Supplementary Material for
Proto2Proto: Can you recognize the car, the way I do?

In this supplementary, we include the following details which could not be included in the main paper due to space constraints.

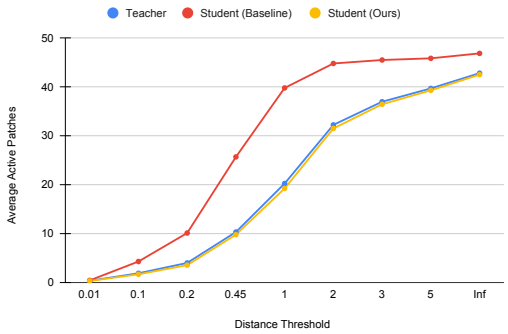
1. Additional Ablations (Appendix A)
2. Additional Results (Appendix B)
3. Modified Jaccard Similarity (Appendix C)
4. Few-Shot Experiments (Appendix D)
5. Hyperparameter Details (Appendix E)
6. Visualizations (Appendix F)

A. Additional Ablations

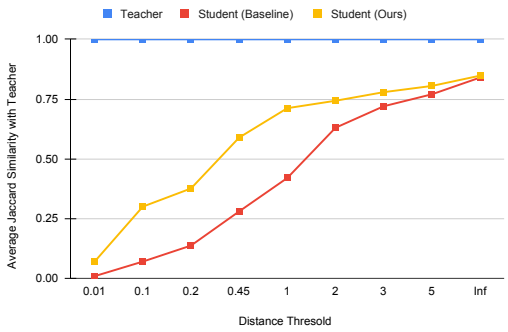
Ablation on distance threshold (τ_{test}) Figure S1a shows variation of AAP for different values of τ_{test}. As evident, our student model is close to the teacher model over the entire range of τ_{test}. Figure S1b shows variation of AJS for different values of τ_{test}. As observed, the gap between our student model and teacher is smaller as compared to baseline student and teacher for all values of τ_{test}. In results, we report the performance by doing a grid search on the values of $\tau_{test} = [\inf, 5.0, 1.0, 0.45, 0.1]$ for L_2 distance metric.

B. Additional Results

Table S1 summarizes the results of our proposed method on VGG metwork. The experiment was performed on VGG19, VGG16 and VGG11 models. Our proposed student model clearly outperforms the baseline student model in all the experiments as indicated by Top-1 Accuracy. For example, the VGG16 (Teacher)→VGG11 (Student), there is 3.33 percent of absolute improvement in accuracy compared to baseline student. Eventhough there is marginal improvement for VGG19 (Teacher)→VGG16 (Student), the interpretability evaluation metrics of AAP, AJS and PMS indicate a significant difference. This shows that proposed method performs well on different architectures and for measures with respect to accuracy and interpretability.



(a) Distance Threshold (τ_{test}) v/s AAP



(b) Distance Threshold (τ_{test}) v/s AJS with Teacher

Figure S1. Effect of τ_{test} on AAP and AJS. Observe the closeness of the student with the teacher. The AAP score of our student is almost same as that of teacher and AJS of our student is much better than baseline student.

C. Modified Jaccard Similarity (MJS)

Consider two sets A and B. The Jaccard Similarity between A and B is given by,

$$\text{JS(A, B)} = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

As the size of the sets A and B increases, the denominator tends to increase at a faster rate than the numerator, which decreases the the value of JS(A,B). We observed this drawback was prevalent in Prototype Matching Score where we maintain a list of active patches across all images, for

Datasets	Method	Setting	AAP	AJS (↑)	PMS (↑)	Top-1 Accuracy (↑)
CUB	ProtopNet	VGG16 (Teacher)	32.80	1.0	1.0	76.35
	ProtopNet	VGG11 (Student)	37.92	0.63	0.34	71.62
	Ours	VGG16 → VGG11 (KD)	33.16	0.76 **(+0.13)**	0.85 **(+0.51)**	74.95 **(+3.33)**
	ProtopNet	VGG19 (Teacher)	29.10	1.0	1.0	77.97
	ProtopNet	VGG16 (Student)	32.80	0.57	0.39	76.35
	Ours	VGG19 → VGG16 (KD)	29.22	0.77 **(+0.20)**	0.88 **(+0.49)**	77.33 **(+0.98)**

Table S1. Results of Proto2Proto student (Ours) on ProtoPNet [1] for VGG architecture on CUB. Evaluated performance using Top-1 Accuracy and interpretability using metrics AAP, AJS and PMS

each prototype. For e.g., if number of test images is in the range of 10000 and if 1000 active patches of two prototypes match. JS will be less despite 1000 active patches matching. Hence, we introduce Modified Jaccard Similarity (MJS). The MJS between sets A and B is given by,

$$\text{MJS(A, B, } \alpha, m) = \frac{min(|A \cap B|, \alpha m)}{min(|A \cup B|, \alpha m)}, \qquad (2)$$

where $0.0 \leq \alpha \leq 1.0$ and $m = max(|A|, |B|)$. We take $\alpha = (0.1, 0.2, 0.3, ...1.0)$ and average the results.

D. Few-shot Experiments

ProtoNet [2] is a widely used method for Few-shot recognition. They learn class wise prototypes unlike ProtoTree and ProtoPNet. Our method can be easily adapted for ProtoNet. We follow the implementation of FRN [3] for ProtoNet. The Teacher model is ResNet-12 and Student model is Conv-4. Table S2 summarize the results. As evident, our method improves the performance of the student model signficantly.

	CUB	
Method	1-shot	5-shot
ProtoNet (T)	79.09	90.59
ProtoNet (S)	66.98	86.12
Ours (T → S)	72.37 **(+5.39)**	88.51 **(+2.39)**

Table S2. Results on Few-shot Recognition

E. Hyperparameter Details

We follow the same hyperparemeters as ProtoPNet and ProtoTree. We set $\lambda_{global} = 10$ and $\lambda_{ppc} = 10$. For training $\tau_{train} = 100$ and for testing, we choose τ_{test} from the set $\{0.1, 0.45, 1, 5, inf\}$. We found setting τ as high for training and low testing to be optimal. For VGG architectures the dimensions of the prototype is set to $d = 128$ whereas for ResNet architectures it is set to $d = 256$.

F. Visualizations

We visualize the top-k prototypes of a test image for ProtoPNet in Figure S2 (similar to Figure 1. in the main draft). As observed, the prototypes of Proto2Proto student are very similar to that of teacher's and the prototypes of baseline student are very different from that of teacher's. Also, the prototypes of baseline student seem to be less relevant compared to that of teacher's. For Prototree, we visualize (in Figure S3 and Figure S4) the subtree of teacher, baseline student and our student. The non-leaf nodes of the subtree represent the prototypes and leaf nodes represent the distribution of the classes. Note that, at the top of each non-leaf node, the prototype numbers are also mentioned for a fair comparison. In Figure S3, the nodes 2179, 2180 and 2243 of our student are identical to that of teacher's. The node 2181 is focusing on the same car but different parts of it. The leaf distribution of our student is same as that of teacher's. The baseline subtree, however, is completely different. Similar inference can be made from Figure S4. We can conclude that the decision process of our student is very similar to that of teacher's.

References

[1] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: Deep learning for interpretable image recognition. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 2

[2] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 2

[3] Davis Wertheimer, Luming Tang, and Bharath Hariharan. Few-shot classification with feature map reconstruction networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8012–8021, 2021. 2

Figure S2. Comparison of sample prototypes of test image between Teacher, Baseline Student and Proto2Proto (P2P) Student.

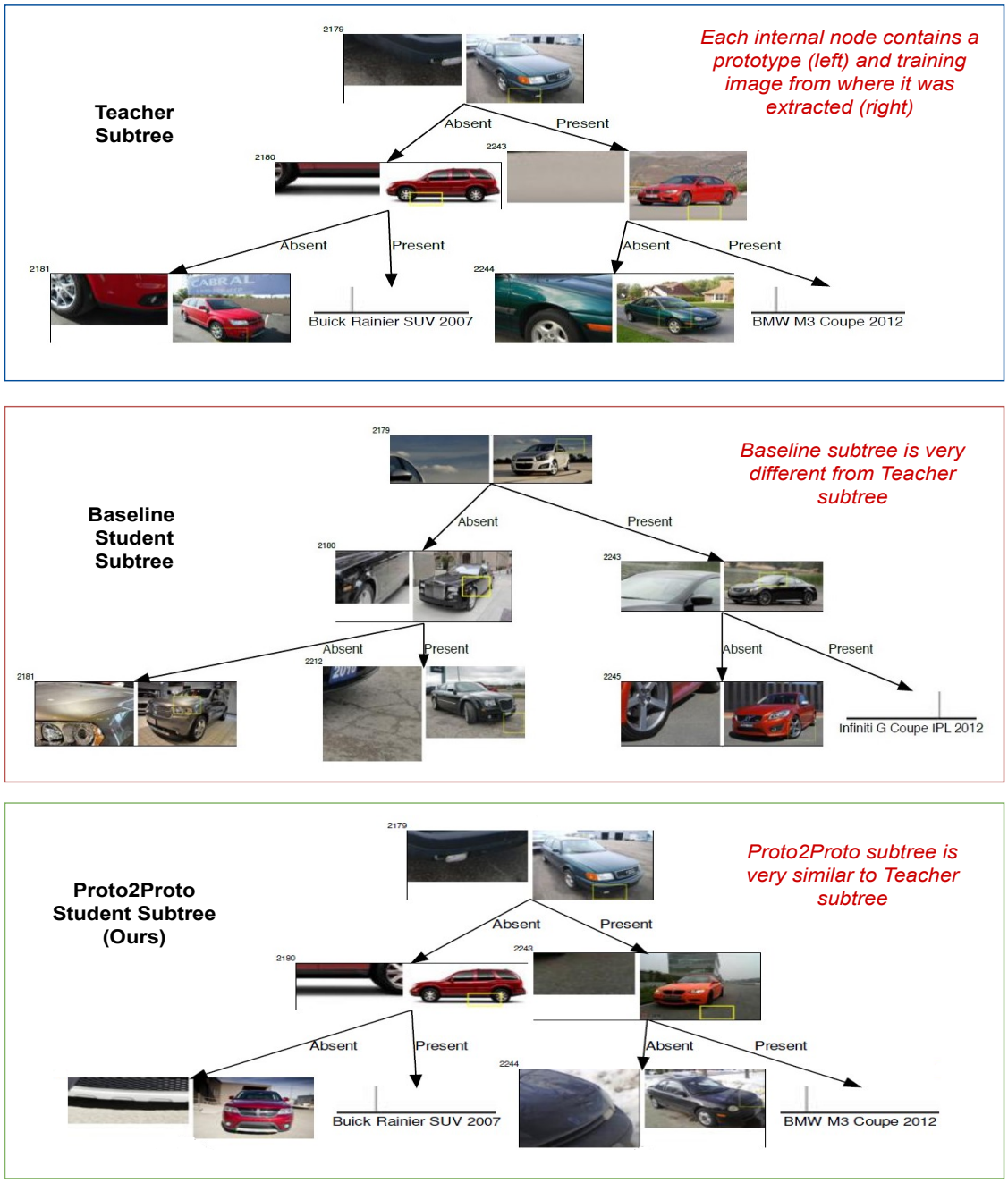


Figure S3. Comparison of subtrees between Teacher, Baseline Student and Proto2Proto (P2P) Student.

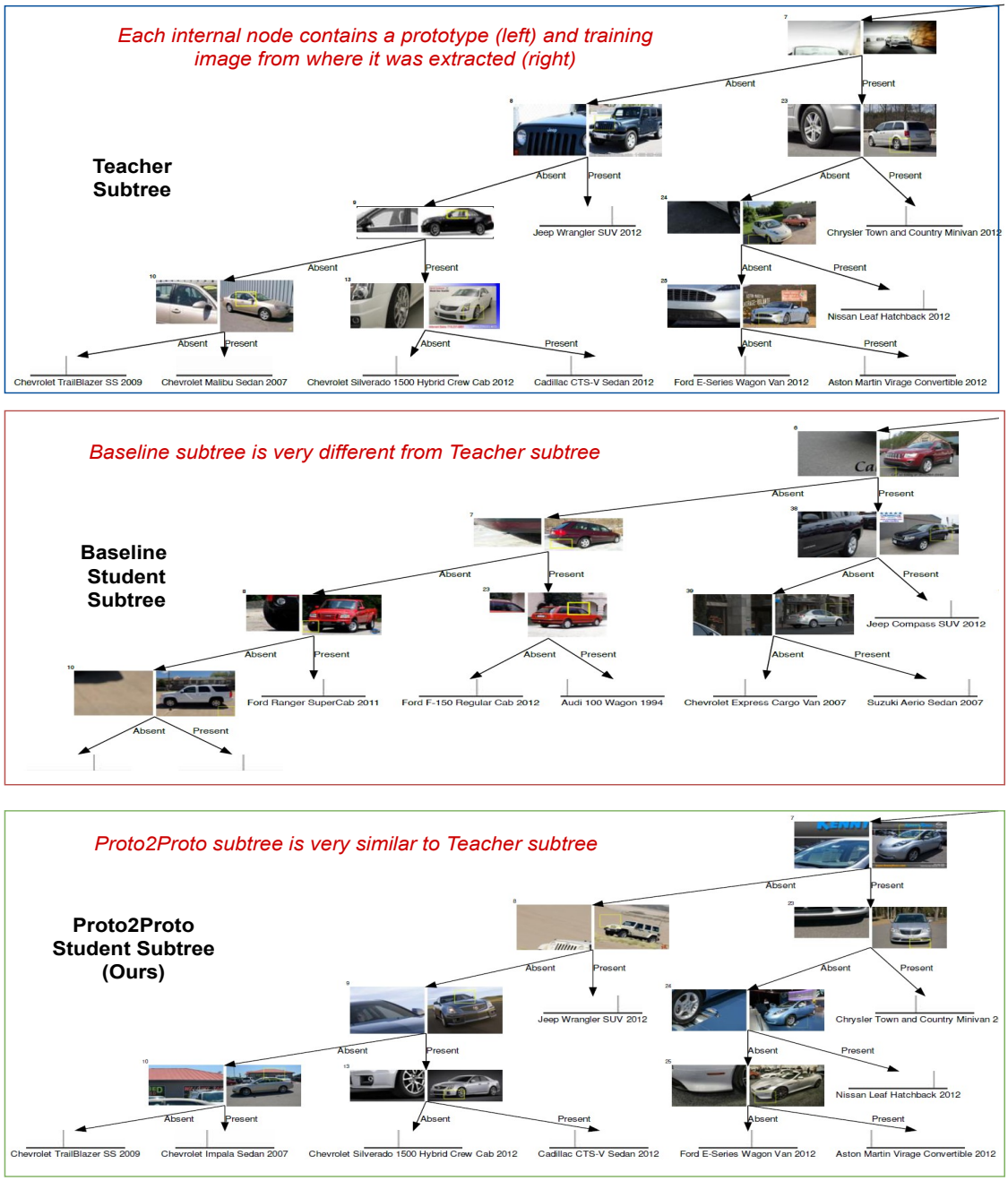


Figure S4. Comparison of subtrees between Teacher, Baseline Student and Proto2Proto (P2P) Student.