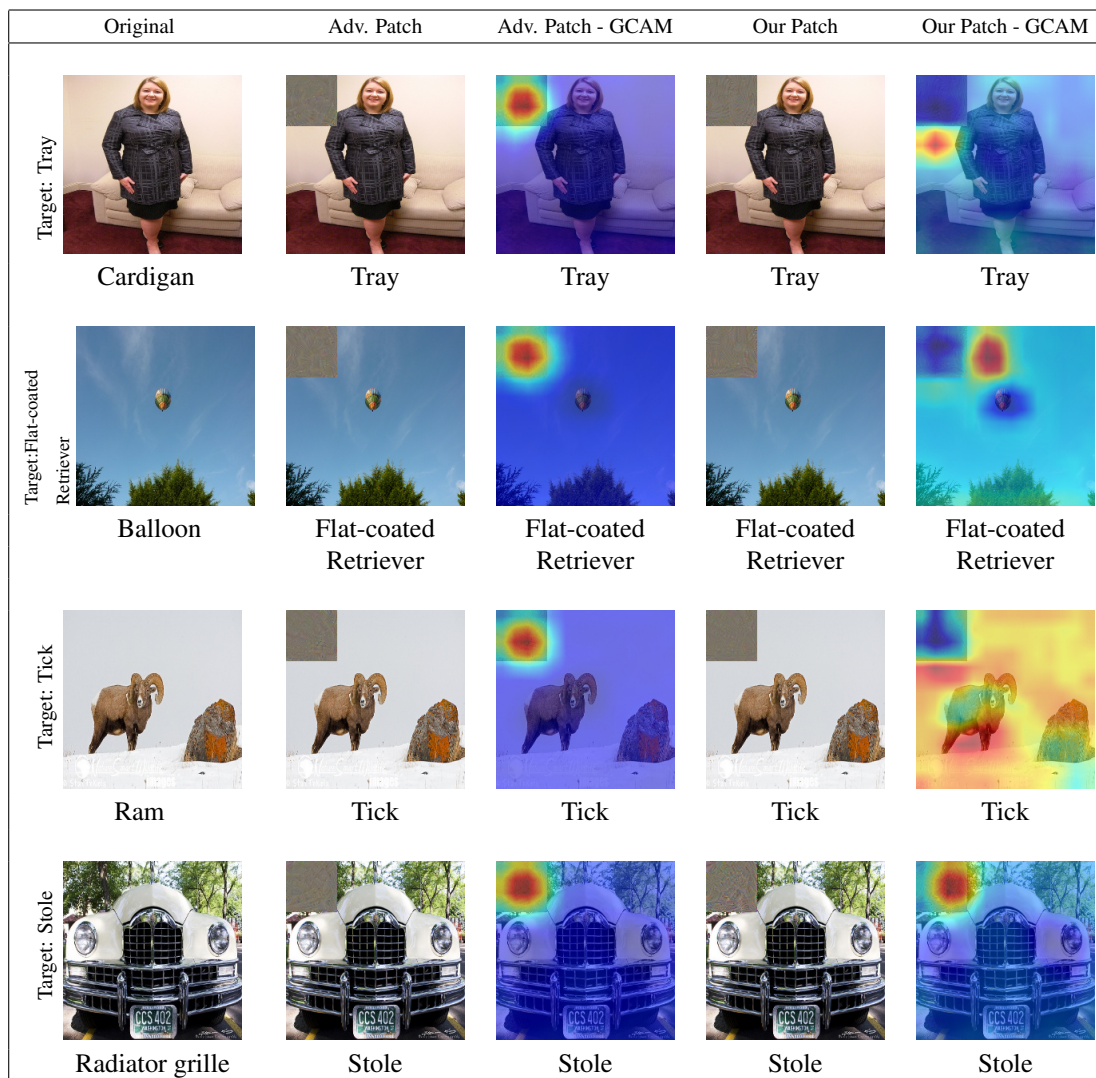


Figure S1: **Targeted Patch Attack:** Additional results for targeted patch attack for ImageNet pretrained VGG19-BN network similar to Figure 2 of the main paper. Images come from ImageNet validation set. The last row shows a failure case where our patch is not completely hidden in the interpretation.










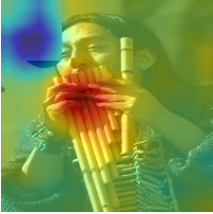







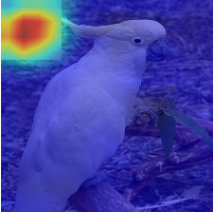

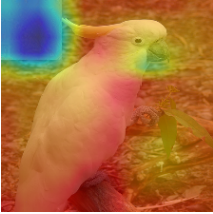
Original	Adv. Patch	Adv. Patch - GCAM	Our Patch	Our Patch - GCAM
				
Black grouse	Partridge	Partridge	Baseball	Baseball
				
Panpipe	Bath tissue	Bath tissue	Hornbill	Hornbill
				
Street sign	Lampshade	Lampshade	Patio	Patio
				
Cockatoo	Lycaenid	Lycaenid	Stinkhorn	Stinkhorn

Figure S2: **Non-targeted Patch Attack:** Comparison of Grad-CAM results for non-targeted patch attacks using our method vs regular adversarial patch. The description is in **Section 4.3** and the quantitative results are in **Table 1** of the main paper. We use ImageNet pre-trained VGG19-BN. The predicted label is written under each image, the non-targeted attack was successful for all images, and Grad-CAM is always computed for the predicted category. Images come from ImageNet validation set.








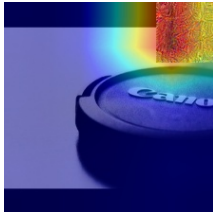

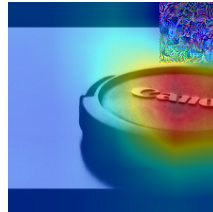
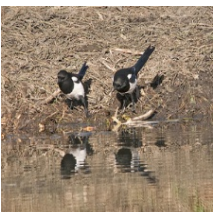
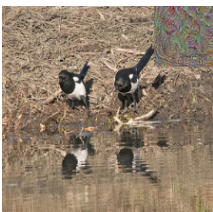
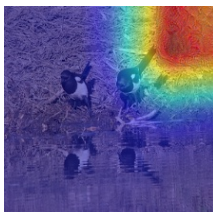

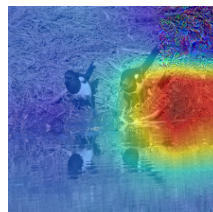


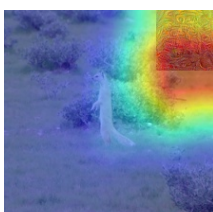

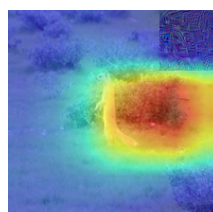
	Original	Adv. Patch	Adv. Patch - GCAM	Our Patch	Our Patch - GCAM
Target: White Shark					
	Beach Wagon	White Shark	White Shark	White Shark	White Shark
Target: Standard Poodle					
	Lens Cap	Standard Poodle	Standard Poodle	Standard Poodle	Standard Poodle
Target: Thimble					
	Magpie	Thimble	Thimble	Thimble	Thimble
Target: Wooden Spoon					
	Mongoose	Wooden Spoon	Wooden Spoon	Wooden Spoon	Wooden Spoon

Figure S3: **Different networks and patch locations:** Comparison of Grad-CAM visualization results for our targeted patch attack vs regular adversarial patch. It uses ImageNet pretrained **ResNet-34** network with the patch on the top right corner. The description is in **Section 4.4** and the quantitative results are in **Table 2** of the main paper. The predicted label is written under each image, the targeted attack was successful for all images in this figure, and Grad-CAM is always computed for the target category. Images are from ImageNet validation set.





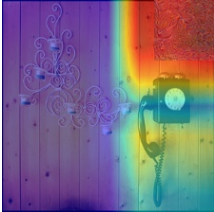

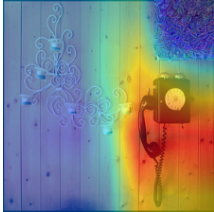


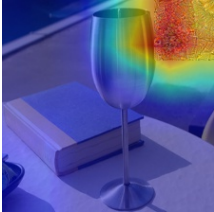
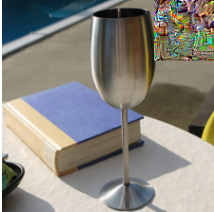
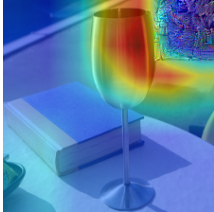






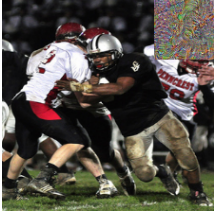



	Original	Adv. Patch	Adv. Patch - GCAM	Our Patch	Our Patch - GCAM
Target: Corkscrew					
	Dial Phone	Corkscrew	Corkscrew	Corkscrew	Corkscrew
Target: Toilet Seat					
	Goblet	Toilet Seat	Toilet Seat	Toilet Seat	Toilet Seat
Target: Doberman					
	Jeep	Doberman	Doberman	Doberman	Doberman
Target: Centipede					
	Football Helmet	Centipede	Centipede	Centipede	Centipede

Figure S4: Similar to Figure S3 of the supplementary, but for **DenseNet-121** network.









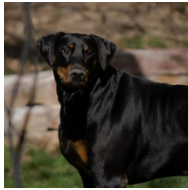
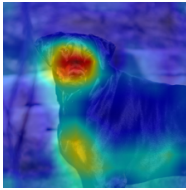
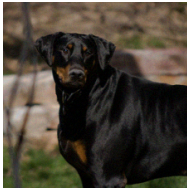
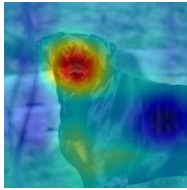

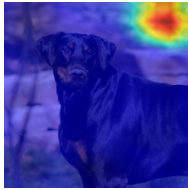





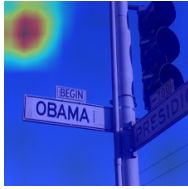
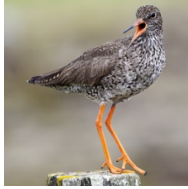
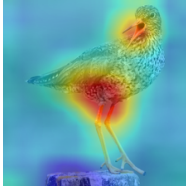

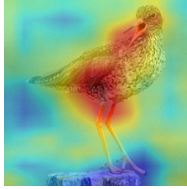
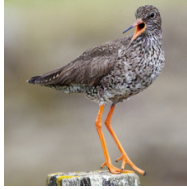
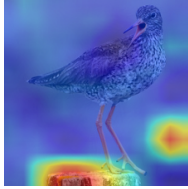
	Orig	Orig - GCAM	PGD Adv	PGD Adv - GCAM	Our Adv	Our Adv - GCAM
Target: European gallinule						
	Pred: Lifeboat	Grad-CAM "Lifeboat"	Pred: European gallinule	Grad-CAM "Lifeboat"	Pred: European gallinule	Grad-CAM "Lifeboat"
Target: Dowitcher						
	Pred: Doberman	Grad-CAM "Doberman"	Pred: Dowitcher	Grad-CAM "Doberman"	Pred: Dowitcher	Grad-CAM "Doberman"
Target: Leonberg						
	Pred: Street Sign	Grad-CAM "Street Sign"	Pred: Leonberg	Grad-CAM "Street Sign"	Pred: Leonberg	Grad-CAM "Street Sign"
Target: Porpie						
	Pred: Redshank	Grad-CAM "Redshank"	Pred: Potpie	Grad-CAM "Redshank"	Pred: Potpie	Grad-CAM "Redshank"

Figure S5: **Targeted regular adversarial examples:** As described in **Section 4.6** of the main paper, we use an ImageNet pretrained VGG 19-BN network to perform a targeted attack using our method as well as using standard PGD method. Note that in this case, unlike other experiments, we compare Grad-CAM for the *original* category and not the target one. The predicted label is written under each image. The attack was successful for all images. Note that compared to the original image and the PGD adversarial image, the Grad-CAM for our adversarial image fires less on the object. This attack not only reduces the probability of the original category, but also changes its interpretation. Images are from ImageNet validation set. The quantitative results are in **Table 4** of the main paper.


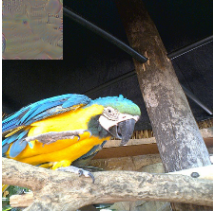
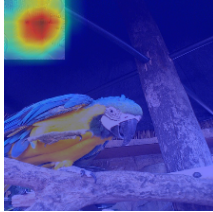
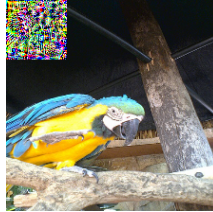
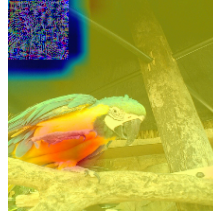



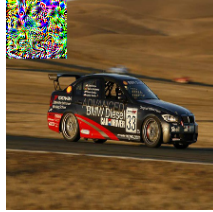



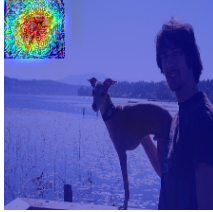
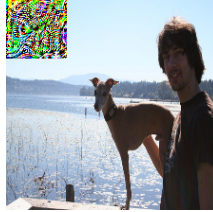




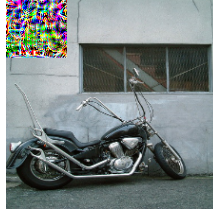
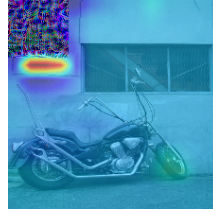
	Original	Adv. Patch	Adv. Patch - GCAM	Our Patch	Our Patch - GCAM
Target: Aeroplane	 Bird	 Aeroplane	 Aeroplane	 Aeroplane	 Aeroplane
Target: Bike	 Car	 Bike	 Bike	 Bike	 Bike
Target: Bird	 Person	 Bird	 Bird	 Bird	 Bird
Target: Boat	 Motorbike	 Boat	 Boat	 Boat	 Boat

Figure S6: **Universal targeted patch attack:** As described in **Section 4.9** of the main paper, we compare Grad-CAM of our universal attack on  $GAIN_{ext}$  with the regular adversarial patch. The predicted label is written under each image, the targeted attack was successful for all images, and Grad-CAM is always computed for the target category. Note that each row shows the result for a different category chosen as the universal target. Images are from PASCAL VOC-2012 validation set. The quantitative results are in **Table 7** of the main paper.

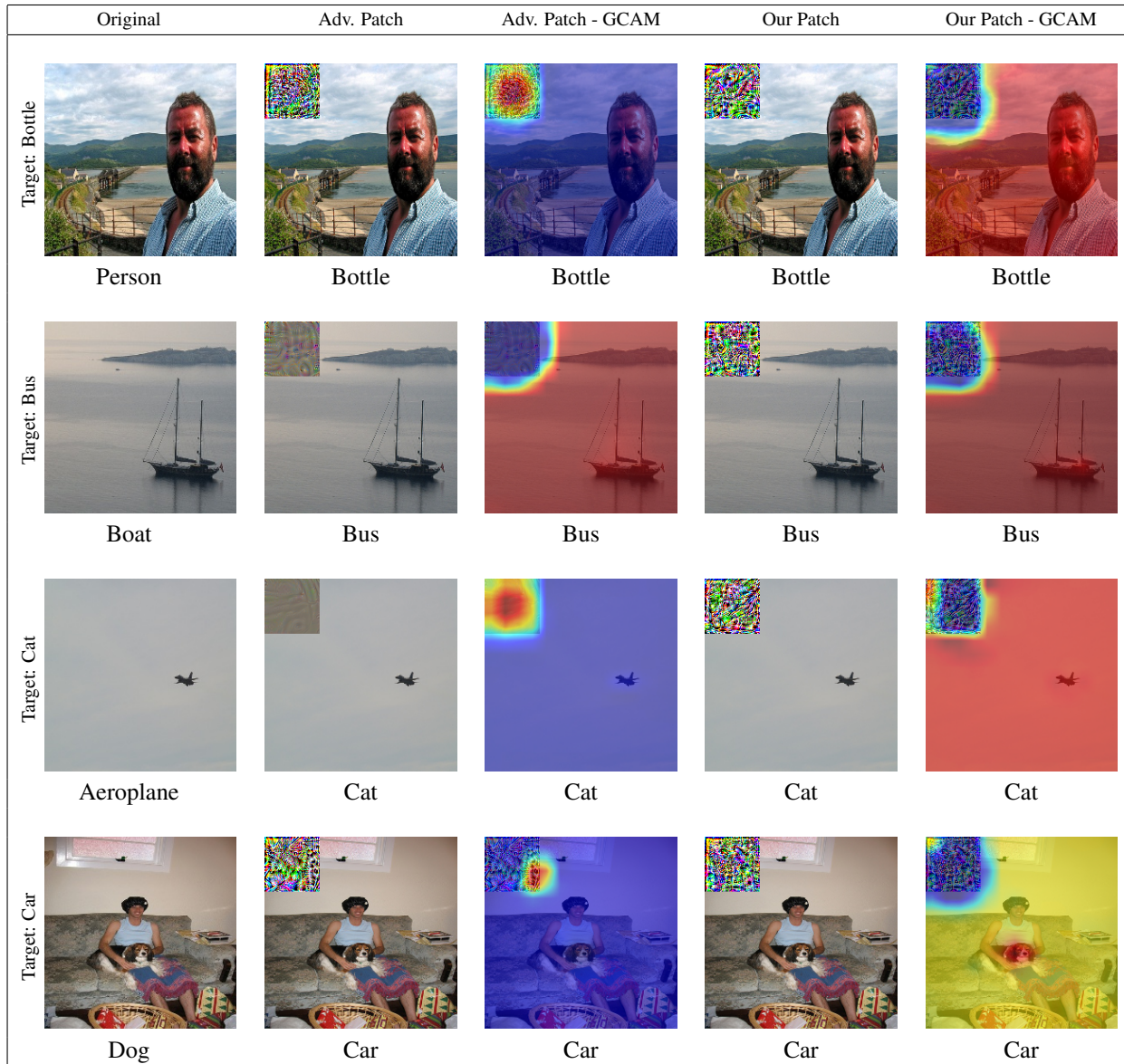


Figure S7: Similar to Figure S6 of the supplementary, but for different target categories. Interestingly, in the second row, regular adversarial patch is already hidden in Grad-CAM although it is not optimized for.



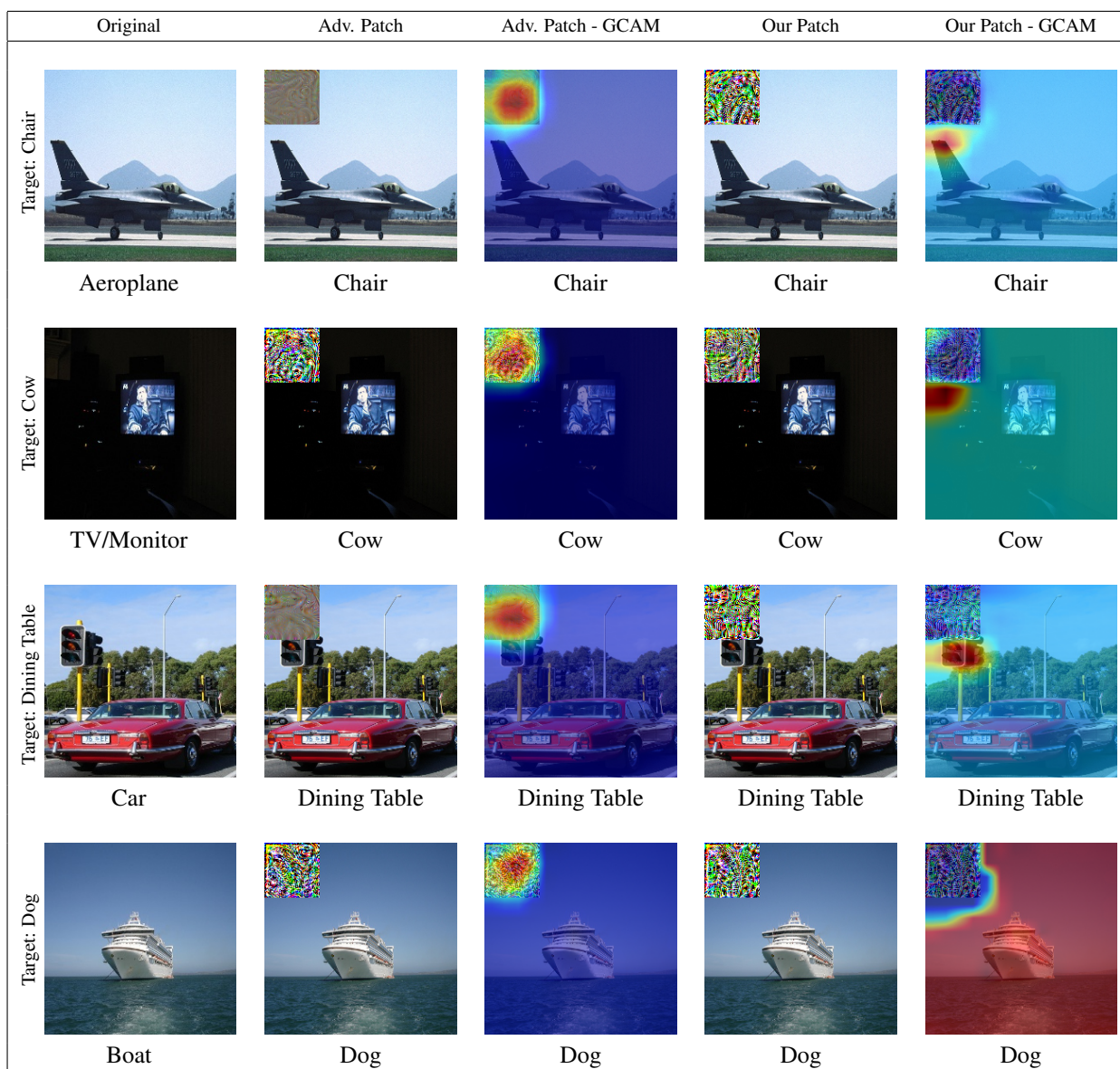


Figure S8: Similar to Figure S6 and S7 of the supplementary, but for different target categories.