Relational Embedding for Few-Shot Classification
- *Supplementary Materials* -

In this supplementary materials, we provide additional details and results of our method.

Contents

1. Alternative derivation of relational embedding

Equations (4), (5), and (6) in the main paper describe the process of deriving relational embeddings, $\mathbf{q}$ and $\mathbf{s} \in \mathbb{R}^C$, using pre-computed co-attention maps, $\mathbf{A}_\text{q}$ and $\mathbf{A}_\text{s} \in \mathbb{R}^{H\times W}$, where the attention maps themselves provide interpretable visualization, *e.g.*, Fig. 1(c) in the main paper. In this section, we derive $\mathbf{q}$ and $\mathbf{s}$ in an alternative way of not explicitly introducing the attention maps, $\mathbf{A}_\text{q}$ and $\mathbf{A}_\text{s}$, but *multiplying a feature map by cross-correlation*, which is used in spatial attention work [3, 14, 16]. Let us denote the normalized cross-correlation tensor in Eq. (4) by

$$\tilde{\mathbf{C}} = \frac{\exp\left(\hat{\mathbf{C}}(\mathbf{x}_\text{q}, \mathbf{x}_\text{s})/\gamma\right)}{\sum_{\mathbf{x}'_\text{q}} \exp\left(\hat{\mathbf{C}}(\mathbf{x}'_\text{q}, \mathbf{x}_\text{s})/\gamma\right)} \tag{s.1}$$

and reshape it to a 2D matrix: $\tilde{\mathbf{C}} \in \mathbb{R}^{HW\times HW}$.

The relational embedding $\mathbf{q}$ is equivalently derived by multiplying two matrices $\tilde{\mathbf{C}}^\top$ and $\mathbf{F}_\text{q} \in \mathbb{R}^{HW\times C}$ followed by average pooling:

$$\begin{aligned}
\mathbf{q} &= \sum_{\mathbf{x}_\text{q}} \left(\underbrace{\left(\frac{1}{HW} \sum_{\mathbf{x}_\text{s}} \tilde{\mathbf{C}}(\mathbf{x}_\text{q}, \mathbf{x}_\text{s}) \right)}_{\text{Eq. (4)}} \mathbf{F}_\text{q}(\mathbf{x}_\text{q}) \right) \quad \text{(Eq. (5))} \\
&= \frac{1}{HW} \sum_{\mathbf{x}_\text{q}} \left(\sum_{\mathbf{x}_\text{s}} \tilde{\mathbf{C}}(\mathbf{x}_\text{q}, \mathbf{x}_\text{s}) \right) \mathbf{F}_\text{q}(\mathbf{x}_\text{q}) \\
&= \frac{1}{HW} \sum_{\mathbf{x}_\text{s}} \sum_{\mathbf{x}_\text{q}} \tilde{\mathbf{C}}(\mathbf{x}_\text{q}, \mathbf{x}_\text{s}) \mathbf{F}_\text{q}(\mathbf{x}_\text{q}) \\
&= \frac{1}{HW} \sum_{\mathbf{x}_\text{s}} \underbrace{\sum_{\mathbf{x}_\text{q}} \tilde{\mathbf{C}}^\top(\mathbf{x}_\text{s}, \mathbf{x}_\text{q}) \mathbf{F}_\text{q}(\mathbf{x}_\text{q})}_{\text{matrix multiplication}} \\
&= \frac{1}{HW} \sum_{\mathbf{x}_\text{s}} \underbrace{\tilde{\mathbf{C}}^\top \mathbf{F}_\text{q}}_{\mathbb{R}^{HW\times C}}(\mathbf{x}_\text{s}).
\end{aligned} \tag{s.2}$$

Here, $\tilde{\mathbf{C}}^\top \mathbf{F}_\text{q}$ is considered as softly-aligning the query feature map $\mathbf{F}_\text{q}$ in the light of each position of the support using the cross-correlation $\tilde{\mathbf{C}}^\top$.

Likewise, the relational embedding $\mathbf{s}$ is computed as

$$\mathbf{s} = \frac{1}{HW} \sum_{\mathbf{x}_\text{q}} \tilde{\mathbf{C}} \mathbf{F}_\text{s}(\mathbf{x}_\text{q}). \tag{s.3}$$

2. Comprehensive details on implementation

For training, we use an SGD optimizer with a momentum of 0.9 and a learning rate of 0.1. We train 1-shot models for 80 epochs and decay the learning rate by a factor of 0.05 at each $\{60, 70\}$ epoch. To train 5-shot models, we run 60 epochs and decay the learning rate at each $\{40, 50\}$ epoch. We randomly construct a training batch of size 128 for the ImageNet family [11, 19] and 64 for CUB [20] & CIFAR-FS [1] to compute $\mathcal{L}_\text{anchor}$. This objective is jointly optimized from scratch with $\mathcal{L}_\text{metric}$ and $\mathcal{L}_\text{anchor}$ as described in Sec. 4.4. For a fair comparison, we adopt the same image sizes, the backbone network, the data augmentation techniques, and the embedding normalization following the recent work of [23, 24].

self-correlation computation	category of neighbors	*mini* ImgNet	CUB
✗ (GAP baseline)	✗	65.33	77.54
$\mathbf{R} \in \mathbb{R}^{H\times W\times H\times W\times C}$	absolute	66.41	76.34
$\mathbf{R} \in \mathbb{R}^{H\times W\times U\times V\times C}$	relative	**66.66**	**78.69**

Table s.1: Comparison between absolute and relative neighborhood space in computing the self-correlation tensor $\mathbf{R}$.

4D convolution kernels	*mini* ImgNet	CUB	GPU time (*ms*)
✗ (GAP baseline)	65.33	77.54	27.74
vanilla 4D [12]	65.59	78.89	60.35
separable 4D [22]	**65.90**	**78.49**	34.97

Table s.2: Comparison between 4D convolutions for $h(\cdot)$.

3. Ablation studies

We provide more ablation studies on CUB [20] and *mini*ImageNet [19] in the 5-way 1-shot setting.

3.1. Self-correlation computation with relative *vs.* absolute neighbors

We validate the importance of *relative* neighborhood correlations of a self-correlation tensor $\mathbf{R}$ in Table s.1. We set $H = W = U = V$ such that the two models have the same input sizes for a fair comparison. The results show the superiority of the relative neighborhood correlation. An advantage of the relative correlation over the absolute one is that relative correlations provide a translation-invariant neighborhood space. For example, let us consider a self-correlation between a reference position $\mathbf{x}$ and its neighbors. While an absolute correlation $(\mathbf{x}, \mathbf{x}') \in \mathbb{R}^{H\times W\times H\times W}$ provides a variable neighborhood space as $\mathbf{x}$ translates by $\mathbf{t}$: $(\mathbf{x}+\mathbf{t}, \underline{\mathbf{x}'+\mathbf{t}})$, a relative correlation $(\mathbf{x}, \mathbf{p}) \in \mathbb{R}^{H\times W\times U\times V}$ provides a consistent view of the neighborhood space no matter how $\mathbf{x}$ moves: $(\mathbf{x}+\mathbf{t}, \underline{\mathbf{p}})$.

3.2. Separable *vs.* vanilla 4D convolution on CCA

Comparison between the original vanilla 4D convolutional kernels [12] and separable 4D kernels [22] is summarized in Table s.2, where we adopt the separable one for its efficiency. Note that the separable 4D kernels approximate the vanilla $3\times3\times3\times3$ kernels by two sequential $3\times3\times1\times1$ and $1\times1\times3\times3$ kernels followed by a point-wise convolution. The reported GPU time in Table s.2 is an average time for processing an episode and is measured using a CUDA event wrapper in PyTorch [10]. While the two kinds of kernels closely compete with each other in terms of accuracy, the separable one consumes less computational costs.

method	5-way 1-shot accuracy (%)	# add. params
CAN [5]	63.85 ± 0.48	0.3K
RENet (ours)	$\mathbf{67.60 \pm 0.44}$	203.2K
LEO [13]	61.76 ± 0.08	248.8K
CTM [7]	64.12 ± 0.82	305.8K
FEAT [23]	66.78 ± 0.20	1640.3K
MTL [17]	61.20 ± 1.80	4301.1K
wDAE [4]	61.07 ± 0.15	11273.2K

Table s.3: Performance comparison in terms of model size and accuracy (%) on *mini*ImageNet.

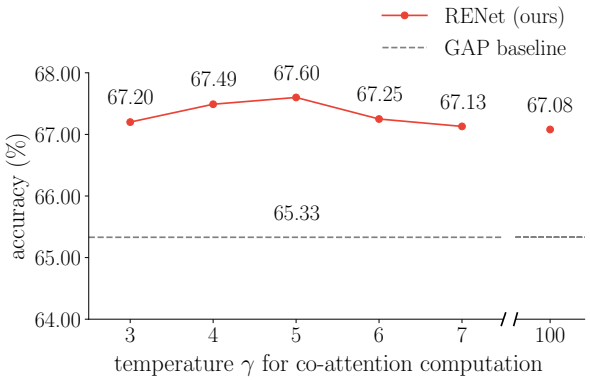

Figure s.1: Accuracy (%) of varying γ on *mini*ImageNet.

3.3. Number of parameters

We measure the number of additional model parameters of recent methods and compare them with RENet in Table s.3. Table s.3 studies the effect of *additional* parameters only so we collect publicly available codes of methods that use additional parameterized modules [4, 5, 7, 13, 17, 23], and intentionally omit [2, 6, 8, 9, 15, 18, 21, 24] as their trainable parameters are either in the backbone network or in the last fully-connected layer. Compared to the largest model [4], ours performs significantly better (67.60 *vs.* 61.07) while introducing 55 times less additional capacity (203.2K *vs.* 11.2M).

3.4. Temperature γ for co-attention computation

We investigate the impact of the hyper-parameter γ that controls the smoothness of the output attention map (Eq. (4)). As its name "temperature" suggests, a higher temperature outputs a smoother attention map, while a lower temperature outputs a peakier one. Figure s.1 shows that the temperature γ has a certain point that maximizes the accuracy by appropriately balancing the smoothness factor. Interestingly, an extremely high temperature $\gamma = 100$ degrades accuracy by making all attention scores evenly distributed. It is noteworthy that our full model RENet with a range of $\gamma \in \{3, 4, 5, 6, 7\}$ outperforms all existing methods on the dataset.

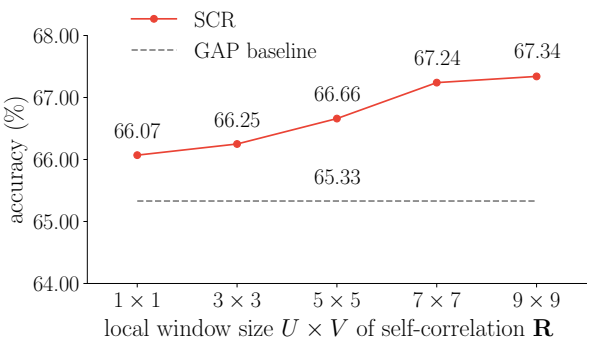

Figure s.2: Accuracy (%) of varying $U \times V$ on *mini*ImageNet.

3.5. Local window size UV for SCR

To evaluate the effectiveness of learning relational features from local neighborhood correlation, we vary the local window size UV of a self-correlation tensor $\mathbf{R} \in \mathbb{R}^{H \times W \times U \times V \times C}$. As shown in Fig s.2, the accuracy steadily increases as more neighborhood correlations are learned, which indicates that learning relational structures is favorable for few-shot recognition. Note that SCR with $U = V = 1$ already outperforms the GAP baseline, which is an effect of learning from $l2$-normalized features (Eq. 1). Despite the consistent accuracy gain from observing wide local window, we choose $U = V = 5$ for all experiments *to limit the space complexity* increased by a factor of UV.

4. Qualitative results

To demonstrate the effects of our method, we present additional qualitative results. All images are sampled from the *mini*ImageNet validation set in the 5-way 1-shot setting.

4.1. Effects of SCR

We ablate the SCR module and demonstrate the effects of SCR in Fig. s.3. The results show that "CCA w/ SCR" successfully attends to fine characteristics than "CCA w/o SCR" does, implying that the SCR module provides reliable representation for the subsequent CCA module.

4.2. Co-attention maps on multi-object queries

Given a multi-object image as a query, we examine if the object regions can be adaptively highlighted depending on the support semantics in Fig. s.4. The CCA module successfully captures query regions that are semantically related with each support image. This effect accords with the motivation of the CCA module, which is to adaptively provide "where to attend" between two image contexts.

4.3. Cross-correlation refinement via $h(\cdot)$

We demonstrate the effect of 4D convolutional block $h(\cdot)$ that filters out unreliable matches in the initial cross-correlation by analyzing neighborhood consensus patterns. We visualize the top 10 matches among $2HW$ matching candidates computed by argmax of matching scores from each side. As shown in Fig. s.5, the initial cross-correlation $\mathbf{C}$ exhibits many spurious matches misled by indistinguishable appearance, *e.g.*, matching two regions of the sky, whereas the updated cross-correlation $\hat{\mathbf{C}}$ shows reliable and meaningful matches, *e.g.*, matching two sails.

References

[1] Luca Bertinetto, Joao F Henriques, Philip Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. In *Proc. International Conference on Learning Representations (ICLR)*, 2018. 1

[2] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *International Conference on Learning Representations (ICLR)*, 2019. 2

[3] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[4] Spyros Gidaris and Nikos Komodakis. Generating classification weights with gnn denoising autoencoders for few-shot learning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[5] Ruibing Hou, Hong Chang, MA Bingpeng, Shiguang Shan, and Xilin Chen. Cross attention network for few-shot classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 2

[6] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[7] Hongyang Li, David Eigen, Samuel Dodge, Matthew Zeiler, and Xiaogang Wang. Finding task-relevant features for few-shot learning by category traversal. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[8] Bin Liu, Yue Cao, Yutong Lin, Qi Li, Zheng Zhang, Mingsheng Long, and Han Hu. Negative margin matters: Understanding margin in few-shot classification. In *Proc. European Conference on Computer Vision (ECCV)*, 2020. 2

[9] Puneet Mangla, Nupur Kumari, Abhishek Sinha, Mayank Singh, Balaji Krishnamurthy, and Vineeth N Balasubramanian. Charting the right manifold: Manifold mixup for few-shot learning. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020. 2

[10] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *Advances in Neural Information Processing Systems (NeurIPS) Workshop Autodiff*, 2017. 2

[11] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, Hugo Larochelle, and Richard S Zemel. Meta-learning for semi-supervised few-shot classification. In *Proc. International Conference on Learning Representations (ICLR)*, 2018. 1

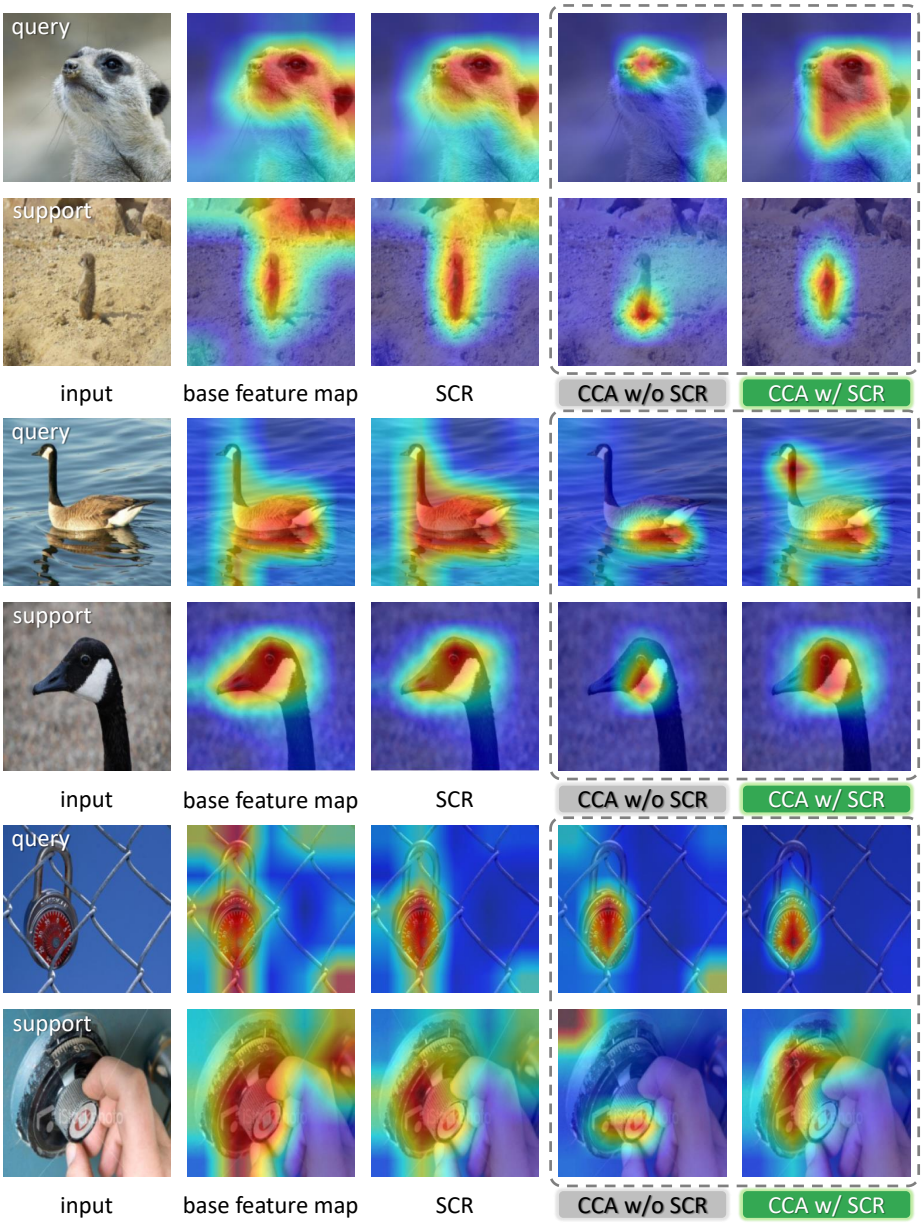

Figure s.3: Effects of SCR on *mini*ImageNet. "CCA w/ SCR" captures fine details between two images while "CCA w/o SCR" often fails. The "base feature map" and "SCR" columns visualize average channel activations. The "CCA w/ SCR" and "CCA w/o SCR" columns visualize co-attention maps.

[12] Ignacio Rocco, Mircea Cimpoi, Relja Arandjelović, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Neighbourhood consensus networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 2

[13] Andrei A Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. In *Proc. International Conference on Learning Representations (ICLR)*, 2018. 2

[14] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[15] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2

[16] Guolei Sun, Wenguan Wang, Jifeng Dai, and Luc Van Gool. Mining cross-image semantics for weakly supervised semantic segmentation. In *Proc. European Conference on Computer Vision (ECCV)*, 2020. 1

[17] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele.

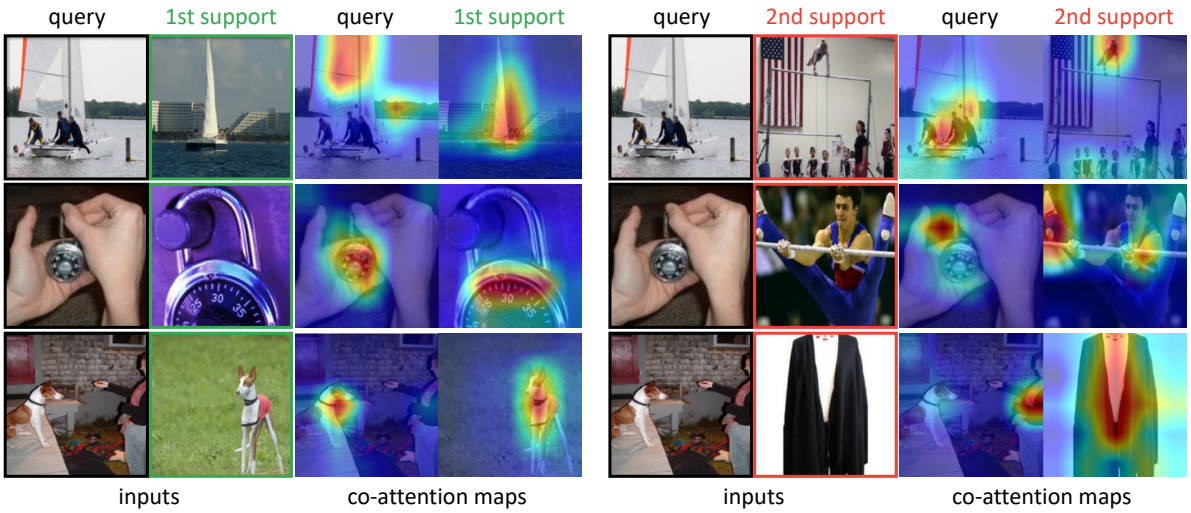

Figure s.4: Co-attention maps on multi-object queries on *mini*ImageNet. The proposed CCA module can adaptively capture multiple objects in a query depending on the context of each support instance.

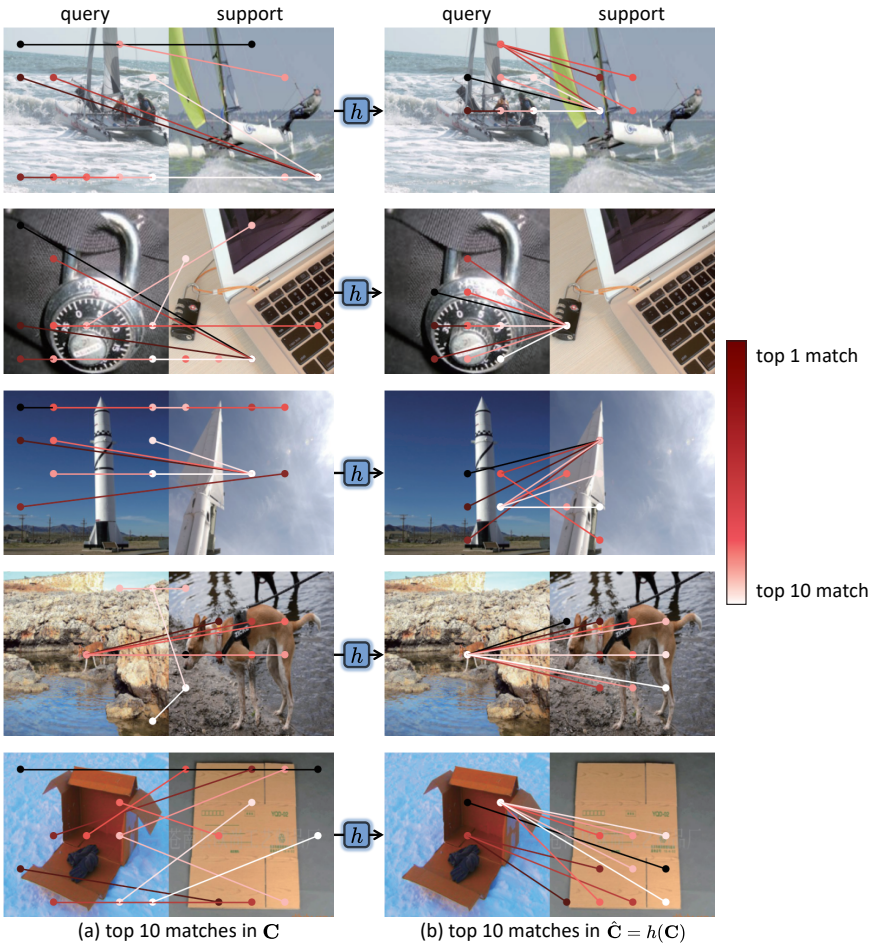

Figure s.5: Visualization of cross-correlation on *mini*ImageNet. **(a)**: Top 10 matches in $\mathbf{C}$ (initial cross-correlation). **(b)**: Top 10 matches in $\hat{\mathbf{C}} = h(\mathbf{C})$ (refined cross-correlation). Unreliable matches are filtered through $h(\cdot)$.

Meta-transfer learning for few-shot learning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[18] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? In *Proc. European Conference on Computer Vision (ECCV)*, 2020. 2

[19] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, and Daan Wierstra. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. 1, 2

[20] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. Technical report, 2011. 1, 2

[21] Yan Wang, Wei-Lun Chao, Kilian Q Weinberger, and Laurens van der Maaten. Simpleshot: Revisiting nearest-neighbor classification for few-shot learning. *arXiv preprint arXiv:1911.04623*, 2019. 2

[22] Gengshan Yang and Deva Ramanan. Volumetric correspondence networks for optical flow. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 2

[23] Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, and Fei Sha. Few-shot learning via embedding adaptation with set-to-set functions. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2

[24] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2