Supplementary Material for "Hierarchical Image-to-image Translation with Nested Distributions Modeling"

Shishi Qiao[a,b,c], Ruiping Wang[b,c,*], 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*

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

1. Network Architectures

Following the backbone designs in [1] for image-to-image translation task, let *c7s1-k* denotes a 7×7 convolution block with k filters and stride 1. *dk* means a 4×4 convolution block with k filters and stride 2. *Rk* denotes a residual block that contains two 3×3 convolution blocks with k filters. The last layer *c1s1-8* in the style encoder is a 1×1 convolution block with 8 filters and stride 1. Therefore, we obtain 8 dimensions of style codes. Similarly, the mean and diagonal elements of the covariance matrix for each Gaussian are also parameterized with 8 dimensions to be optimized with the generator simultaneously. *uk* denotes a $2\times$ nearest-neighbor upsampling layer followed by a 5×5 convolution block with k filters and stride 1. GAP denotes a global average pooling layer. Instance Normalization (IN) and Adaptive Instance Normalization (AdaIN) are adopted to the content encoder branch and decoder respectively. No normalization is used

*Corresponding author
Email address: wangruiping@ict.ac.cn (Ruiping Wang)

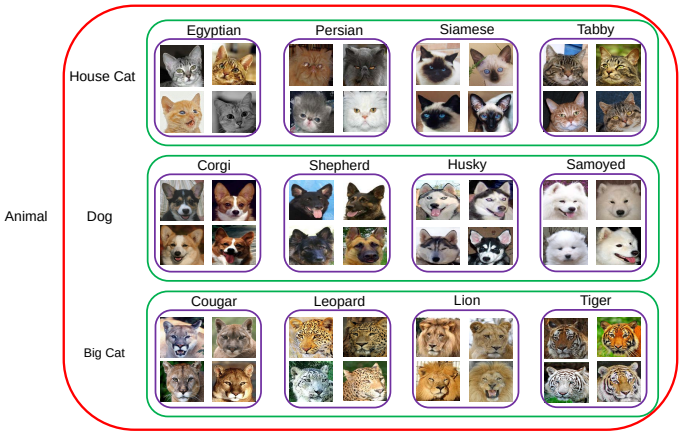


Figure 1: Typical samples of hierarchical data on ImageNet. Images within a purple rectangular box are some instances of a leaf-level category. Categories within a green rectangular box belong to one common super-category. The super-categories within a red rectangular box share one common ancestor.

in the style encoder branch. Use ReLU activations in the encoder-decoder and Leaky ReLU with slope 0.2 in the discriminator and classifier. Multi-scale discriminators with 3 scales and objective of LSGAN [2] are used to ensure both realistic details and global structure preserved. The last layer of the decoder is equipped with a *tanh* activations to normalize the values of generated images to the range of $[-1, 1]$. In the following, we give detailed architectures of each module.

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

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

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

Discriminator & Classifier: d64, d128, d256, d512

2. Hierarchical Data Construction

In this section, Fig. 1, Fig. 2 and Fig. 3 provide leaf-level examples used in the main paper for better understanding the nested relationships among categories in different hierarchy levels. Take the ImageNet animal for example,

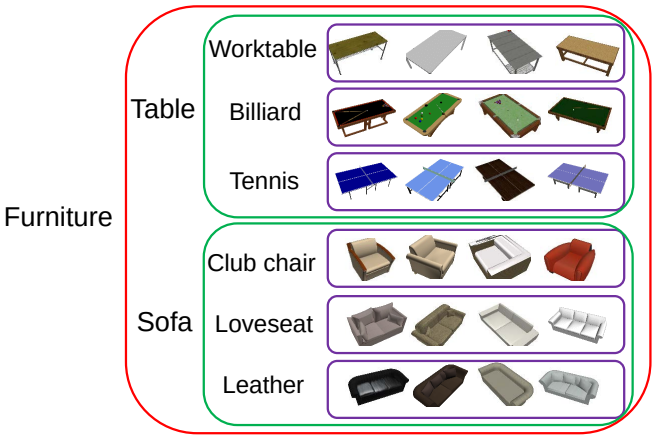


Figure 2: Typical samples of hierarchical data on ShapeNet. Images within a purple rectangular box are some instances of a leaf-level category. Categories within a green rectangular box belong to one common super-category. The super-categories within a red rectangular box share one common ancestor.

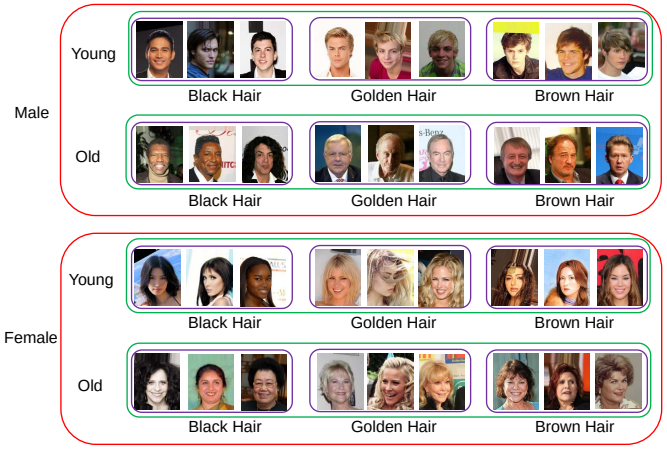
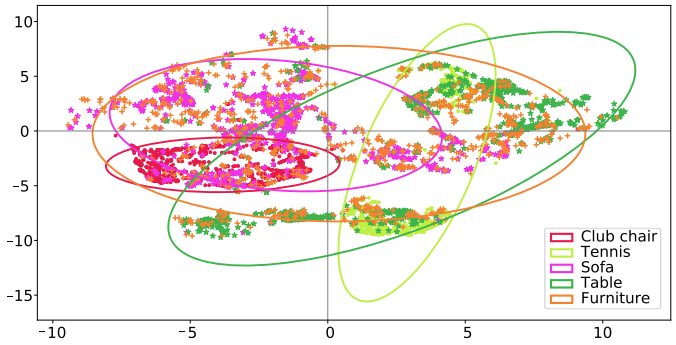
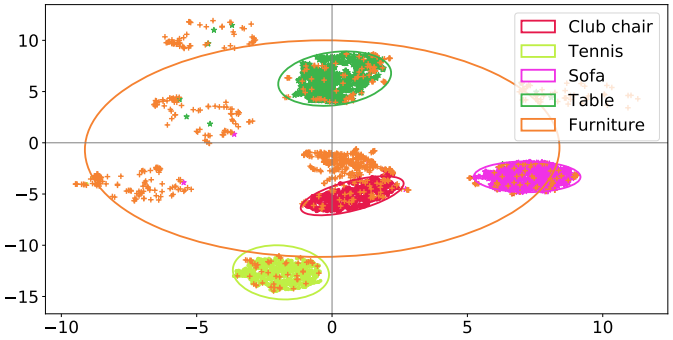


Figure 3: Typical samples of hierarchical data on CelebA. Images within a purple rectangular box are some instances of a leaf-level category. Categories within a green rectangular box belong to one common super-category. The super-categories within a red rectangular box share one common ancestor.



(a) Full HIT



(b) HIT w/o $\mathcal{L}_{nest}$

Figure 4: 2D UMAP visualization of learned Gaussian distributions of domains in different hierarchy levels on ShapeNet for (a) Full HIT and (b) HIT w/o $\mathcal{L}_{nest}$. For each domain, $1,000$ points are sampled and fitted for a Gaussian ellipse.

the root category *animal* has three children, i.e. *house cat, dog* and *big cat*. For each of the three super domains, it includes four finer-grained children. Within each leaf-level category, samples mainly contain intra-class variations caused by shapes, colors, poses, etc.

3. More Details of User study

In this section, we introduce more details about the user study settings in our experiments. The 30 users are selected from both undergraduate and postgraduate students. The distribution of them is half-and-half. The undergraduate

students are mainly majored in computer science and technology, and the post-graduate ones are learning in computer applications. On each test set, we first randomly select 100 images as the source. Second, for each source image, a different domain is randomly chosen as the translation target. Third, for compared multi-modal methods, we randomly choose one translated result. To make it stable, such random steps are conducted 3 times, resulting in 300 source images and 300 translated results for each method. Besides, for each participant user, the results of the second and third steps are different by using different random seeds.

During the user study, we show users the source image, target domain name, and translated images of all compared methods (methods order is shuffled in each case). Users are asked to choose the best one he/she regards with unlimited time, based on the three tips shown on the annotation tool: 1). Results should be distinguished as the target domain easily (most important) 2). Results should perceptually have high image quality, i.e. looks as real as possible. 3). Results should share some common information with the sources, such as pose, background (i.e. domain irrelevant).

4. Additional Experimental Results

In this part, we show additional experimental results which are complementary to our main paper. Fig. 4 shows a 2D visualization of learned Gaussians of some categories at different hierarchy levels on the ShapeNet dataset. Similar conclusions with the ones in the main paper can be drawn, we do not repeat that here.

4.1. Facial Attributes Transfer

In this subsection, we extend our HIT to the task of facial attributes transfer on CelebA [3] dataset. CelebA provides more than 200K face images with 40 attribute annotations. To make a comparison with existing attribute editing methods, we define a hierarchy by imitating the category hierarchy. Specifically, all faces are first clustered into male and female, and further classified

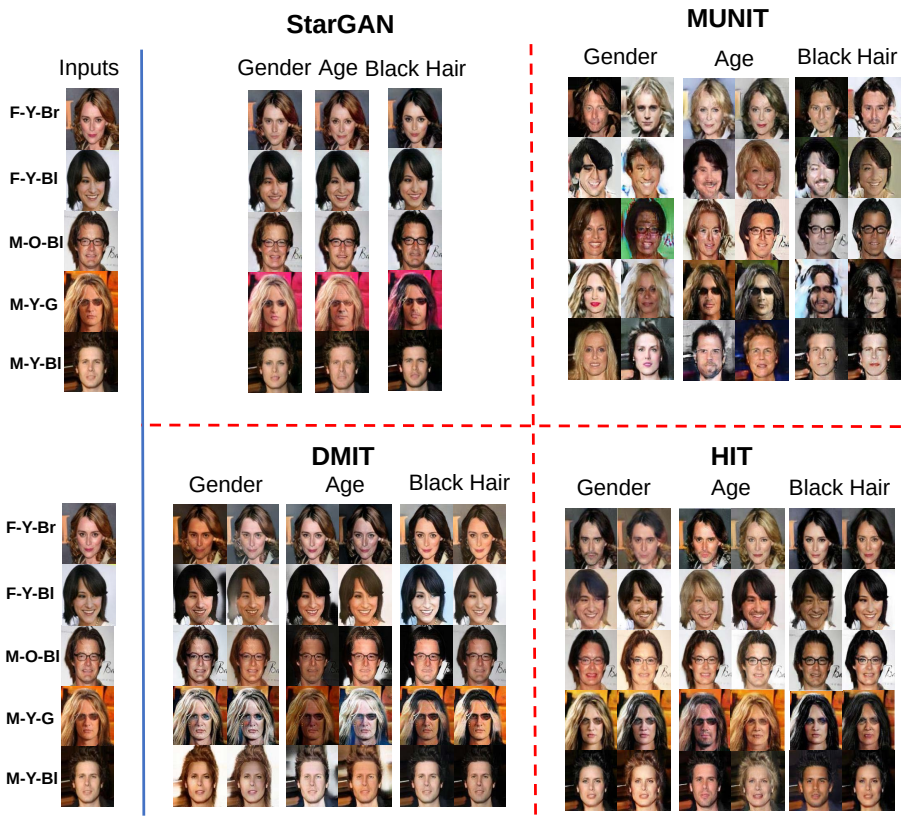
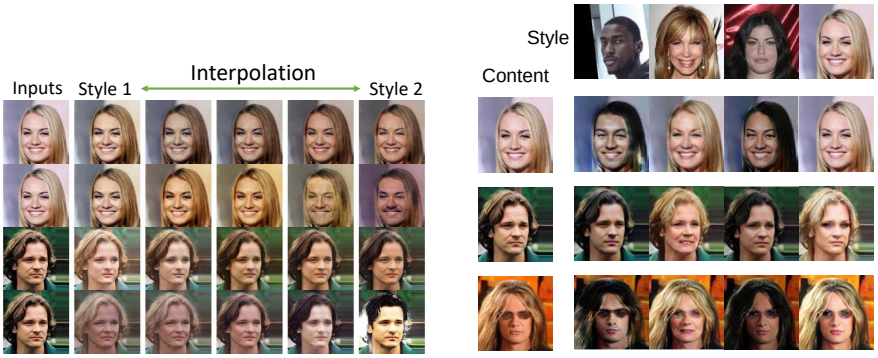


Figure 5: Qualitative comparison on CelebA. The inputs are translated to their reversed value for gender and age attributes, and to black hair color (the attributes shown on the left are the raw values of inputs, and abbreviations are: Female(F), Male(M), Young(Y), Old(O), Brown (Br), Black (Bl) and Golden (G) hair.). StarGAN learns one-to-one mapping. MUNIT, DMIT, and our HIT can generate multi-modal results (2 outputs for each input are randomly sampled from distributions). StarGAN and DMIT only change the target attribute, leaving others unchanged by modifying that attribute dimension of the one-hot vector. MUNIT models one attribute as a domain with other attributes as intra-domain variations and thus may reach the target domain with other attributes changed. As for HIT, styles of each target attribute can be sampled from any domain node containing that attribute in the hierarchy, e.g. styles of the old attribute can be sampled from 6 different combined domains (i.e. old $\times$ [male or female] $\times$ [brown, black or golden] hair colors), which can be controlled by users (i.e. one can sample from the distribution node with/without changing other raw attribute values). Best viewed in colors and zoom-in.

according to the age and hair color in the next two levels, typical examples of such hierarchy are shown in the supplementary materials. Note that the order of attributes selection is not unique, one can define other data split orders. On this fine-grained translation task, the entropy loss defined in Eqn.(4) of the main paper is dropped as we find it may do harm to the preserving of face identity. MUNIT is trained on domain pairs of male↔female, young↔old and black hair↔golden hair on CelebA.

Fig. 5 shows qualitative results on CelebA. It is observed that StarGAN (FID:21.50) and DMIT (FID:28.20) achieve outstanding image quality on such fine-grained translations among attribute domains, while MUNIT (FID:88.19) is not satisfactory for this. The reasons may be that using only the adversarial learning (MUNIT) to find fine-grained attribute differences between domains is not stable while additional multi-domain classifier (StarGAN, DMIT, and HIT) is good at such objective. In terms of image quality, our HIT (FID:38.71) performs comparably with StarGAN and DMIT. Considering that StarGAN is elaborately designed for this task, the results of a simple extension of our HIT are overall acceptable. Besides, our HIT considers both multi-modal distributions fitting and multi-domain classification by splitting attributes in a hierarchy, one can control the degree of attributes change by sampling from domains containing the target attribute at different hierarchy levels, e.g. change a female to male by sampling from domain of male, old male or old male with black hair, the progress of which is as simple as changing the one-hot attribute vector in StarGAN and DMIT.

In Fig. 6(a), we further study the smoothness of learned distributions. It is observed one can conduct smooth translation via linear interpolations between styles from different attribute domains. Besides, our method can provide style transfer (also called example guided translation) as done in [1] and [4], i.e. use the styles of referenced real images instead of sampling them from distributions. Fig. 6(b) shows some results. We can find that the semantics of gender, age, and hair colors in the source images are correctly transferred to the specified styles of the target images.



(a) Style interpolation (b) Style transfer

Figure 6: Translations by (a) interpolations of sampled styles from different domain distributions and (b) style transfer between real images. In (a), inputs are translated to the target styles which are the interpolations between two sampled styles, e.g. in the fourth case style 1 means old female with golden hair and style 2 denotes young male with black hair, and the interpolations between them are mixed attributes of style 1 and 2. In (b), images in the left-most column provide content parts (encoding pose, identity, background, etc.) and the ones in the top row bring style parts (encoding gender, age, and hair colors) for image generation, e.g. in the first case the hair color and gender of the source image (young female with golden hair) is transferred to the style of the first target image (young male with black hair).

4.2. Failure Case Analysis

The assumption of single Gaussian for all category domains in the common space has some limits, the sampling of distributions at high levels may fail. As shown in Fig. 7, it is found that the failure cases (highlighted by red bounding boxes) present blurred and miss some local details. On one hand, though Gaussian distribution prior is a good approximation for many real data, it may not be completely qualified when the scale of available training data is not enough to capture the full data manifold with large intra-domain variations, such as the used hierarchical data of ImageNet and ShapeNet in this paper. On the other hand, the parent distributions should be better modeled by a mixture of Gaussians given multiple single Gaussians of its children. As illustrated in Fig. 8, this issue would lead to sparse sampling around the center of parent distribution

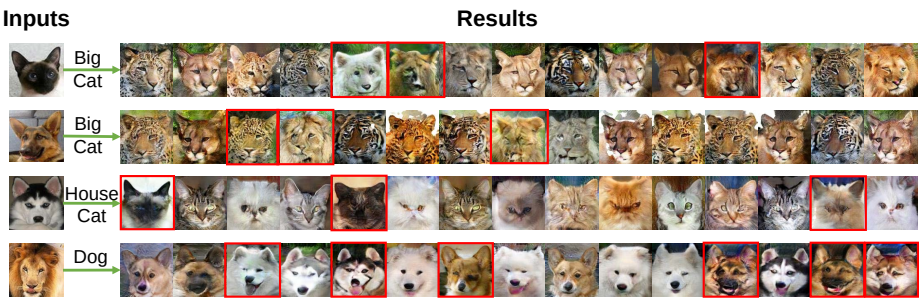


Figure 7: Qualitative generation results of multiple random style samplings. For each input, 15 images are generated by randomly sampling in a particular target super-domain (House cat, dog, or big cat). Images with poor quality are highlighted by red rectangular bounding boxes. Best viewed in colors.

and thus poor generation quality.

To verify this, we conduct experiments to compare the image quality of translations of different distribution modeling schemes on the ImageNet dataset. Specifically, the parent distribution of the house cat, dog, and big cat is modeled by the Gaussian mixture model (GMM) fitted with the EM algorithm, the learned single Gaussian model (SMM) of our HIT method, and the offline-fitted SMM as Fig. 8, respectively. By sampling from each of the three distributions, the input images are translated to the parent category. The FID results of such three schemes are 49.80, 56.44, and 64.92, respectively. We can infer that the GMM might be more reasonable than the SMM for the hierarchical distribution space modeling. To address this issue, we made efforts to exploit the idea of the mixture of Gaussians to train our HIT, but found that it is hard to compute the KL divergence between two mixture of Gaussians which does not have an analytical solution. Besides, the re-parameterize trick for distribution sampling during SGD optimization can not be transferred to the case of the mixture of Gaussians. A more principled assumption to realize the nested relationships among parent-children distributions is a promising direction for our future research.

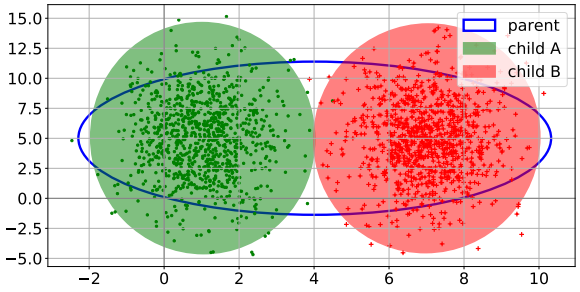


Figure 8: 2D illustration of a possible situation using the single Gaussian assumption for nested modeling. Given single Gaussians of two children, sparse sampling would occur around the fitted single Gaussian distribution center of the parent.

References

[1] X. Huang, M. Liu, S. J. Belongie, J. Kautz, Multimodal unsupervised image-to-image translation, in: ECCV, 2018, pp. 179–196.

[2] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, S. P. Smolley, Least squares generative adversarial networks, in: IEEE , ICCV, 2017, pp. 2813–2821.

[3] Z. Liu, P. Luo, X. Wang, X. Tang, Deep learning face attributes in the wild, in: IEEE ,ICCV, 2015, pp. 3730–3738.

[4] Y. Chen, X. Xu, Z. Tian, J. Jia, Homomorphic latent space interpolation for unpaired image-to-image translation, in: IEEE, CVPR, 2019, pp. 2408–2416.