

Figure 7: Semantic error visualization on the test set of SUN. (a) is the class-averaged semantic error matrix for the baseline, and (b) is the class-averaged semantic error matrix for CVsC.

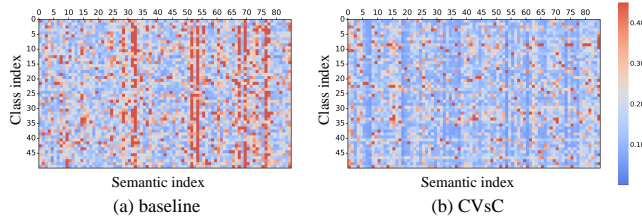


Figure 8: Semantic error visualization on the test set of AWA2. (a) is the class-averaged semantic error matrix for the baseline, and (b) is the class-averaged semantic error matrix for CVsC.

Table 6: Semantic error statistics on the test set of SUN and AWA2.

Method	SUN	AWA2
	average error	average error
baseline	0.2128	0.2208
CSvC	0.1312	0.1665

A COMPARISON OF DIFFERENT LOSS FUNCTIONS

We compared two classic loss functions used in ZSL, namely Mean Square Error (MSE) loss and Cross-Entropy (CE) loss, to establish visual-semantic correlation. ZSL allows for the direct use of attribute annotations as supervision, utilizing MSE loss to guide model training. Alternatively, it can also use the labels as supervision, which employs Cross-Entropy loss to optimize the model to learn visual-semantic correlation. Here, we take ViT as the backbone and use a fully connected neural network for semantic embedding as the baseline. Subsequently, we separately train the baseline using the two aforementioned loss functions. The results are presented in Table 5.

It shows that the performance of the CE loss significantly surpasses that of the MSE loss. This demonstrates our observation in the introduction, wherein directly enforcing alignment between mismatched vision and semantics leads to spurious visual-semantic correlation, resulting in poor zero-shot prediction performance. Although the CE loss utilizes strictly accurate image labels as supervision to enable the model to adaptively learn visual-semantic correlation, it lacks further effective semantic guidance and exploration of substantive visual-semantic correlation. This inspires us to propose CVsC, which establishes substantive visual-semantic correlation for ZSL.

Table 5: Comparison of different loss functions on CUB, SUN and AWA2.

Loss type	CUB		SUN		AWA2	
	GZSL	CZSL	GZSL	CZSL	GZSL	CZSL
	H	Acc	H	Acc	H	Acc
MSE	56.8	60.5	41.2	59.9	60.5	53.0
CE	67.3	72.6	52.5	72.9	72.0	65.5

B ADDITIONAL EXPERIMENTS FOR VERIFICATION OF SUBSTANTIVE CORRELATION

To provide a more comprehensive demonstration of the effectiveness of our CVsC in establishing substantive visual-semantic correlation, we conducted semantic error visualization experiments on additional datasets, namely SUN and AWA2. Figures 7 and 8 respectively visualize the semantic error matrices on the SUN and AWA2 datasets. Furthermore, statistical values of average semantic errors on these two datasets are presented in Table 6. It is evident that CVsC leads to a significant reduction in semantic errors. This finding indicates that CVsC achieves superior results on both benchmark datasets, highlighting its consistent ability to enhance the substantive correlation between vision and semantics for ZSL.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009