Supplementary Material for “Hierarchical Disentangling Network for Object Representation Learning”

Shishi Qiao[a,b,c], Ruiping Wang[b,c,d], Shiguang Shan[b,c], Xilin Chen[b,c,*]

[a] *College of Information Science and Engineering, Ocean University of China, QingDao, 266100, China*
[b] *Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China*
[c] *University of Chinese Academy of Sciences, Beijing, 100049, China*
[d] *Beijing Academy of Artificial Intelligence, Beijing, 100084, China*

In this document, we give additional implementation details and experimental results for the corresponding sections in the main paper to support the method we proposed.

1. Network Architecture Details

Let *c7s1-k* denotes a convolution block with k filters of 7 × 7 size and stride 1. *dk* means a convolution block with k filters of 4 × 4 size and stride 2. *Rk* denotes a residual block that contains two convolution blocks with k filters of 3 × 3 size . The last layers at non-root levels in the bottom encoder branch (note that common features of the root level are encoded by the upper encoder branch) are implemented by multiple *c1s1-8* (i.e., the convolution block with 8 filters of 1 × 1 size and stride 1. *uk* denotes a 2× nearest-neighbor upsampling layer followed by a convolution block with k filters of 5 × 5 size and stride 1. GAP denotes a global average pooling layer. Instance Normalization (IN) is adopted to the upper encoder branch. In the following, we give the detailed architectures of each module for training our HDN.

On general datasets with 128 × 128 **resolution inputs**:

Upper encoder: c7s1-64, d128, d256, R256, R256, R256

[*] Corresponding author
Email address: xlchen@ict.ac.cn (Xilin Chen)

Bottom encoder: c7s1-64, d128, d256, d256, d256, GAP, c1s1-8

Decoder: R256, R256, R256, u128, u64, c7s1-3

Discriminator & Classifier: d64, d128, d256, d512

On the Fashion-MNIST with 28×28 resolution inputs:

Upper encoder: c7s1-32, d64, d128, R128, R128, R128

Bottom encoder: c7s1-32, d64, d128, R128, R128, R128, GAP, c1s1-8

Decoder: R128, R128, R128, u64, u,32 c7s1-1

Discriminator & Classifier: d32, d64, d128, d256

2. Disentanglement Results on Challenging ImageNet-animal Data

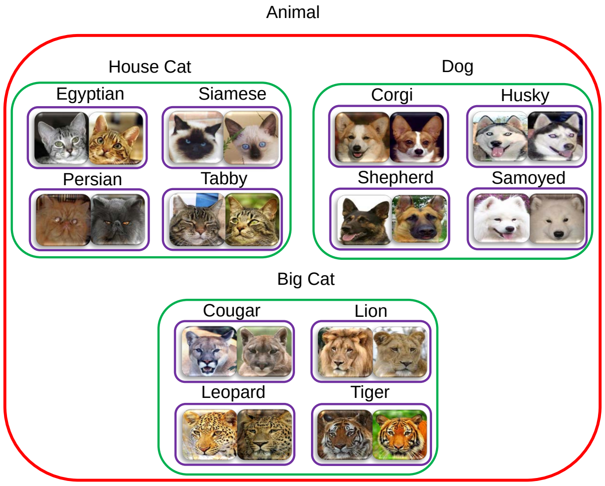
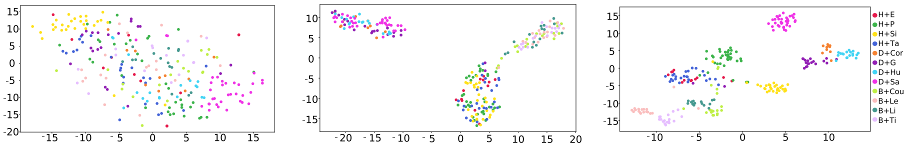


Figure 1: Typical samples of the constructed hierarchical data on ImageNet. Images within a purple rectangular box are instances of a leaf-level category. Images within a green rectangular box belong to one common super-category. The super-categories within the red rectangular box share one common ancestor, i.e. the *animal*.

In this section, we show results of HDN on the collected challenging ImageNet-animal data and analyze the limitations of our method. To be specific, We collect images from 3 super categories including the *house cats, dogs* and *big cats* on the ImageNet. Each super category contains 4 fine-grained sub-categories, and we thus construct a three-level hierarchical structure (the root is *animal*). To train and test our HDN, all images are split by the official train/test protocol and preprocessed by a pre-trained faster-rcnn head detector, the detection results of which are then cropped and resized to 128*128 resolution. Examples of the preprocessed hierarchical data are shown in Fig.1. Network architecture and training hyper-parameters are same with the settings on CelebA, CADCars and ShapeNet, as introduced in Sec.3.4 of the main paper and the text in Sec.1 of this document.



(a) Root

(b) Level 2

(c) Level 3

Figure 2: 2D tSNE visualization of the disentangled $\mathbf{F}_l$ on the test set of ImageNet-animal at different levels. H, D and B mean House cat, Dog and Big cat, respectively. E, P, Si, Ta, Cor, G, Hu, Sa, Cou, Le, Li and Ti mean Egyptian, Persian, Siamese, Tabby cat, Corgi, German shepherd, Husky, Samoyed, Cougar, Leopard, Lion, and Tiger, respectively.

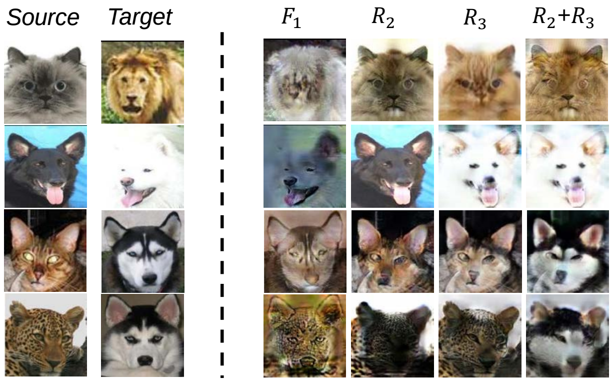


Figure 3: Semantic translation results of the source images controlled by hierarchically disentangled features of the targets. On such hierarchical data, all levels are complex categorical variation, i.e., R_2 should encode the information of the image being one of the three super-categories (e.g., the *house cat*), while R_3 should contain the information of the image being one of the further divided fine-grained sub-categories (e.g., one kind of house cat).

We first quantitatively evaluate the hierarchical classification accuracy of generated images conditioned on the disentangled features as we did in Sec.4.1 of the main paper. The results on real test images at level 2 and 3 are 0.9293 and 0.8760, and on the generated images are 0.9493 and 0.8160, respectively. Fig.2 shows the tSNE embedding results using different levels of $\mathbf{F}_l$. From these results, we may infer that our method has successfully disentangled the desired semantic features at different levels, since the discriminability are progressively increased and the generalization ability on the generated images also seems satisfactory. However, qualitative investigation reveals that it is not the truth of all. Fig.3 shows some semantic translation results of the source object images conditioned on the disentangled hierarchical features of the targets. It is observed that the disentangled semantic information from the targets can only change partial appearance (e.g., the textures or colors) of the source images, while lose the necessary and even the key information to recognize the objects at that level (e.g., the shape of the lion rather than the skin color). Besides, by purely changing only one particular level of features, the generated images sometimes look strange.

Reasons for these phenomena are mainly in two aspects. On the one hand, there exists too much information that can be leveraged for classification. Since these ImageNet categories themselves are too complex, the differences within them are in many aspects. Consequently, the classifiers can easily find "shortcuts" and extract only partial discriminative primitives from the objects at that level. Sometimes these "shortcuts" are even the wrong evidence, which is the so-called bias problem in many powerful ImageNet classification models (e.g., the images containing black man are predicted as basketball) [1, 2]. From the qualitative results of HDN in Fig.3, we also find that the semantic information of disentangled features is not sufficient to interpret the objects of being the categories at that level, and sometimes the semantic meanings are very difficult to be understood by humans. This tells us that sometimes deep features can perform well in terms of certain quantitative measurements, but may not work in the manner as we expected. Even so, our HDN can diagnose this kind of problems as done in Fig.3. On the other hand, the poor image quality is partially owing to the capacity of GAN. Generating high-quality images on the ImageNet is notoriously difficult for GAN-based methods until now, due to the much complex data distribution. In our HDN framework, in order to disentangle semantic commonality and individuality among categories, it is required to synthesize *nonexistent* categories combined by semantics from different levels, which further makes the objectives of distribution fitting harder for the GAN-based framework. We believe that the performance of HDN would be improved on the ImageNet dataset with the development of generative frameworks.

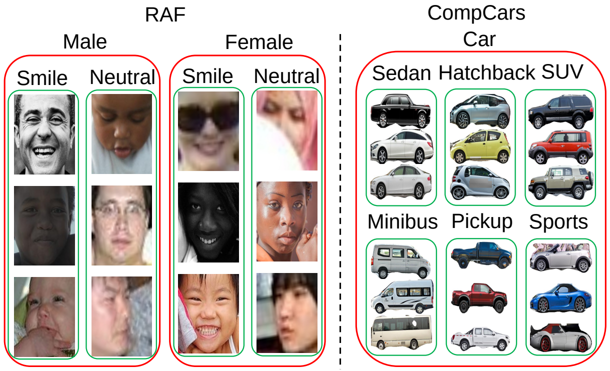


Figure 4: Typical samples of the hierarchical data on the RAF and CompCars datasets. Each image represents one leaf-level category (e.g., different races or ages, different car make models, etc.). Images within a green rectangular box belong to one common super-category (e.g., the smile, the sedan, etc.). The super-categories within a red rectangular box share one common ancestor.

3. Cross Dataset Study

Learning general representation that can be applied across datasets is one of the long goals for machine learning and computer vision. In this section, we briefly evaluate our method on datasets which have similar categorical annotaions but quite different domain styles, compared to the datasets we have evaluated in previous experiments. To be specific, we evaluate HDN on a quite challenging facial expression dataset named RAF [3] and a car dataset named CompCars [4], using models trained on CelebA and CADCars, respectively. The

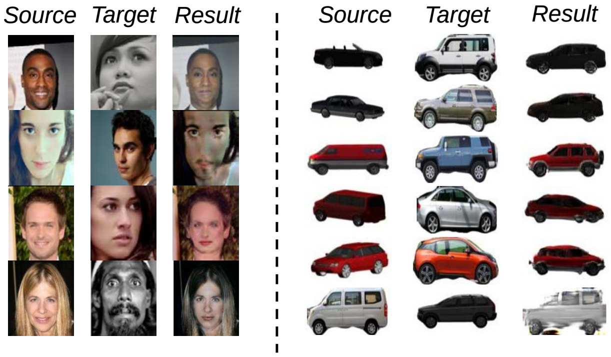


Figure 5: Semantic translation results between image pairs from different datasets (i.e., CelebA and RAF, and CADCars and CompCars.), using all levels of disentangled features of the targets to replace those of the sources. Compactly cropped face images without hair information are from RAF, while the other faces are from CelebA. It is found that the information of gender and smile (i.e., non-leaf level semantics) is correctly transferred. Car images in the right second column (i.e., the Target column) except the one in the last row are from CompCars, and the other cars are from the CADCars. It is observed that the car types and poses are changed accordingly.

RAF dataset provides expressions, race, age range and gender attributes annotations. Besides, the released images are compactly aligned which have little information about hair colors. Therefore, the leaf-level categories are organized according to the race or age range for the RAF dataset, instead of hair colors originally defined on the CelebA. As for the CompCars dataset, it contains 163 car makes with 1,716 car models, and also provides the car type annotations (i.e., SUV, Sedan, Sports, etc.). Based on these annotations, we replace the pose labels at the leaf-level on CADCars with different car model annotations for the CompCars dataset[1]. Typical examples of the two hierarchical data are shown in Fig.4.

Fig.5 shows the semantic translation results across datasets. It is observed that the information of gender and smile (i.e., the semantic at non-leaf levels) is correctly disentangled and transferred. For the translation between image

[1] It does not provide the same pose annotations as used on the CADCars.

pairs from the CADCars and the CompCars, given the unseen target image of the *Hatchback* car type from the CompCars (the fifth row), the translated result of the source SUV image looks like nothing on the earth (like a SUV but has cambered shape). Besides, we also find it difficult to translate the source images which are from unseen dataset, as shown in the last car case, which is mainly due to the domain shift for the generator (i.e., the upper encoder branch can not extract meaningful basic information from the unseen dataset for the subsequent hierarchical feature aggregation and image generation).

References

[1] P. Stock, M. Cissé, Convnets and imagenet beyond accuracy: Explanations, bias detection, adversarial examples and model criticism, CoRR abs/1711.11443.

[2] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, W. Brendel, Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness, in: ICLR, 2019.

[3] S. Li, W. Deng, J. Du, Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild, in: IEEE, CVPR, 2017, pp. 2584–2593.

[4] L. Yang, P. Luo, C. C. Loy, X. Tang, A large-scale car dataset for fine-grained categorization and verification, in: IEEE, CVPR, 2015, pp. 3973–3981.