Fully Exploiting Every Real Sample: SuperPixel Sample Gradient Model Stealing

Supplementary Material

1. Why does Sample Gradient contain model information?

'Dark knowledge' is a term used to describe knowledge that is implicitly embedded in a model but doesn't manifest directly, such as predictive logits. By learning 'dark knowledge', a surrogate model can inherit the characteristics and mimic the functionalities of the victim model. In prior work, sample gradients have been interpreted as a reflection of a model's local sensitivity to a specific input, guiding the perturbation direction in adversarial attacks. However, due to the inclusion of variance, instability, and a lack of interpretability, sample gradients are rarely considered as a form of dark knowledge. Numerous studies [20, 21, 23–26, 33] have attempted to utilize sample gradients to aid interpretability, often through altering the model, employing the gradient of the model's feature maps, or introducing additional inputs to propose interpretability methods. These methods fall short of establishing the interpretability of the original sample gradients. In this section, we introduce SG-Map, a method for interpreting sample gradients, designed without modifying the original model architecture, utilizing feature maps, or adding any additional inputs. We demonstrate that the sample gradients, processed and visualized as heatmaps, exhibit interpretability comparable to Grad-CAM.

For an individual input image, after backpropagation through the loss function, each pixel is assigned a gradient value, collectively forming the sample gradient. The SG-Map algorithm initiates by preprocessing the sample gradients: taking the absolute value of the gradient for each pixel and normalizing the pixels in each channel independently. This preprocessing ensures that the sample gradients meet the requirements for image display and eliminates numerical discrepancies in the sample gradients, which are tied to the parameter values of all neurons in the model and do not accurately reflect the model's decision-making characteristics. Channel-wise normalization is preferred over whole-image normalization due to the more substantial inter-pixel connections within channels than between them. SG-Map then combines the pixel gradients from the three different channels according to the specifications for a grayscale image, resulting in a single-channel sample gradient. In a crucial final step, we apply average pooling to this single-channel sample gradient, mitigating the impact of erratic behaviors from specific instances of the model on the pixel gradients. The resulting sample gradient is then presented as a heatmap. The visualization result is shown in Figure 1. Comparing SG-Map with CAM methods, we observe that SG-Map focuses on similar pixel locations, reflecting the model's sensitivity and attention allocation across different areas. Unlike CAM methods, the visualization of sample gradients through SG-Map provides a more stringent expression of sensitivity, manifesting as more concentrated yet precise high-temperature areas in the heatmap. Our proposed SG-Map thereby conclusively demonstrates that sample gradients encapsulate deep-seated information of the model, qualifying as a form of dark knowledge that can guide the training of surrogate models.

2. More Experiment Result

2.1. Hard-label experiments of Indoor-Scene

Under the query setup with hard labels, we use baselines to steal the resnet34 model trained on Indoor-Scenes. As shown in Table 3, SPSG still maintains the highest performance metrics.

2.2. Experiments on different Proxy's architecture

In the main text, we default to using the same proxy architecture as the victim, that is, resnet34. However, in practice, we cannot obtain information about the victim's model. Therefore, we experiment with different neural network architectures as proxys. The experimental results, as shown in Table 4, indicate that SPSG can effectively extract the performance of the victim across various neural network architectures. Due to the different performance ceilings inherent to each neural network architecture, the results present varying degrees of difference.

2.3. Impact of hyperparameter β

β is used to remove gradients outside of extreme values. The larger the β, the more gradients are removed, indicating a stricter selection of extremes. We conduct experiments under different β values. When the value of β is small, more gradient variance is introduced, leading to a decrease in the proxy model's performance. When β value is larger, there are hardly any superpixel gradients left. When β is at its maximum, the sample gradient contains only one superpixel gradient. When β is greater than 0.8, the performance of the proxy model remains at a lower level. This is because at this point, what is left are the most representative superpixel gradients, so the performance of the proxy model remains unchanged as it always mimics the most important

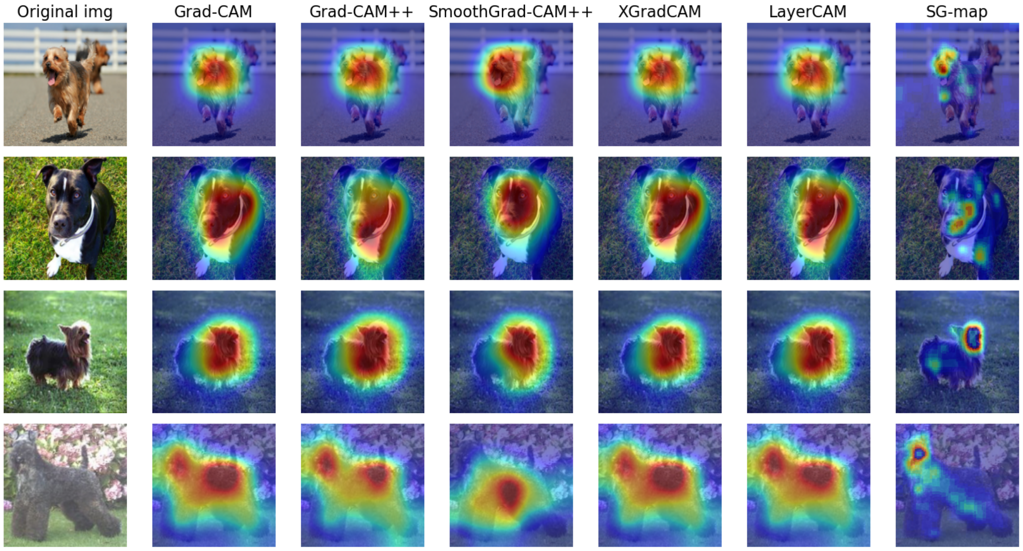

Figure 1. Partial visualization results of grad-CAM [20], grad-CAM++ [3], Smooth-gradCAM [14], X-gradCAM [4], layer-CAM [10], and SG-map. The neural network is ResNet34 pre-trained on ILSVRC-2012.

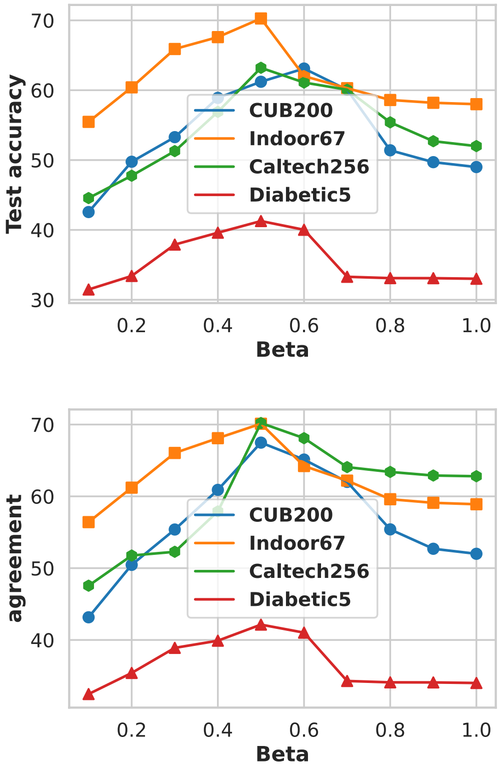

Figure 2

superpixel gradients. when β is 0.5, SPSG gets the best performance. The visulaization result is shown in Figure 2

2.4. Offline Training of SPSG

We observe the changes in simulated-superpixel gradients of samples during the offline training process of the proxy model, as shown in Figure 4. As the training epochs increase, the similarity between the pseudo-superpixel gradients obtained from the mean of model backpropagation pixel SG and the superpixel gradients queried from the victim model gets higher and higher. More similar sample gradients indicate that our loss function setting is reasonable. The proxy model sufficiently learns SG knowledge of the victim model.

2.5. Impact of Sample Selection Strategy

Our method does not conflict with sample selection strategies. Therefore, we compare the performance of SPSG, knockoff, and Black Dissector under two different sample selection strategies. These two strategies are the Reinforcement learning strategy [16] and the K-center strategy [17], as shown in Table 5. Both strategies improve the performance of MS to varying degrees. SPSG achieves the highest accuracy and similarity under both sample selection strategies. It is important to note that in the main text, we have already found that SPSG also obtains the best performance compared to other methods under a random sample selection strategy.

2.6. Ability to evade the SOTA defense method.

We conducted experiments with protection measures similar to [28] including Adaptive Misinformation [11], Prediction Poisoning [15], Gradient Redirection[2], External Feature[1]. The victim model is ResNet34 model trained on the CUBS-200-2011 dataset. The real sample number is 20k. The attack set employed ILSVRC-2012. As shown in Table 1, SPSG demonstrates significantly higher resistance to these two defenses compared to other methods.

Table 1. The larger the threshold, the better the defense effect (0.0 means no defense). ”False” and ”True” respectively correspond to evading monitoring and being detected by monitoring.

Method	No defence	AM	GR	EF	PP
Threshold	-	0.5	-	-	0.5
KnockoffNets	54.21±0.11	49.13±0.28	47.17±0.13	True	49.22±0.11
ActiveThief	55.24±0.12	50.12±0.11	47.71±0.21	True	49.11±0.21
Black-Box Dissector	56.98±0.21	51.21±0.31	48.81±0.14	False	49.03±0.28
Inversenet	55.17±0.19	49.21±0.73	51.28±0.31	False	48.72±0.49
DFMS	52.17±0.32	53.27±0.21	47.79±0.32	False	49.92±0.09
EDFBA	55.32±0.21	43.22±0.13	49.82±0.34	False	48.87±0.73
DS	51.33±0.12	54.33±0.37	51.11±0.94	True	50.01±0.56
DFME	53.28±0.23	52.26±0.61	50.87±0.34	False	51.06±0.72
SPSG	**61.33** ±0.03	**54.95** ±0.22	**52.15** ±0.34	False	**59.02** ±0.41
victim model	77.10	71.29	74.16	-	71.46

2.7. More study about SPGQ

We investigate the effectiveness of SPGQ and finite difference query methods. We compare the similarity of the sample gradients obtained by different methods to the real sample gradients. We use the average pair-wise distance to each real sample gradient as the evaluation metric and record the average number of queries required to query a sample. For superpixel gradients, we compare them by averaging the real samples within the corresponding superpixels. As shown in Table 2, our method exhibits a smaller distance compared to the finite difference method, indicating a higher similarity.

3. More study about SGP

Empirical validation of SGP's efficacy is demonstrated through two categories of experiments. On one hand, knowledge distillation experiments were conducted. When the student model distills unpurified sample gradients and logits knowledge, the resultant accuracy exhibits a decline due to irregular variance in the sample gradients, as compared to training without distillation. Conversely, distillation using purified sample gradients in conjunction with logits culminates in accuracy surpassing that achieved by distilling logits alone. On the other hand, T-SNE visualization was employed, concatenating the model-extracted sample features with the purified sample gradients. Comparative analysis reveals that purified sample gradients significantly enhance the final visualization outcome, as opposed to scenarios involving no concatenation or concatenation with unpurified sample gradients. The aforementioned experiments collectively attest to the effectiveness of the purification mechanism in eliminating variance from sample gradients.

3.1. knowledge distillation

We conducted experiments on image classification knowledge distillation. As shown in Figure 3, the sample gradients obtained from passing the samples through the teacher and student models are processed by SGP and then associated through the loss function for distillation. We selected ICKD [13], Overhaul [6], AT [31], FitNet [19], and FSP [30], KD [7], RKD [18], DIST [9], SRRL [29], and CRD [27] as the baselines. Initially, we compared the effects of each baseline with or without SGKD used individually in CIFAR100 [12]. The training strategies for CIFAR100 is shown in Table 6. We selected a series of teacher-student model combinations from VGG [22], ResNet [5], and their variants [32]. CIFAR100 Experimental Results: As depicted in Table 7, using SGP individually results in improved accuracy for student models. To further examine the influence of SGP on the knowledge distillation task, we compare the accuracy of the student model with and without SGP on the CIFAR100 dataset. As demonstrated in Table 8, the considerable numerical difference between the original sample gradients of the student and teacher hinders the accurate transfer of the teacher model's dark knowledge to the student model. This results in a decrease in the student model's performance. With SGP, the student model can effectively learn the teacher's dark knowledge.

3.2. T-SNE visualization

SGP allows pixel-level sample gradients to possess more class information. We first train a resnet34 on CIFAR-10. Then, we obtain the test sample feature vectors through the final layer before the output of resnet34. The feature vectors are visualized using T-SNE. Next, we concatenate the purified sample gradients or the original sample gradients behind the feature vectors and visualize them again. The four visualization results show that the sample gradients purified by SGP can effectively aid in classification. In contrast, the original sample gradients, containing variance and having low informational content, provide no benefit to the representation of feature vectors. The result is shown in Figure 5.

Result	Gradient Query Method				
	Finite Difference	SPGQ(QuickShift)	SPGQ(Slic)	SPGQ(Felzenszwalb)	Grid Query
Distance	2.987	2.178	2.089	2.078	16.679
Queries	150528	137	256	457	900

Table 2. The effectiveness of SPGQ and finite difference query methods.

Table 3. The agreement (in %), test accuracy (in %), and queries of each method with hard label. For our model, we report the average result as well as the standard deviation computed over 5 runs. (**Boldface**: the best value.)

Method (hard-label)	Indoor (10k)			Indoor (15k)			Indoor (20k)		
	Agreement	Acc	Queries	Agreement	Acc	Queries	Agreement	Acc	Queries
ZSDB3KD	27.55	26.43	1109k	29.52	30.07	1002k	34.21	33.71	1229k
DFMS	28.75	27.13	1321k	30.12	29.35	993k	34.23	33.15	989k
EDFBA	27.56	26.55	345k	30.34	29.48	477k	34.12	33.72	531k
DS	27.52	26.55	1200k	30.53	29.88	1090k	35.24	34.18	996k
knockoff	25.31	23.66	10k	27.19	25.73	15k	31.23	29.93	20k
ActiveThief	25.01	24.19	10k	27.59	26.13	15k	30.98	30.11	20k
Black-Box Dissector	25.91	23.57	20k	27.43	26.26	30k	31.59	30.46	40k
SPSG(Ours)	**27.86**±0.16	**26.79**±0.21	132k±0.01k	**31.43.27**±0.34	**30.29**±0.13	195k±0.01k	**38.27**±0.34	**36.32**±0.13	371k±0.01k

Table 4. The agreement (in %) and test accuracy (in %)s of each method with different proxy architecture [5, 8, 22] in CUBS-200-2011. The real sample number is 20k. For our model, we report the average result as well as the standard deviation computed over 5 runs. (**Boldface**: the best value.)

Method (probability)	ResNet-18		ResNet-50		VGG-16		DenseNet	
	Agreement	Acc	Agreement	Acc	Agreement	Acc	Agreement	Acc
ZSDB3KD	48.55	47.23	51.51	50.04	48.11	47.61	51.23	50.93
DFMS	49.55	48.11	51.46	50.25	48.63	47.85	52.13	51.93
EDFBA	49.76	48.55	51.32	50.11	48.72	47.84	51.13	49.98
DS	48.72	46.59	50.77	50.23	48.95	48.22	50.33	49.71
knockoff	45.39	43.98	47.28	45.63	49.78	48.23	50.21	49.28
ActiveThief	47.01	46.19	48.59	46.19	50.18	50.11	50.23	49.13
Black-Box Dissector	46.77	45.87	48.49	47.56	50.19	50.16	51.23	50.93
SPSG(Ours)	**49.96**±0.16	**49.70**±0.19	**52.27**±0.34	**51.39**±0.19	**51.21**±0.14	**51.12**±0.11	**52.71**±0.34	**52.32**±0.13

Table 5. The agreement (in %) and test accuracy (in %) of each method with different sample selection Strategies in CUBS-200-2011. For our model, we report the average result as well as the standard deviation computed over 5 runs. (**Boldface**: the best value.)

Method (probability)	CUBS200(10k)		CUBS200(15k)		CUBS200(20k)		CUBS200(25k)	
	Agreement	Acc	Agreement	Acc	Agreement	Acc	Agreement	Acc
knockoff (K-center)	55.65	52.71	59.32	55.71	61.77	57.61	63.21	61.93
knockoff (Reinforce)	54.37	50.11	56.76	54.21	59.67	57.15	61.13	59.92
Black-Box Dissector (K-center)	57.76	53.53	59.38	56.47	61.22	58.81	63.10	62.99
Black-Box Dissector (Reinforce)	55.71	52.12	58.71	56.23	61.95	59.11	61.31	60.71
SPSG (K-center)	59.27±0.11	**56.81**±0.11	**61.49**±0.17	**60.56**±0.14	63.17±0.14	62.16±0.12	**65.21**±0.31	**63.93**±0.11
SPSG (Reinforce)	**59.96**±0.16	56.70±0.19	61.27	60.39±0.19	**63.23**±0.14	**62.19**±0.11	64.11±0.34	62.32±0.13

Table 6. Strategies for CIFAR100

Strategy	Dataset	Epochs	Batch size	Initial LR	Optimizer	Weight decay	LR scheduler	Data augmentation
A1	CIFAR-100	240	64	0.05	SGD	0.0005	X0.1 at 150,180,210 epochs	crop+flip

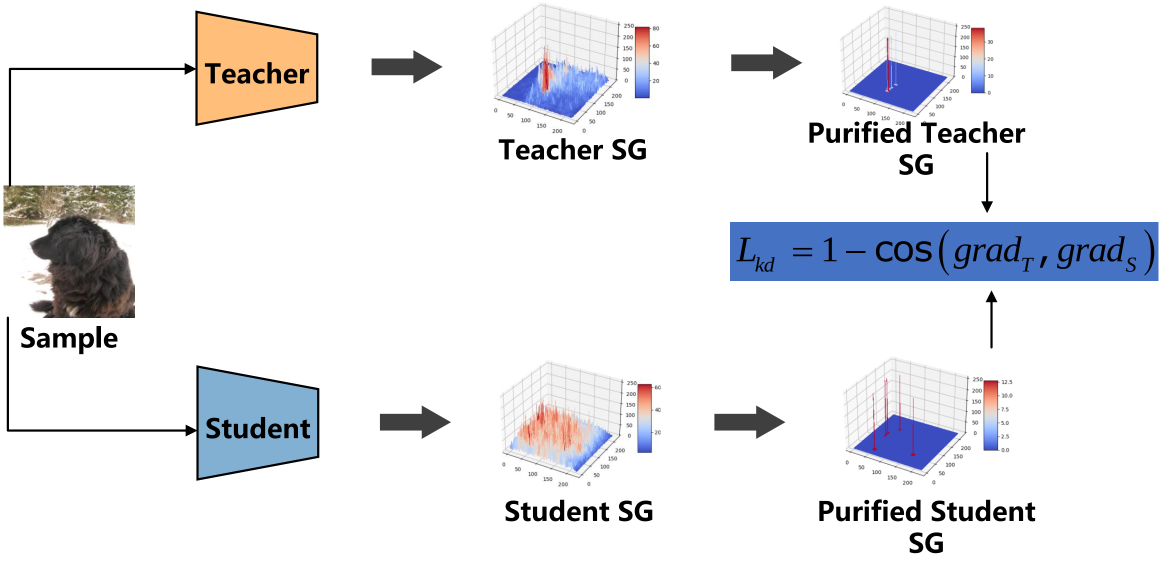

Figure 3. The framework of sample gradient knowledge distillation with SGP

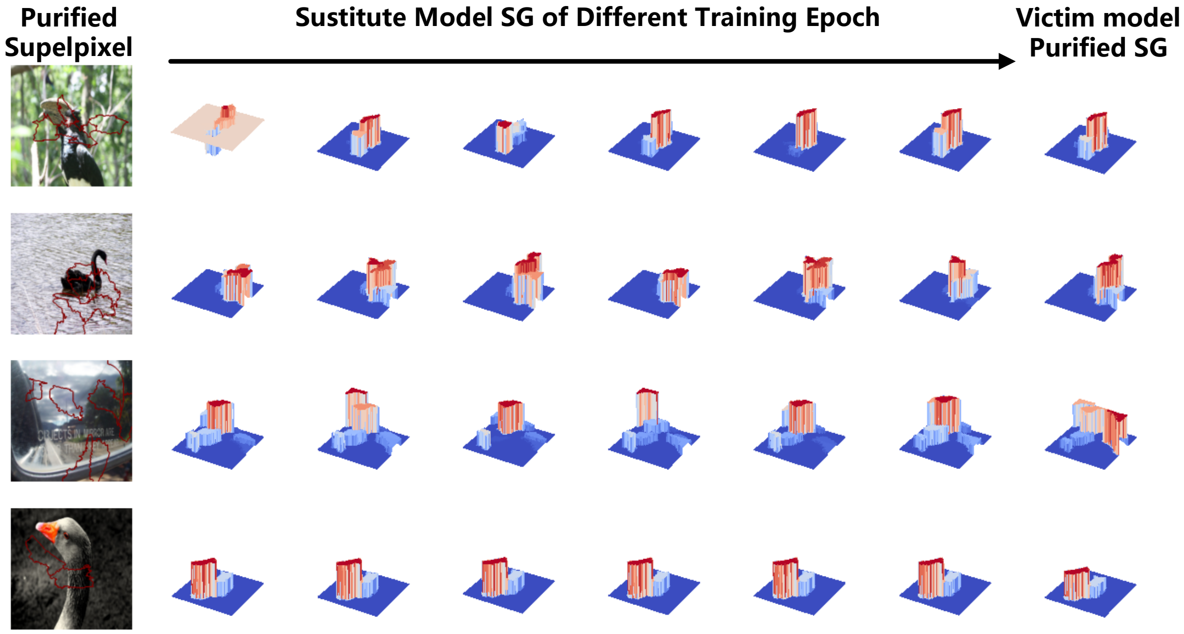

Figure 4. Simulated-superpixel gradients of samples during the offline training process

Table 7. Results (accuracy:%) in CIFAR100

Teacher Student	WRN-40-2 WRN-16-2	WRN-40-2 WRN-16-2 (with SGP)	ResNet110 ResNet20	ResNet110 ResNet20 (with SGP)	Vgg13 Vgg8	Vgg13 Vgg8 (with SGP)
Teacher Student	75.61 73.26	/ /	74.31 69.06	/ /	74.64 70.36	/ /
kD [7]	74.92 ± 1.1	75.48 ± 1.3↑	70.67 ± 2.5	71.78 ± 1.2↑	72.98 ± 1.2	72.78 ± 1.1↓
RKD [18]	73.35 ± 2.2	74.97 ± 2.1↑	69.25 ± 1.6	70.28 ± 1.8↑	71.48 ± 1.4	71.64 ± 1.3↑
DIST [9]	73.78 ± 1.3	74.79 ± 1.4↑	71.86 ± 1.2	71.91 ± 3.5↑	71.62 ± 1.7	71.67 ± 1.6↑
SRRL [29]	73.71 ± 1.7	75.16 ± 3.1↑	70.91 ± 1.4	71.23 ± 2.7↑	71.45 ± 1.8	71.78 ± 2.1↑
CRD [27]	75.48 ± 2.1	75.61 ± 1.1↑	71.46 ± 1.7	71.57 ± 3.6↑	73.94 ± 1.2	73.62 ± 2.3↓
SGKD	74.91 ± 1.1	/	71.48 ± 1.7	/	72.61 ± 2.1	/
ICKD [13]	75.34 ± 1.2	75.44 ± 1.6↑	71.91 ± 1.3	72.01 ± 2.1↑	73.88 ± 2.2	74.09 ± 1.3↑
overhaul [6]	75.52 ± 1.3	75.48 ± 1.4↓	71.21 ± 1.5	71.34 ± 1.2↑	73.42 ± 1.1	73.57 ± 1.9↑
AT [31]	74.08 ± 1.7	74.61 ± 2.3↑	70.22 ± 2.1	71.24 ± 1.8↑	71.43 ± 1.9	72.86 ± 2.1↑
FitNet [19]	73.58 ± 2.3	73.67 ± 2.1↑	68.99 ± 1.2	71.67 ± 2.3↑	71.02 ± 2.5	72.21 ± 2.4↑
FSP [30]	72.91 ± 2.1	73.28 ± 1.7↑	70.11 ± 1.5	71.94 ± 1.5↑	70.23 ± 1.3	71.47 ± 1.5↑

Table 8. Ablation study on CIFAR100.

Teacher Student	WRN-40-2 WRN-16-2 (with SG)	WRN-40-2 WRN-16-2 (with SGP)	ResNet110 ResNet20 (with SG)	ResNet110 ResNet20 (with SGP)	Vgg13 Vgg8 (with SG)	Vgg13 Vgg8 (with SGP)
Teacher Student	75.61 73.26	/ /	74.31 69.06	/ /	74.64 70.36	/ /
KD[7]	72.13 ± 1.4↓	75.48 ± 1.3↑	68.77 ± 2.1↓	71.78 ± 1.2↑	69.58 ± 1.2↓	72.78 ± 1.1↓
RKD [18]	71.39 ± 2.0↓	74.97 ± 2.1↑	68.45 ± 2.6↓	70.28 ± 1.8↑	69.18 ± 1.2↓	71.64 ± 1.3↑
DIST [9]	72.44 ± 1.2↓	74.79 ± 1.4↑	68.76 ± 1.3↓	71.91 ± 3.5↑	69.52 ± 1.3↓	71.67 ± 1.6↑
SRRL [29]	73.16 ± 1.2↓	75.16 ± 3.1↑	68.45 ± 1.8↓	71.23 ± 2.7↑	70.16 ± 1.4↓	71.78 ± 2.1↑
CRD [27]	72.18 ± 1.8↓	75.61 ± 1.1↑	68.66 ± 2.7↓	71.57 ± 3.6↑	70.11 ± 1.2↓	73.62 ± 2.3↓
ICKD [13]	72.39 ± 1.2↓	75.44 ± 1.6↑	71.91 ± 1.5↓	72.01 ± 2.1↑	68.34 ± 2.7↓	74.09 ± 1.3↑
overhaul [6]	71.98 ± 2.1↓	75.48 ± 1.4↓	71.21 ± 1.3↓	71.34 ± 1.2↑	69.67 ± 1.4↓	73.57 ± 1.9↑
AT [31]	72.78 ± 1.1↓	74.61 ± 2.3↑	70.22 ± 2.7↓	71.24 ± 1.8↑	69.89 ± 1.5↓	72.86 ± 2.1↑
FitNet[19]	72.18 ± 1.2↓	73.67 ± 2.1↑	68.99 ± 1.2↓	71.67 ± 2.3↑	69.78 ± 2.1↓	72.21 ± 2.4↑
FSP [30]	72.11 ± 2.1↓	73.28 ± 1.7↑	70.11 ± 1.3↓	71.94 ± 1.5↑	69.35 ± 1.3↓	71.47 ± 1.5↑

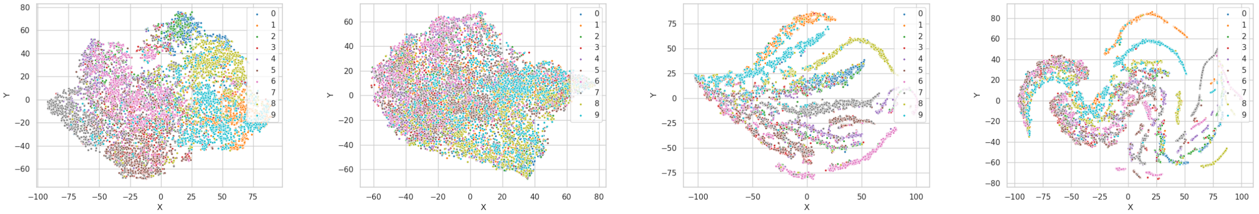

Figure 5. The images from left to right are original feature vector classification, feature vector classification with original SG, feature vector classification with purified SG as $\beta = 0.5$, and classification with purified SG as $\beta = 0.2$.

References

[1] Defending against model stealing via verifying embedded external features. 2

[2] How to steer your adversary: Targeted and efficient model stealing defenses with gradient redirection. 2

[3] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847. IEEE, 2018. 2

[4] Kun Fu, Quan Jin, Runze Cui, Fei Sha, and Changqing Zhang. Axiom-based grad-cam: Towards accurate visualization and explanation of cnns. *Pattern Recognition*, 110: 107638, 2021. 2

[5] 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. 3, 4

[6] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1921–1930, 2019. 3, 6

[7] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 3, 6

[8] 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. 4

[9] Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from a stronger teacher. *arXiv preprint arXiv:2205.10536*, 2022. 3, 6

[10] Ziqi Jiang, Li Zhang, Chuang Zhang, Chunxia Li, Fei Li, and Kuiyuan Yang. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 30:5875–5888, 2021. 2

[11] Sanjay Kariyappa and Moinuddin K Qureshi. Defending against model stealing attacks with adaptive misinformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2020. 2

[12] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 3

[13] Li Liu, Qingle Huang, Sihao Lin, Hongwei Xie, Bing Wang, Xiaojun Chang, and Xiaodan Liang. Exploring inter-channel correlation for diversity-preserved knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8271–8280, 2021. 3, 6

[14] Daniel Omeiza, Simon Speakman, Celia Cintas, and Komminist Weldemariam. Smooth grad-cam++: An enhanced inference level visualization technique for deep convolutional neural network models. In *2019 15th International Conference on Signal-Image Technology Internet-Based Systems (SITIS)*, pages 272–279. IEEE, 2019. 2

[15] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. Prediction poisoning: Towards defenses against dnn model stealing attacks. *arXiv preprint arXiv:1906.10908*, 2019. 2

[16] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. Knockoff nets: Stealing functionality of black-box models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4954–4963, 2019. 2

[17] Soham Pal, Yash Gupta, Aditya Shukla, Aditya Kanade, Shirish Shevade, and Vinod Ganapathy. Activethief: Model extraction using active learning and unannotated public data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 865–872, 2020. 2

[18] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3967–3976, 2019. 3, 6

[19] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014. 3, 6

[20] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 1, 2

[21] Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*, 2016. 1

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

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

[24] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

[25] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

[26] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017. 1

[27] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*, 2019. 3, 6

[28] Yixu Wang, Jie Li, Hong Liu, Yan Wang, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Black-box dissector: Towards erasing-based hard-label model stealing attack. In *European Conference on Computer Vision*, pages 192–208. Springer, 2022. 2

[29] Jing Yang, Brais Martinez, Adrian Bulat, Georgios Tzimiropoulos, et al. Knowledge distillation via softmax regression representation learning. International Conference on Learning Representations (ICLR), 2021. 3, 6

[30] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4133–4141, 2017. 3, 6

[31] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016. 3, 6

[32] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. 3

[33] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 1