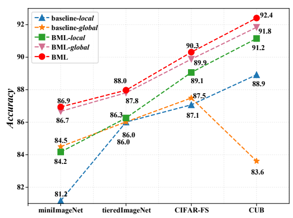
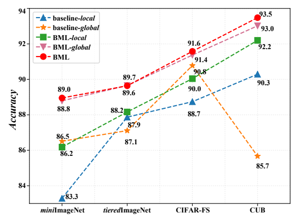
S-1. More Experimental Results

S-1.1. More Shots

As the number of visible samples (support shots) increases (Figure S-1), performance gradually improves, and BML is steadily higher than the two single-view baselines. Besides, the performance of BML-*global* and BML-*local* under the binocular mode is superior to baseline-*global* and baseline-*local* under single view mode.



(a) 10-shot



(b) 20-shot

Figure S-1. Comparison between BML and the two baselines with more shots.

S-1.2. More Benchmarks

To further verify the performance of BML, we do experiments on another public few-shot classification benchmark: FC100. FC100 is derived from CIFAR-100, which has a total of 100 classes. Among them, 60 classes are used for training, 20 are used for verification, and the remaining 20 are used for testing. Since the division is carried out at the superclass level, the information overlap between splits is minimized, thus more challenging.

As is shown in Table S-1, on FC100, BML is still superior to the two single-view baselines, and stays ahead of the other six competitors. Specifically, three key points are conveyed which have been emphasized in Section 4:

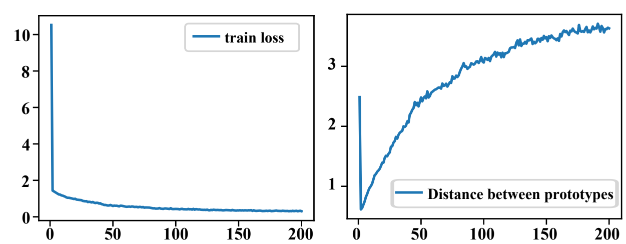
- Binocular learning is better than single-view mode. BML is 2% higher than the single global view and 5%-9% higher than the single local view.
- On coarse-grained dataset, global view performs better than local view.
- The two complementary views can promote each other (*i.e.*, BML-*global* vs. baseline-*global*, BML-*local* vs. baseline-*local*), and the global impact on the local view is more obvious.

S-1.3. More Analyze of elastic loss

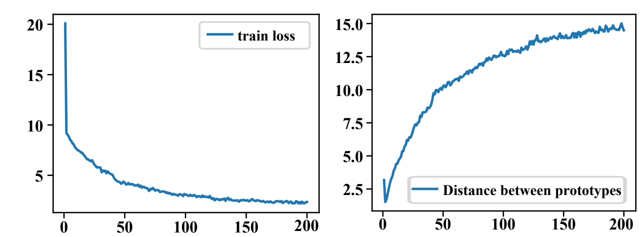
We carefully monitor the elastic loss and further explore its mechanism. Figure S-2 shows the trend of *training loss* and *distance between prototypes* with or without elastic loss (On *mini*ImageNet). Obviously, comparing the left subfigure of Figure 2(a) with the one of Figure 2(b), we can find that when no elastic loss is applied, the loss value quickly drops to a low point, and the subsequent decline has been very slow. On the contrary, after applying the elastic loss, the initial loss value is increased significantly, and the downward trend is more obvious. This shows that elastic loss does increase the difficulty of optimization. Furthermore, as shown in right subfigures of Figure 2(a) and Figure 2(b), the distance between N prototypes (first-order moment) shows a similar change. With the help of elastic loss, the distance between prototypes is gradually expanded, and the features are more dispersed in the embedding space. This shows that the network is learning to amplify the difference between prototypes to improve matching accuracy.

Table S-1. Comparison on FC100.

Method	FC100	
	5-way 1-shot	*5-way 5-shot*
MAML	38.10±1.70	50.40±1.00
MetaOptNet	41.10±0.60	55.50±0.60
ProtoNet	35.30±0.60	48.60±0.60
TADAM	40.10±0.40	56.10±0.40
Rethink	42.60±0.70	59.10±0.60
DC	42.04±0.17	57.05±0.16
Baseline-*local*	38.88±0.38	54.25±0.40
Baseline-*global*	42.61±0.39	61.03±0.40
BML-*local*	43.25±0.41	58.70±0.39
BML-*global*	43.88±0.40	62.06±0.39
BML	45.00±0.41	63.03±0.41



(a) *w/o* elastic loss



(b) *w/* elastic loss

Figure S-2. Loss value and mean distance between prototypes on the base set.

S-1.4. More Visualization of tasks

We randomly visualize two tasks in Figure S-3, from left to right, they represent BML, baseline-*global* and baseline-*local*. Obviously, the prototypes (highlight with *star*) computed by BML are more dispersed in the embedding space, which proves that BML helps to obtain more discriminative features.

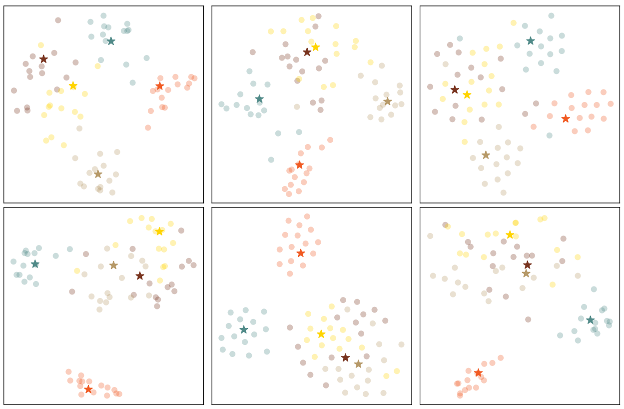


Figure S-3. t-SNE visualization results on two tasks.

S-1.5. More analyze of mutual interaction

In order to further verify the influence of mutual interaction on performance, we design a series of ablation experiments, including the impact of the number of shared blocks and the influence of mutual interaction. Here is the result (S:share, I:independent).

Table S-2. Analysis of the number of shared blocks.

Methods	Accuracy	*Params.*
(a) Ensemble	81.08 ± 0.31	24,930,688
(b) BML$(S^0 I^4)$	83.10 ± 0.30	24,930,688
(c) BML$(S^1 I^3)$	83.24 ± 0.30	24,813,504
(d) BML$(S^2 I^2)$	83.30 ± 0.29	24,249,024
(e) BML$(S^3 I^1)$	$\mathbf{83.63 \pm 0.29}$	**21,891,264**

According to the results shown in Table S-2, comparing (a) and (b), the simple integration of baseline-*global* and baseline-*local* without interactive learning has almost no benefit since the difference between two models is relatively large (see in Table 2), while (b) still has good performance, it is mutual interactive learning ensures that the features of the two branches have similarities while maintaining appropriate differences. Comparing (b)-(e), the performance of BML changes relatively gently, which shown the main factor that affects the performance is whether performing binocular mutual learning. To reduce the amount of parameters, BML only separates the last block.

S-2. Efficient Implementation of BML

To fully unleash the power of binocular framework, during training, we adopt uniform sampling strategy. Specifically, a batch contains $N = 15$ randomly sampled classes.