Appendix

In this appendix, Section A summarizes the architecture details of HSV A. Section B provides the key hyperparameter analyses and setting in our experiments.

A Network Topology

Our proposed HSV A consists of two partially-aligned variational autoencoders, which include three encoders (i.e., $E_x$, $E_a$, and $E_z$), and two decoders (i.e., $D_x$ and $D_a$). As described in the paper that $E_x$, $E_a$, $E_z$, $D_x$, and $D_a$ are MLP architectures, we present the architecture details of them as shown in Table 3.

| Visual encoder ($E_x$) | Input: $x$, size=2048;  
hidden layer: Fully connected, neurons=4096; LeakyReLU;  
Output: Fully connected, neurons=2048; |
|------------------------|-----------------------------------------------------------------|
| Semantic encoder ($E_a$) | Input: $x$, size=$|Att|$;  
hidden layer: Fully connected, neurons=4096; LeakyReLU;  
Output: Fully connected, neurons=2048; |
| Classifiers ($CLS^1/CLS^2$) | Input: $E^x(x)$ or $E^a(a)$, size=2048;  
hidden layer: Fully connected, neurons=512; BachNorm, LeakyReLU;  
Output: Fully connected, neurons=$|Seen|$; |
| Common encoder ($E_z$) | Input: $E^x(x)$ or $E^a(a)$, size=2048;  
hidden layer: Fully connected, neurons=2048; LeakyReLU;  
hidden layer: Fully connected, neurons=64*2; LeakyReLU;  
encoding layer: $\mu^x=64$ and $\delta^x=64$, or $\mu^a=64$ and $\delta^a=64$;  
Output: Reparametrization, $z^x$ or $z^a$, neurons=64; |
| Visual decoder ($D_x$) | Input: $z^x$, size=64;  
hidden layer: Fully connected, neurons=4096; LeakyReLU;  
Output: Fully connected, neurons=2048; |
| Semantic decoder ($D_a$) | Input: $z^a$, size=64;  
hidden layer: Fully connected, neurons=4096; LeakyReLU;  
Output: Fully connected, neurons=$|Att|$; |

Table 3: Network topology of HSV A. $|Att|$ is the dimensionality of semantic vectors per class, e.g., $|Att|=312$ in CUB. $|Seen|$ denotes the numbers of seen classes, e.g., $|Seen|=150$ in CUB.

B Hyperparameter Analysis

Features of Per Unseen Class in CZSL Setting ($N_u$). We evaluate the effect of the number of latent features per unseen class in CZSL. Since we only need to synthesize unseen features of unseen classes for training a classifier, We try a wide range of $N_u$ (i.e., $N_u = \{200, 400, 800, 1200, 1600, 2000\}$) for evaluation on CUB, SUN and AWA1 datasets as shown in Figure 4. Overall, the performance of HSV A is insensitive to the number of latent features per unseen class in CZSL. Targeting on better results, we set $N_u$ as 400, 200 and 800 for CUB, SUN and AWA1, respectively.
Features of Per Seen and Unseen Class in GZSL Setting ($N_s$ and $N_u$). We analyze the effect of the number of latent features per class in GZSL. We try a wide range of $N_s$ and $N_u$ (i.e., $N_s = \{100, 200, 400\}$ and $N_u = \{100, 200, 400, 800, 1200\}$) for evaluation on CUB and AWA1 datasets, resulting in a total of 15 pairs of ($N_s$, $N_u$), as shown in Figure 7. Since the visual features possess more discriminative information, we should set $N_u$ larger than $N_s$. Compared to HSV A using $N_s/N_u = 1/1$, HSV A improves classification accuracy using $N_s/N_u = 1/2$, achieving top-1 accuracy on unseen classes (Harmonic mean) improvement at least 17.5%(9.5%) and 36.3%(30.5%) on fine-grained dataset (e.g., CUB) and coarse-grained dataset (AWA1), respectively. Note that HSV A achieves better results on seen classes when $N_s/N_u$ is set to larger than 1/2. To trade-off top-1 accuracy on seen and unseen, we set ($N_s, N_u$) = (200, 400) to conduct all experiments.

Here we show how to set the dimensionality of the latent features in structure- and distribution-aligned common space, denoted as $Dim_S$ and $Dim_D$, respectively. As shown in Figure 5 and Figure 6, HSV A perform steadily on the coarse-grained dataset (e.g., AWA1) while it is sensitive to $Dim_S$ and $Dim_D$ on fine-grained datasets (e.g., CUB). On the fine-grained datasets, HSV A increases its accuracy on seen classes and decreases its accuracy on unseen classes when $Dim_S$ and $Dim_D$ are increased. We note that HSV A achieves significant results when $Dim_S = 2048$ and $Dim_D = 64$, thus we set $Dim_S$ and $Dim_D$ to 2048 and 64 respectively on CUB and AWA1.
Figure 7: Evaluating the effect of the number of synthesized latent features per seen/unseen class on (a) CUB and (b) AWA1 in GZSL setting.