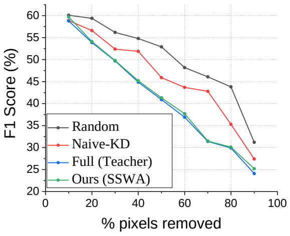
We thank the reviewers for their constructive feedback. We are encouraged that all reviewers found our observation and motivation - attribution map distortion while compressing networks - to be novel and interesting [**R1**,**R2**,**R3**,**R4**] and our approach to be intuitive [**R1**], readily applied [**R2**], and evaluated with convincing experiments [**R1**,**R3**]. We address the reviewer comments as much as space permits and will include all feedback in the final version.

	Predictive Performance		Attribution Score	
Method	mAP	F1 Score	AUC	Point Acc
Full (Teacher)	91.79±0.16	78.44±0.23	88.68±0.17	80.16±0.16
Naive-KD	81.44±0.19	62.71±0.28	80.51±0.25	69.18±0.17
EWA	82.31±0.22	63.16±0.31	84.23±0.22	79.19±0.22
SWA	83.63±0.19	64.94±0.24	87.73±0.19	79.91±0.21
SSWA	**84.37**±**0.26**	**66.27**±**0.35**	**88.13**±**0.22**	**80.14**±**0.24**

[R1,R2] Evaluation of predictive performance and attribution score with *Error Bound*. We reported the (multi-label) predictive performance measure in main the text with *mean-average-precision (mAP)*. We apologize for not elaborating this [**R1**,**R2**]. Also, "accuracy" in Appendix table 3,4 refers to the ImageNet top-1 accuracy [**R1**]. Moreover, we will add one more predictive performance measure: F1 score. Following the [**R2**]'s comments, we additionally report the performances with standard variation through the 5 same experiments to be more convincing results. In above the table, we report the predictive performances and attribution scores of various methods including our methods for knowledge distillation with vgg/4. For other compression cases, we will incorporate in the final version.



[R1,R2,R3,R4] Other evaluation of attribution map & Reliability of Grad-Cam. Following **R1**'s suggestion, we will add the perturbation metric, RemOve And Retrain (ROAR) [A], to evaluate the how well attribution maps from compressed networks explain the model behavior. This test is as follows: remove the top-k pixels of an image ranked by the attribution map for the entire training data, and retrain a classifier on this perturbed dataset. If the attribution map accurately represents the importance of the pixels, the classifier must have lower predictive performance. Here, we show a quick experiment using vgg-11 classifier. **(a)** In the right graph, all Grad-Cam perturbations (from different models) were able to lower the F1 score more than random perturbations, which verifies that Grad-Cam indeed reflects a model's decision making process. **(b)** The student trained with our method scored almost in par with the full network. This indicates that the attributions (which reflect a model's decision process) are indeed preserved by our method.

AUC/PointAcc	Excitation Bp	$\text{LRP}_{\alpha=1\beta=0}$	RAP
Full (Teacher)	84.2/74.8	85.3/65.5	84.5/69.5
Naive-KD	76.3/66.3	79.6/53.4	80.9/56.9
SSWA (Ours)	**82.3/71.2**	**82.5/64.1**	**83.5/65.7**

[R4] Preserving differentiable attribution map also preserves other attributions. We observed the deformation of various attribution maps in Appendix A. Following **R4**'s suggestion, we evaluated other attribution maps for the model using our method. We observe that it also helps preserving other attribution maps. Since most attribution maps [14, 15] are generated with gradients and activations, preserving one may help others.

[R2,R3,R4] Technical novelty. In [4], the authors introduce attention map transfer and gradient transfer. Although our method and theirs share certain aspects, we believe that there are enough differences to them. **(a)** The problem of focus is distinct. [4] focuses on boosting the *predictive performance* of a student network, while our method focuses on preserving the *attribution* power of a network while being compressed. **(b)** The function of the loss functions are different. Attribution map used in attention matching in [4] lacks label-specific attribution information since all activation maps are equally weighted and aggregated. In other words, this form of attribution map may tell where the network is looking, but it holds no "meaning". Thus, this regularizer may teach the student how to look and distinguish objects, but does not pass on the information of "what" and "why" it should look at a certain region. Gradient matching of [4] does hold some label-specific information. However, since they only match the gradients at the input level, the information in the intermediate layers (and thus their decision processes) are not appropriately transferred while our method is able to transfer information from intermediate layers by collapsing the channels.

[R2,R3] Motivation of stochastic matching. The intent behind our stochastic matching regularizer is to facilitate the transfer of relevant information and prevent overfitting. Several recent works utilize this concept to boost the performance of knowledge transfer between a teacher and a student. In [23], they encourage the teacher-student information transfer by using a gaussian dropout to maximize the mutual information between teacher and student features. Injection of stochasticity can also be found on other similar fields such as continual learning [B] and domain adaptation [C]. Based on the empirical evidences of performance increase presented by the literature mentioned above, we believe that it is plausible to motivate our stochastic matching method.

[R3] Hypothetical explanation for the attribution deformation problem. As a network is compressed, it must cram its decision procedures(information) inside a smaller memory. If so, it would be unable to use "standard" decision procedures, but must resort to using shortcuts and inklings that means less to humans. Thus, its decision procedures would become harder to interpret which is reflected in its distorted attribution map.

Clarity and minor details. [**R2**] Is the "full net" finetuned or trained from scratch?: *Fine-tuned.* [**R3**]$\cdots$ is Grad-Cam used for some other purpose?: *the qualities of Grad-CAM maps extracted from EWA, SWA, SSWA were evaluated.* [**R3**] We apologize for the mis-reporting: In Appendix Table3, the 'Full Network''s AUC should be changed: *75.92* → *81.64*

[A] Hooker, Sara, et al. "A benchmark for interpretability methods in deep neural networks." NeurIPS. 2019.; [B] Lee, et al. "Overcoming Catastrophic Forgetting by Incremental Moment Matching", NeurIPS. 2017.; [C] Saito, et al. Adversarial Dropout Regularization, ICLR. 2018.;