Appendix

1. Alignment Loss

The alignment loss encourages the model to associate the correct visual features with the corresponding textual features, which is critical for improving performance in zero-shot and generalized zero-shot learning tasks. Our proposed approach of aligned representation learning yields significant improvements in zero-shot learning and generalized zero-shot learning tasks. However, the scales of the alignment loss and classification loss values in our model are typically dissimilar. To reduce the overall magnitude of these losses, it becomes necessary to weight them appropriately.

By balancing these losses, we can also selectively emphasize one loss over the other to guide the model's focus towards specific learning objectives. This not only optimizes performance but also allows for greater interpretability of the model's learning process. Our study shows that this approach yields superior results compared to traditional methods in zero-shot and generalized zero-shot learning The weighted loss can be represented as L, whereas Lc is composed of L_{clip} and L_c. Mathematically we can write L as

$$L = \alpha \quad L_{clip} + \beta \quad L_c \tag{1}$$

Eq. 1 shows the loss formulation where L_{clip} is the CLIP alignment loss while L_c indicates the cross entropy classification loss. To optimize our approach for aligned representation learning, we experimented with different combinations of weights for the alignment loss L_{clip} and the classification loss L_c. After testing several options, we found that assigning a weight of α = 0.1 to L_{clip} and β = 0.9 to L_c yielded the best results. This combination brings the larger L_c loss closer to the smaller L_{clip} loss, allowing for better optimization of the function and providing the network with greater control over the loss regulation process. We also attempted to improve performance by utilizing distillation loss via logits matching. However, we found that this technique did not result in any significant improvements.

2. Pre-processing: Prompts design

After the ablation study in ZSL for one kind of prompt we also evaluated the results with a different prompt, further we applied and evaluated our approach in GZSL, where we add the seen and unseen classes together and try to predict the unseen classes out of them. All the previous settings of ZSL stayed the same for all the experiments to perform the GZSL multi-label classification. As we can see, we got an overall increase in MAP of 3.7% with good competitive F1 score values for k = 3 and k = 5, as given in Table 2. In GZSL, we train just like in ZSL. But, in the testing or evaluation phase, we test our model from a whole set of seen and unseen classes together to assess our model's ability to relate relevant images present in the combined set of seen and unseen classes with corresponding labels. We compare the variation in results by using different prompts, and the results of two prompts are given in Table 1. It indicates that using "a photo of [class],[class] and [class] gives us better results.

Figure 1. Examples of prompt templates for sample images: For both images, we use template, "a photo of [Class], [class] and [class]" and " a picture of [class], [class] and [class]". We fill these as, 1a: a photo of elephant, sky, and grass. 1b: b a photo of the book, man, and phone. Similarly for the second template as a picture of an elephant, sky, and grass

Task	Prompts	mAP	F1 k=3	F1 k=5
ZSL	A picture of [class], [class] and [class]	33.43	34.80	31.10
	A photo of [class], [class] and [class]	33.14	33.02	29.61
GZSL	A picture of [class], [class] and [class]	23.80	24.87	27.53
	A photo of [class], [class] and [class]	22.90	23.21	25.90

Table 1. Multilabel prompts: Performance comparison for ZSL and GZSL when used to benchmark NUS-WIDE dataset. The evaluation metric is in terms of mAP and F1 score for k 2 3; 5. We demonstrate that overall results may change with slight variation in context.