Supplementary Material

A. Additional Experiments and Analysis

A.1. Activation Function

We use ReLU as the activation function to achieve nonlinearities for each expert module, which is a key part of achieving a more uniform embedding space. We also tried some other activation functions including Sigmoid and Tanh, but they cannot achieve as good performance on Au as ReLU. Table 4 and Figure 7 show the comparison results of CRnet with ReLU (abbreviated as ReLU) and CRnet with Sigmoid (abbreviated as Sigmoid) on each dataset.

As shown in Table 4, As of CRnet with Sigmoid on each dataset increases compared with CRnet with ReLU (increases by 12.3%, 10.8%, 10.7%, 5.8%, 16.9% respectively), whereas its Au decreases a lot (decreases by 13.3%, 12.3%, 12.2%, 5.7%, 14.0% respectively), resulting in a lower H (decreases by 6.3, 7.5, 5.9, 1.3, 13.7 respectively).

It turns out that ReLU is the most suitable activation function for expert module. ReLU is actually a piecewise linear function of 2 pieces and what the cooperation module finally learns is a piecewise linear function of $K+1$ pieces for each dimension of the embedding space. Therefor ReLU helps to achieve the highest local linearity (every piece is linear), whereas other activation functions like Sigmoid reduce local linearity as shown in Figure 6 and lead to a decrease on Au. The experimental results prove the point that high local linearity causes less bias problem as clamied in Section 5.2.

Table 4. Classification accuracies on various datasets of CRnet with different activation functions. As/Au: Average per-class top-1 accuracy in % on seen/unseen classes. H: Harmonic mean accuracy.

Dataset	Model	Accuracy As	Au	H
AwA1	ReLU	74.7	58.1	65.4
	Sigmoid	87.0	44.8	59.1
AwA2	ReLU	78.8	52.6	63.1
	Sigmoid	89.6	40.3	55.6
CUB	ReLU	56.8	45.5	50.5
	Sigmoid	67.5	33.3	44.6
SUN	ReLU	36.5	34.1	35.3
	Sigmoid	42.3	28.4	34.0
aPY	ReLU	68.4	32.4	44.0
	Sigmoid	85.3	18.4	30.3

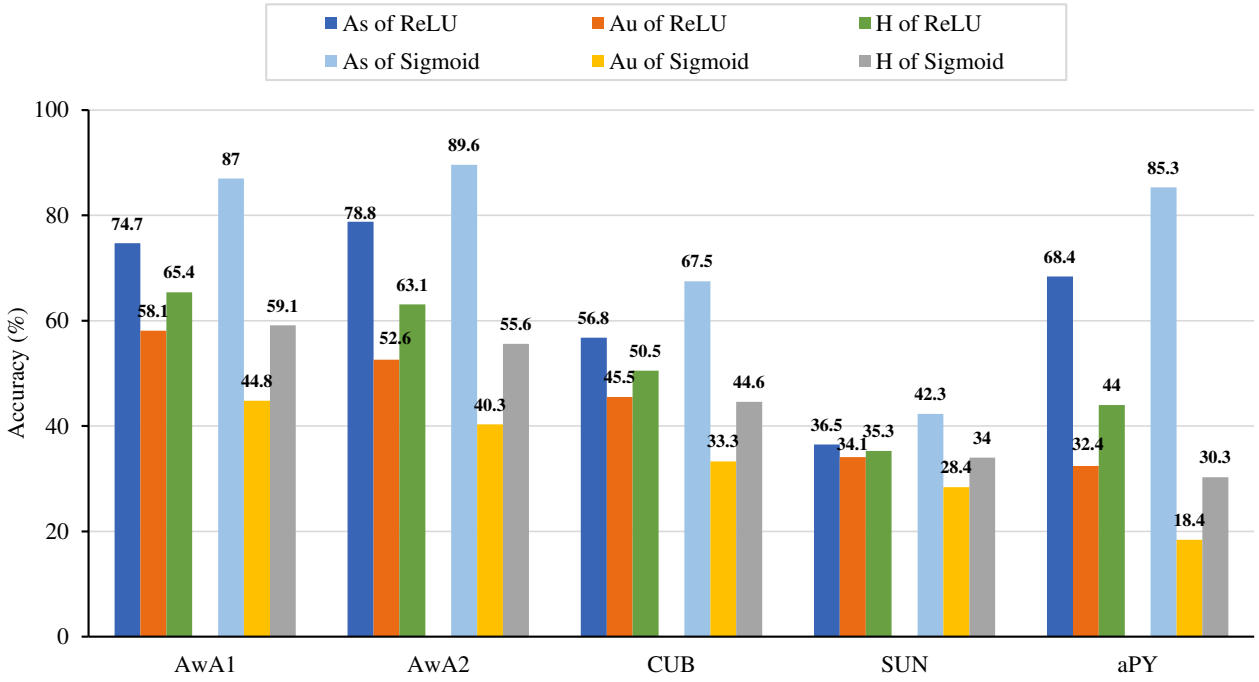

Figure 7. Bar chart of classification accuracies on various datasets of CRnet with different activation functions.

Table 5. Classification accuracies on various datasets at different K values. As/Au: Average per-class top-1 accuracy in % on seen/unseen classes. H: Harmonic mean accuracy.

	AwA1			AwA2			CUB			aPY				SUN		
K	**As**	**Au**	**H**	**As**	**Au**	**H**	**As**	**Au**	**H**	**As**	**Au**	**H**	***K***	**As**	**Au**	**H**
1	76.3	52.0	61.8	**87.5**	46.9	60.9	65.3	33.8	44.5	**82.8**	24.7	38.0	**3**	**42.1**	22.2	29.1
2	77.1	55.0	64.2	82.6	49.9	62.2	**66.5**	35.4	46.2	78.1	30.3	43.7	**6**	37.0	33.1	34.9
3	74.7	58.1	65.4	78.8	**52.6**	63.1	58.8	43.5	50	68.4	32.4	44.0	**9**	38.1	32.1	34.8
4	78.7	55.5	65.1	83.9	51.5	**63.9**	56.8	45.5	**50.5**	67.9	32.2	43.7	**12**	36.5	34.1	**35.3**
5	76.3	**59.4**	**66.8**	84.2	50.1	62.8	55.8	44.3	49.4	69.8	32.7	44.5	**15**	34.2	34.7	34.4
6	75.3	57.6	65.3	85.0	48.0	61.3	54.7	**46.2**	50.1	73.2	31.1	43.7	**18**	31.6	**38.5**	34.7
9	75.4	57.6	65.3	84.4	48.5	61.6	51.1	44.4	47.5	67.2	**33.4**	**44.6**	**21**	34.3	35.6	34.9
12	**79.7**	54.5	64.7	84.9	49.8	62.8	51.3	42.4	46.4	67.1	32.0	43.8	**24**	35.9	32.3	34.0

A.2. Hyperparameter Analysis

In our method K is an important hyperparameter that reprents the number of expert modules as well as the number of clusters for semantic vectors. We explore the impact of different K values on network performance on all five datasets to study the robustness of CRnet with respect to K.

For AwA1, AwA2, CUB and aPY, we test the accuracy of CRnet when $K = 1, 2, 3, 4, 5, 6, 9, 12$ respectively. And for SUN, whose number of seen classes reaches 645, we test the accuracy of CRnet when $K = 3, 6, 9, 12, 15, 18, 21, 24$ respectively. The results are shown in Tabel 5 and their corresponding curves are shown in Figure 8.

The results show that the hyperparameter K have similar properties with the convolution kernel number in a CNN: When K is small, the model is more likely to achieve a high As but a low Au, and the overall performance is gradually improved with the increase of K until the model is sufficiently representational; Then the accuracy fluctuates within an acceptable range as K continues to increase; If K is too large, performance drops slightly mainly because of redundancy and the model takes a long time to converge. When $K = 1$, the model degenerates into a network with a single expert module, i.e. a traditional GZSL network. When $K > 1$, the obvious improvement of performance again indicates the effectiveness of the cooperation module in solving GZSL problems.

It should be pointed out that the K value given in the main paper follows the principle that the total parameter amount of the model is as small as possible. And the accuracy (both Au and As) of the model trained by the same parameters has a fluctuation of about $\pm 1\%$ because of the randomness of K-means clustering algorithm and the random initialization parameters of the model, which is a common phenomenon in GZSL problems. We chose a more conservative accuracy as the final result for stability considerations, which means that the accuracy we show is lower than the average accuracy of the model and the algorithm has room for further improvement.

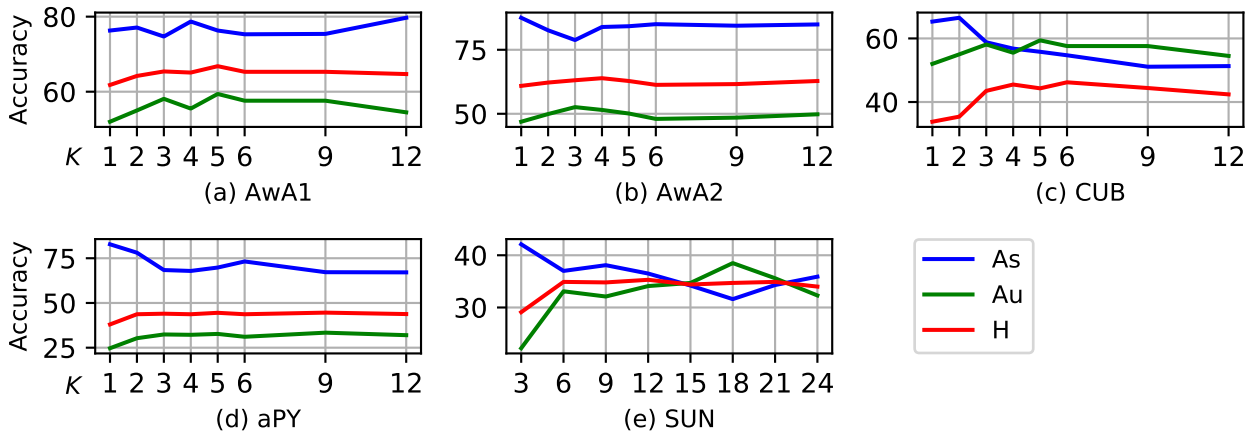

Figure 8. Accuracy(%)-K curves of various datasets.

A.3. Detailed Comparison of Per-class Performance

To further analyze the effectiveness of CRnet in alleviating the bias problem, we compare the per-class performance of CRnet and RN (relation network) on AwA2. For each unseen class, we calculate three indicators including the rate of misclassification into the closest seen class (referred as Bias Rate), per-class classification Error Rate, and LRD, as shown in Table 6. Figure 9 and Figure 10 are its corresponding bar charts.

Here CRnet's LRD is calculated in a slightly different way from the one given in the main paper. CRnet changes the distance between feature anchors to some extent, which may make the closest seen classes for RN no longer the closest ones for CRnet. In this experiment, the closest seen classes of 5 out of 10 unseen classes have changed (Class 1,6,7,8 and 10). Therefore, when calculating CRnet's LRD for these unseen classes, their cloest seen classes follow those in RN. Otherwise, the comparison of Bias Rates is meaningless.

Table 6. Per-class performance of RN and CRnet on AwA2. Index: Unseen class index, in descending order of bias rate; Bias Rate: The rate in % of misclassification into the closest seen class; Error Rate: Per-class classification Error Rate in %.

Index	Model	Bias Rate	Error Rate	LRD
1	RN	57.5	61.5	0.32
	CRnet	2.9	12.6	0.37
2	RN	48.6	81.5	0.89
	CRnet	5.8	59.5	0.87
3	RN	21.1	82.1	0.68
	CRnet	0.1	46.3	0.79
4	RN	14.8	89.7	0.67
	CRnet	2.3	96.2	0.72
5	RN	14.5	64.2	0.78
	CRnet	4.5	59.3	0.79
6	RN	12.3	44.9	0.83
	CRnet	0.8	17.6	1.38
7	RN	11.6	32.3	0.73
	CRnet	4.3	13.8	0.89
8	RN	1.3	46.7	0.88
	CRnet	0.2	39.3	1.88
9	RN	0.5	60.9	0.74
	CRnet	0	22.3	0.76
10	RN	0	97.9	1.02
	CRnet	0	78.9	0.82
Avg	RN	18.2	66.2	0.76
	CRnet	2.1	44.6	0.93

Obviously, for RN, cases that misclassification into the closest seen class (Bias Rate) account for a large part of the total error cases (Error Rate), indicating that bias problem is one of the major causes for its poor performance on GZSL. While CRnet successfully alleviates bias problem and significantly reduces the Bias Rate, thereby reducing the per-class Error Rate. But there are also some failure cases, e.g. Class 4: Its Bias Rate decreases by 12.5% whereas Error Rate increases by 6.5%. This indicates that CRnet may also cause some new problems while solving bais problem. Table 6 also shows that in the same model, the per-class LRD and per-class Bias Rate of different unseen classes are uncorrelated. While in different model, a specific unseen class's Bias Rate is probably positively correlated whit its LRD. A larger LRD means a lower probability of the samples being misclassified into the closest seen class but does not guarantee a smaller Bias Rate. Because accuracy is determined by many factors rather than only the bias problem.

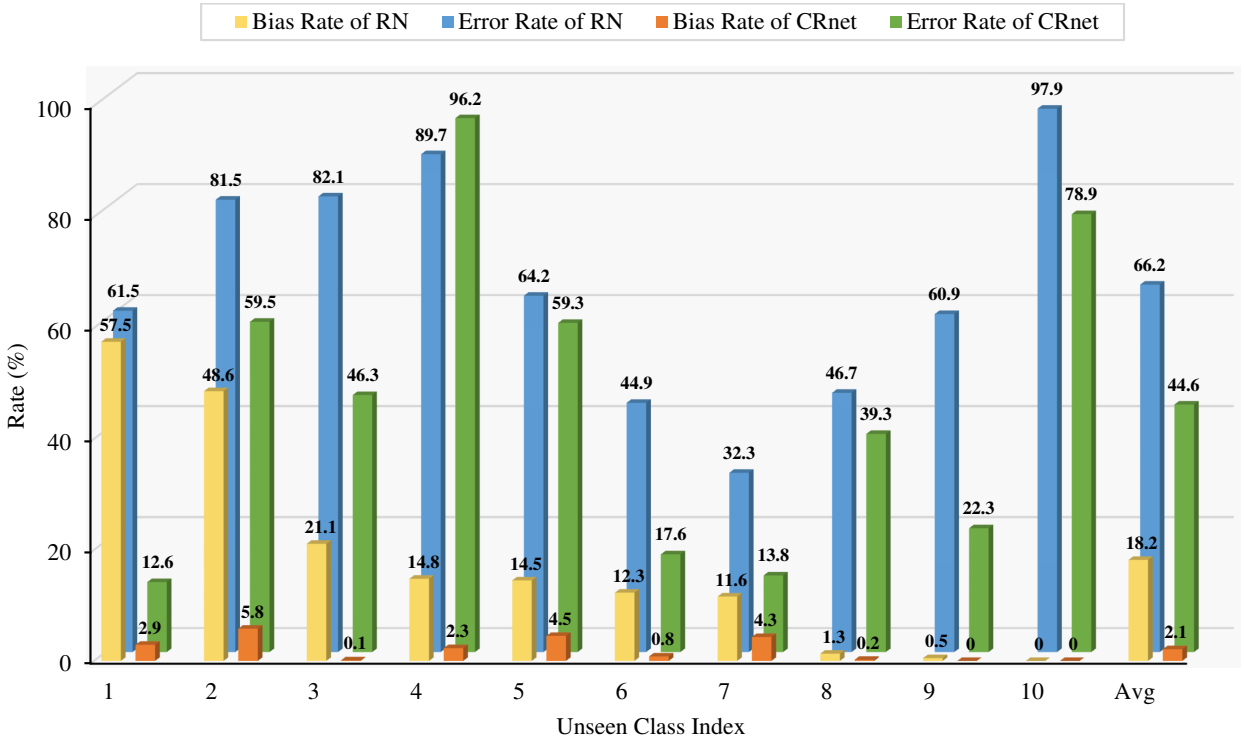

Figure 9. Bar chart of per-class Bias Rate and per-class Error Rate on AwA2.

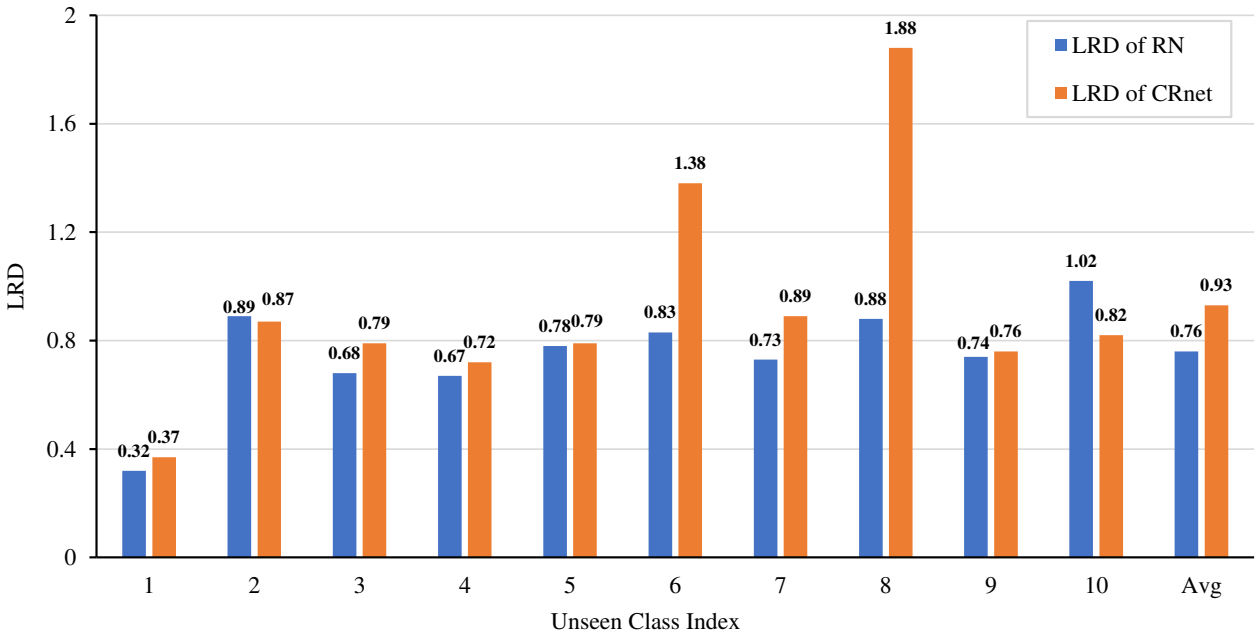

Figure 10. Bar chart of per-class LRD on AwA2.