Hierarchical Visual Categories Modeling: A Joint Representation Learning and Density Estimation Framework for Out-of-Distribution Detection
(*Supplementary Material*)

A. ImageNet OOD Evaluation Protocol

To test the proposed method's ability to identify near OOD samples, we use the remaining images in ImageNet to construct out-of-distribution datasets. We first calculate the semantic distances among 1000 object categories in ImageNet. We utilize a ResNet-50 model pre-trained on ImageNet to extract features for each image and calculate the mean features of each class. As stated in the main paper, the in-distribution dataset contains 100 image categories. We rank the remaining 900 image categories according to their average distances with these in-distribution categories. These ranked image categories are divided into subsets, each containing 100 image categories. Finally, we get nine out-of-distribution datasets.

B. More Experimental Results

Number of attribute groups. We varied the number of attribute groups from 8 to 128 to analyze the components in Gaussian mixture models. In table Table 1, the performances of HVCM are stable with varying attribute groups. It demonstrates that our method is suitable for different numbers of attribute groups. However, setting G to bigger numbers ($G = 64, 128$) does not help us improve the performance. Thus, we finally set G to 32.

Table 1. The detailed performance of HVCM with varying numbers of group centers G. ↑ indicates larger values are better, and ↓ is the opposite. **Bold** numbers are superior results. All values are percentages.

Concepts Number	**OOD Datasets**								**Average**	
	iNaturalist		**SUN**		**Places**		**Textures**			
	FPR95↓	**AUROC↑**	**FPR95↓**	**AUROC↑**	**FPR95↓**	**AUROC↑**	**FPR95↓**	**AUROC↑**	**FPR95↓**	**AUROC↑**
G=8	22.18	90.62	18.46	94.22	20.75	93.39	36.09	88.68	24.37	91.73
G=16	**20.98**	**92.38**	**17.16**	**94.54**	20.44	92.76	33.70	89.26	23.07	92.24
G=32	21.56	92.19	17.20	94.44	**19.98**	**93.62**	**29.22**	**90.68**	**21.99**	**92.73**
G=64	23.00	90.82	18.06	93.90	21.46	92.54	38.54	90.77	25.27	92.01
G=128	23.26	89.53	18.98	94.10	20.84	93.53	36.20	88.17	24.82	91.33

Table 2. The detailed performance of HVCM with a varying number of dimensions of the attribute space. ↑ indicates larger values are better, and ↓ is the opposite. **Bold** numbers are superior results. All values are percentages.

Attribute Dimension	**OOD Datasets**								**Average**	
	iNaturalist		**SUN**		**Places**		**Textures**			
	FPR95↓	**AUROC↑**	**FPR95↓**	**AUROC↑**	**FPR95↓**	**AUROC↑**	**FPR95↓**	**AUROC↑**	**FPR95↓**	**AUROC↑**
2048	22.76	92.09	**16.40**	**95.37**	19.78	**94.55**	36.16	89.30	23.78	**92.83**
4096	**21.06**	91.89	16.56	94.92	**18.56**	94.27	31.94	89.83	22.03	92.73
8192	21.56	**92.19**	17.20	94.44	19.98	93.62	**29.22**	**90.68**	**21.99**	92.73

Different dimensions of attribute space. In Table 2, we investigate the impact of the different dimensions of attribute spaces. We set the dimensions of attribute space from 2048 to 8192. The experimental results demonstrate that our method works efficiently with different dimensions. The AUROC varies slightly across different attribute dimensions. However, when the feature dimension increases, we achieve a lower FPR95. Therefore, we finally set the dimension of the attribute space to 8192.