HIERARCHICAL IMAGE-TO-IMAGE TRANSLATION WITH NESTED DISTRIBUTIONS MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Unpaired image-to-image translation among category domains has achieved remarkable success in past decades. Recent studies mainly focus on two challenges. For one thing, such translation is inherently multimodal due to variations of domain-specific information (e.g., the domain of house cat has multiple fine-grained subcategories). For another, existing multimodal approaches have limitations in handling more than two domains, i.e. they have to independently build one model for every pair of domains. To address these problems, we propose the Hierarchical Image-to-image Translation (HIT) method which jointly formulates the multimodal and multi-domain problem in a semantic hierarchy structure, and can further control the uncertainty of multimodal. Specifically, we regard the domain-specific variations as the result of multi-granularity property of domains, and one can control the granularity of the multimodal translation by dividing a domain with large variations into multiple subdomains which capture local and fine-grained variations. With the assumption of Gaussian prior, variations of domains are modeled in a common space such that translations can further be done among multiple domains within one model. To learn such complicated space, we propose to leverage the inclusion relation among domains to constrain distributions of parent and children to be nested. Experiments on several datasets validate the promising results and competitive performance against state-of-the-arts.

1 INTRODUCTION

Image-to-image translation is the process of mapping images from one domain to another, during which changing the domain-specific aspect and preserving the domain-irrelevant information. It has wide applications in computer vision and computer graphics Isola et al. (2017); Ledig et al. (2017); Zhu et al. (2017a); Liu et al. (2017); Huang et al. (2018) such as mapping photographs to edges/segments, colorization, super-resolution, inpainting, attribute and category transfer, style transfer, etc. In this work, we focus on the task of attribute and category transfer, i.e. a set of images sharing the same attribute or category label is defined as a domain[1].

Such task has achieved significant development and impressive results in terms of image quality in recent years, benefiting from the improvement of generative adversarial nets (GANs) Goodfellow et al. (2014); Mirza & Osindero (2014). Representative methods include pix2pix Isola et al. (2017), UNIT Liu et al. (2017), CycleGAN Zhu et al. (2017a), DiscoGAN Kim et al. (2017), DualGAN Kim et al. (2017) and DTN Taigman et al. (2017). More recently the study of this task mainly focus on two challenges. The first is the ability of involving translation among several domains into one model. It is quite a practical need for users. Using most methods, we have to train a separate model for each pair of domains, which is obviously inefficient. To deal with such problem, StarGAN Choi et al. (2018) leverages one generator to transform an image to any domain by taking both the image and the target domain label as conditional input supervised by an auxiliary domain classifier.

Another challenge is the multimodal problem, which is early addressed by BicycleGAN Zhu et al. (2017b). Most techniques including aforementioned StarGAN can only give a single determinate output in target domain given an image from source domain. However, for many translation task, the mapping is naturally multimodal. As shown in Fig.1, a *cat* could have many possible appearances such as being a *Husky*, a *Samoyed* or other dogs when translated to the *dog* domain. To address

[1] Since attributes can be treated as fine-grained categories, we denote a category as a domain in the following.

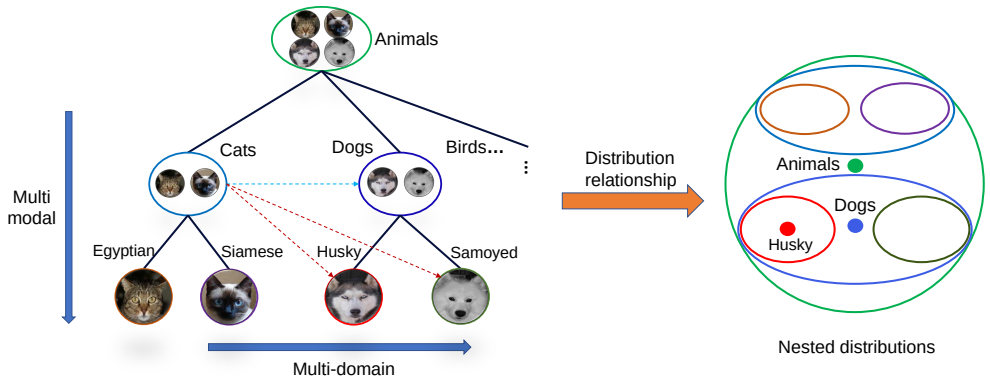


Figure 1: An illustration of a hierarchy structure and the distribution relationship in a 2D space among categories in such hierarchy. Multi-domain translation is shown in the horizontal direction (blue dashed arrow) while multimodal translation is indicated in the vertical direction (red dashed arrow). Since one child category is a special case of its parent, in the distribution space it is a conditional distribution of its parent, leading to the nested relationship between them.

this issue, recent works including BicycleGAN Zhu et al. (2017b), MUNIT Huang et al. (2018) and DRIT Lee et al. (2018) model a continuous and multivariant distribution independently for each domain to represent the variations of domain-specific information, and they have achieved diverse and high-quality results for several two-domain translation tasks.

In this paper , we aim at involving the abilities of both multi-domain and multimodal translation into one model. As shown in Fig.1, it is noted that categories have the natural hierarchical relationships. For instance, the *cat*, *dog* and *bird* are three special children of the animal category since they share some common visual attributes. Furthermore, in the *dog* domain, some samples are named as *husky* and some of them are called *samoyed* due to the appearance variations of being the dog. Of course, one can continue to divide the husky to be finer-grained categories based on the variations of certain visual attributes. Such hierarchical relationships widely exist among categories in real world since it is a natural way for our human to understand objects according to our needs in that time.

We go back to the image translation task, the multi-domain and multimodal issues can be understood from horizontal and vertical views respectively. From the horizontal view as the blue dashed arrow indicates, multi-domain translation is the transformation in a flat level among categories. From the vertical view as the red dashed arrow indicates, multimodal translation further considers variations within target category, i.e. the multimodal issue is actually due to the multi-granularity property of categories. In the extreme case, every instance is a variation mode of the domain-specific information. Inspired by these observations, we propose a Hierarchical Image-to-image Translation (HIT) method which translates object images among both multiple category domains in a same hierarchy level and their children domains. To this end, our method models the variations of all domains in forms of multiple continuous and multivariant Gaussian distributions in a common space. This is different from previous methods which model the same Gaussian distribution for two domains in independent spaces and thus can not work with only one generator. Due to the hierarchical relationships among domains, distribution of a child domain is the conditional one of its parent domain. Take such principle into consideration, distributions of domains should be nested between a parent and its children, as a 2D illustration shown in Fig.1. To effectively supervise the learning of such distributions space, we further improve the traditional conditional GAN framework to possess the hierarchical discriminability via a hierarchical classifier. Experiments on several categories and attributes datasets validate the competitive performance of HIT against state-of-the-arts.

2 Related Works

Conditional Generative Adversarial Networks. GAN Goodfellow et al. (2014) is probably one of the most creative frameworks recently for the deep learning community. It contains a generator and a discriminator. The generator is trained to fool the discriminator, while the discriminator in turn tries to distinguish the real and generated data. Various GANs have been proposed to improve the training stability, including better network architectures Radford et al. (2016); Denton et al. (2015); Zhang et al. (2017); Karras et al. (2017); Brock et al. (2019), more reasonable distribution metrics Mao et al. (2017); Arjovsky et al. (2017); Gulrajani et al. (2017) and normalization schemes Miyato et al. (2018); Karras et al. (2018). With these improvements, GANs have been applied to many

conditional tasks Mirza & Osindero (2014), such as image generation given class labels Odena et al. (2017), super resolution Ledig et al. (2017), text2image Reed et al. (2016), 3D reconstruction from 2D input Wu et al. (2016) and image-to-image translation introduced in the following.

Image-to-image Translation. Pix2pix Isola et al. (2017) is the first unified framework for the task of image-to-image translation based on conditional GANs, which combines the adversarial loss with a pixel-level L1 loss and thus requires the pairwise supervision information between two domains. To address this issue, unpaired methods are proposed including UNIT Liu et al. (2017), DiscoGAN Kim et al. (2017), DualGAN Yi et al. (2017) and CycleGAN Zhu et al. (2017a). UNIT combines the varitional auto-encoder and GAN framework, and proposes to share partial network weights of two domains to learn a common latent space such that corresponding images in two domains can be matched in this space. DiscoGAN, DualGAN and CycleGAN leverage a cycle consistency loss which enforces that we can re-translate the target image back to the original image.

More recently, works in this area mainly focus on the problems of multi-domain and multimodal. To deal with translation among several domains in one generator, StarGAN Choi et al. (2018) takes target label and input image as conditions, and uses an auxiliary classifier to classify translated image into its belonged domain. As for the multimodal issue, BicycleGAN Zhu et al. (2017b) proposes to model continuous and multivariant distributions. However, BicycleGAN requires input-output paired annotations. To overcome this problem, MUNIT Huang et al. (2018) and DRIT Lee et al. (2018) adopt a disentangled representation for learning diverse translation results from unpaired training data. Chen et al. (2019) propose to interpolate the latent codes between input and referred image to realize generation of diverse images. Different from all aforementioned works, we aim at realizing both multi-domain and multimodal translation in one model using the natural hierarchical relationships among domains defined by category or attribute.

Hierarchy-regularized Learning. Hierarchical learning is a natural learning manner for human beings and we often describe objects in the world from abstract to detailed according to our needs of the time. For machine learning and computer vision, such semantic hierarchies have been widely explored in object classification for accelerating recognition Griffin & Perona (2008); Marszalek & Schmid (2008), obtaining multiple granularities of predictions Deng et al. (2012); Ordonez et al. (2013), making use of category relation graphs Deng et al. (2014); Ding et al. (2015), and improving recognition performance as additional supervision Zhao et al. (2011); Srivastava & Salakhutdinov (2013); Hwang & Sigal (2014); Yan et al. (2015); Goo et al. (2016); Ahmed et al. (2016). Apart from these discriminative tasks, Xie et al. (2017); Zhao et al. (2017) propose to use generative models to disentangle the factors from low-level representations to high-level ones that can construct a specific object. Singh et al. (2019) uses an unsupervised generative framework to hierarchically disentangle the background, object shape and appearance from an image. In natural language processing, Athiwaratkun & Wilson (2018) propose a probabilistic word embedding method to capture the semantics described by the WordNet hierarchy. Our method first introduces such semantic hierarchy to learn a both multi-domain and multimodal translation model.

3 Approach

3.1 Problem formulation

Let $x_i \in \mathcal{X}_i$ be an image from domain i. Our goal is to estimate the conditional probability $p(x_j|x_i)$ by learning an image-to-image translation model $p(x_{i\to j}|x_i)$, where $x_{i\to j}$ is a sample produced by translating x_i to domain $\mathcal{X}_j$. Generally speaking, $p(x_j|x_i)$ are multimodal due to the intra-domain variations. To deal with the multimodal problem, similar to Huang et al. (2018), we assume that x_i is disentangled by an encoder E into the content part $c \in \mathcal{C}$ that is shared by all domains (i.e. domain-irrelevant) and the style part $s_i \in \mathcal{S}_i$ that is specific to domain $\mathcal{X}_i$ (i.e. domain-specific). By modeling $\mathcal{S}_j$ as a continuous distribution such as a Gaussian N_j, x_i can be simply translated to domain $\mathcal{X}_j$ by $G(c, s_j)$ where s_j is randomly sampled from N_j and G is a decoder. We further assume G and E are deterministic and mutually inverse, i.e. $E = G^{-1}$ and $G = E^{-1}$. Besides, we assume that c is a high-dimensional feature map while s_i is a low-dimensional vector such that the complex spatial structure of objects can be preserved and the style parts could focus more on the relatively small scale but discriminative domain-specific information.

Different from Huang et al. (2018), we aim to translate not only between two domains but among multiple ones. To this end, we need to model Gaussians of styles for all domains in a common space (not independently in two spaces like Huang et al. (2018)) such that the single decoder G could generate target image based on which Gaussian is sampled. In the multi-domain and multimodal

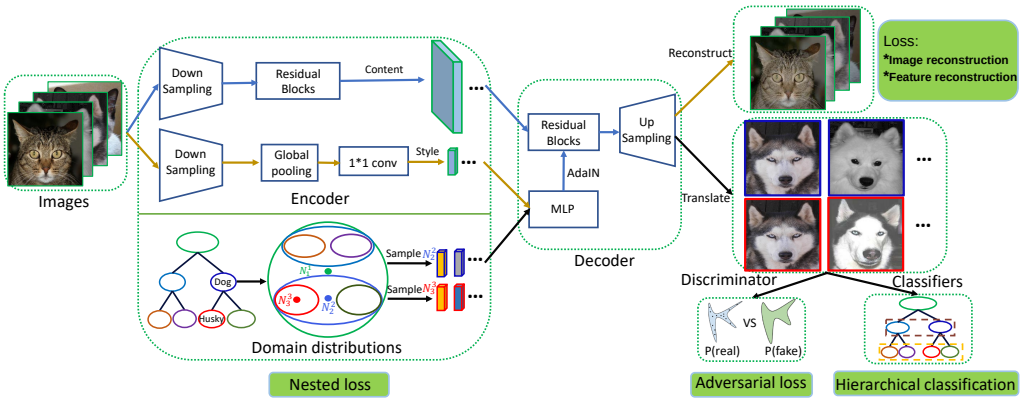


Figure 2: Overview of the whole framework of the proposed method, which mainly consists of five modules: an encoder, a domain distributions modeling module, a decoder, a discriminator and a hierarchical classifier. Given images from different categories, the encoder extracts domain-irrelevant and domain-specific features respectively from the content and style branches. Then the decoder takes them as input to reconstruct the inputs supervised by the reconstruction losses. To realize the multimodal and multi-domain translation, domain distributions are modeled in a common space based on the semantic hierarchy structure and elaborately designed nested loss. Combining the domain-irrelevant features and sampled styles from any distribution, the decoder could translate them to the target domain, guided by the adversarial loss and hierarchical classification loss.

settings, it is noted that categories have the hierarchical relationships. As we introduced in Fig.1, multi-domain translation is in the horizontal direction among categories in a particular hierarchy level, and multimodal translation is in the vertical direction since samples can be further divided into multiple child modes. Therefore, distribution of a parent domain covers several conditional distributions, leading to the nested relationship. In this paper, we model all category domains in a given hierarchy structure as nested Gaussian distributions in a common space, realizing Hierarchical Image-to-image Translation (HIT) between any two domains. In such settings, N_i^l denotes the Gaussian distribution for styles $\mathcal{S}_i^l$ of domain $\mathcal{X}_i^l$ in l-th level ($l = 1, 2, ..., L$).

Fig.2 shows an overview of the proposed HIT method. Our method only contains one pair of encoder and decoder for multi-domain $\mathcal{X}_i^l$. The encoder factorizes x_i into a content code c_i and a style code s_i, i.e. $(c_i, s_i) = E(x_i)$. The decoder can reconstruct them back to the image space via $G(c_i, s_i)$. Image-to-image translation is performed by randomly sampling style codes s_j^l from a domain distribution N_j^l and then using G to output the target image $x_{i\to j}^l = G(c_i, s_j^l)$. The framework is trained with adversarial loss that ensures the translated images approximate the manifold of real images, hierarchical cross-entropy loss that makes the generation conditioned on the sampled domain, nested loss that constrains distributions of domains to satisfy their hierarchical relationships, as well as bidirectional reconstruction losses that ensure enough and meaningful information be encoded.

3.2 Nested distribution loss

In math, the relation between a parent node u and a child node v in the hierarchy is called partial order relation Vendrov et al. (2016), defined as $v \preceq u$. In the application of taxonomy, for concept u and v, $v \preceq u$ means every instance of category v is also an instance of category u, but not vise versa. We call such partial order on probability densities as the notion of nested (encapsulation called by Athiwaratkun & Wilson (2018)). Let g and f be the densities of u and v respectively, if $v \preceq u$, then $f \preceq g$, i.e. f is nested in g. Quantitatively measuring the loss violate the nested relation between f and g is not easy. According to the definition of partial order, strictly measuring that can be done as:

$$\{x : f(x) > \eta\} - \{x : g(x) > \eta\} \tag{1}$$

where $\{x : f(x) > \eta\}$ is the set where f is greater than a nonnegative threshold η. Threshold η indicates the nested degree required by us. Small value of η means high requirement for the overlap between f and g to satisfy $f \preceq g$. Eqn.(1) describes how many regions with densities greater than η of f are not nested in those of g.

Eqn.(1) is difficult to be computed for most distributions including Gaussians. Inspired by the work in word embedding Athiwaratkun & Wilson (2018), we turn to use a thresholded divergence:

$$d_\alpha(f, g) = max(0, D(f||g) - \alpha) \tag{2}$$

where $D(\cdot||\cdot)$ is a divergence measure between densities, we use the KL divergence considering its simple formulation for Gaussians. Such loss is a soft measure of violation of the nested relation. If $f = g$, then $D(f||g) = 0$. In case of $f \preceq g$, $D(f||g)$ would be positive but smaller than a threshold α. As for the effectiveness of using α, please make a reference to Athiwaratkun & Wilson (2018).

To learn the nested distributions for domains in the hierarchy shown in Fig.2, the penalty described by Eqn.(2) between a positive pair of distributions ($N_i \preceq N_j$) should be minimized, while that between a negative pair ($N_i^{'} \npreceq N_j^{'}$) should be greater than a margin m:

$$\mathcal{L}_{nest} = \frac{1}{\mathcal{P}} \sum_{(N_i, N_j) \in \mathcal{P}} d_\alpha(N_i, N_j) + \frac{1}{\mathcal{N}} \sum_{(N_i^{'}, N_j^{'}) \in \mathcal{N}} max\{0, m - d_\alpha(N_i^{'}, N_j^{'})\} \tag{3}$$

where $\mathcal{P}$ and $\mathcal{N}$ denote the numbers of positive and negative pairs respectively.

3.3 Other translation loss functions

Apart from the proposed nested loss in Eqn.(3), our HIT is equipped with an adversarial loss and a hierarchical classification loss to distinguish which domain the generated images belong to, and two general reconstruction losses applied on both images and features.

Adversarial loss. GAN is an effective objective to match the generated images to the real data manifold. The discriminator D tries to classify natural images as real and distinguish generated ones as fake, while the generator G learns to improve image quality to fool D, defined as:

$$\begin{aligned} \mathcal{L}_{GAN}(D) &= \mathbb{E}_{c_i \sim p(E(x_i)), s_j^l \sim N_j^l}[log(D(G(c_i, s_j^l)))] + \mathbb{E}_{x_i \sim p(x)}[1 - log(D(x_i))] \\ \mathcal{L}_{GAN}(E, N, G) &= \mathbb{E}_{c_i \sim p(E(x_i)), s_j^l \sim N_j^l}[log(1 - D(G(c_i, s_j^l)))] \end{aligned} \tag{4}$$

Hierarchical classification loss. Similar to StarGAN Choi et al. (2018), we introduce an auxiliary classifier D_{cls} on top of D and impose the domain classification loss when optimizing G and D, i.e. using real images to train D_{cls} and generated ones to optimize G with such classification loss. Differently, our classifier is hierarchical. In general, the deeper of categories in the hierarchy, the more difficult to distinguish. To alleviate such problem, the loss is cumulative, i.e. classification loss of l-th level is the summation of losses of all levels above l-th with more than two categories.

$$\begin{aligned} \mathcal{L}_{cls}(D) &= \mathbb{E}_{x_i \sim p(x)}[\sum_{k=1}^{L} -log(D_{cls}(y_i^k | x_i))] \\ \mathcal{L}_{cls}(E, N, G) &= \mathbb{E}_{c_i \sim p(E(x_i)), s_j^l \sim N_j^l}[\sum_{k=1}^{l} -log(D_{cls}(y_j^k | G(c_i, s_j^l)))] \end{aligned} \tag{5}$$

where y_j^k is the label of domain $\mathcal{X}_j$ in k-th level.

Bidirectional reconstruction loss. To ensure meaningful information encoded and inverse between G and E, we encourage reconstruction of both images and latent features.

– **Image reconstruction loss:**

$$\mathcal{L}_{recon}^x = \mathbb{E}_{x_i \sim p(x)}[||G(c_i, s_i) - x_i||_1] \tag{6}$$

– **Feature reconstruction loss:**

$$\begin{aligned} \mathcal{L}_{recon}^c &= \mathbb{E}_{c_i \sim p(E(x_i)), s_j^l \sim N_j^l}[||E(G(c_i, s_j^l)) - c_i||_1] \\ \mathcal{L}_{recon}^s &= \mathbb{E}_{c_i \sim p(E(x_i)), s_j^l \sim N_j^l}[||E(G(c_i, s_j^l)) - s_j^l||_1] \end{aligned} \tag{7}$$

Full objectives. To learn E, G and N_j^l, we need to optimize the following terms:

$$\begin{aligned} \mathcal{L}(E, G, N) = &\mathcal{L}_{GAN}(E, N, G) + \mathcal{L}_{cls}(E, N, G) + \lambda_1 \mathcal{L}_{nest} \\ &+ \lambda_2 \mathcal{L}_{recon}^x + \lambda_3 (\mathcal{L}_{recon}^c + \mathcal{L}_{recon}^s) \end{aligned} \tag{8}$$

where λ_1, λ_2 and λ_3 are loss weights of different terms. D is updated with the following losses:

$$\mathcal{L}(D) = \mathcal{L}_{GAN}(D) + \mathcal{L}_{cls}(D) \tag{9}$$

Table 1: Quantitative evaluation on different datasets. For IS and LPIPS, the higher the better. For FID, the smaller the better.

	CelebA			ImageNet-super			ShapeNet-super		
	IS	FID	LPIPS	IS	FID	LPIPS	IS	FID	LPIPS
StarGAN	2.61	21.50	–	6.30	73.80	–	6.04	83.17	–
MUNIT	2.23	92.55	0.298	3.72	81.00	0.491	4.39	157.48	0.400
Our HIT	2.44	38.71	0.094	5.21	93.41	0.323	4.92	72.83	0.168

4 EXPERIMENTS

4.1 IMPLEMENTATION DETAILS

Network architectures. HIT is implemented with Pytorch platform[2]. Images are resized to 128*128 resolution for all datasets. Design of the backbones follows recent proposed image generation Karras et al. (2018) and translation works Huang et al. (2018). As shown in Fig.2, we add a distribution modeling module where a pair of mean vector and diagonal covariance matrix of Gaussian for each domain is parameterized to learn. More training details are given in the Appendix.

Style adversarial loss. Eqn.(4) and Eqn.(5) match the generated images to the distribution of a target domain. Such loss functions can also be applied on the encoded style codes, i.e. matching s_i (act as fake data) of input images to domain Gaussians N_i^l (act as real data) they belong to. By doing so, it is found that the performance of style transfer between a pair of real images would become better. However, such loss would lead to the training collapse on small scale datasets. Therefore, it is recommended to equip it to our framework on datasets with enough training data.

4.2 DATASETS

We conduct experiments on hierarchical annotated data from CelebA Liu et al. (2015), ImageNet Russakovsky et al. (2015) and ShapeNet Chang et al. (2015). Typical examples are shown in Fig.8, Fig.9 and Fig.10 in Appendix. CelebA provides more than 200K face images with 40 attribute annotations. Following the official train/test protocol and imitating the category hierarchy, we define a hierarchy based on attribute annotations. Specifically, all faces are first clustered into male and female and are further classified according to the age and hair color in the next two levels.

Following Huang et al. (2018), we collect images from 3 super domains including house cats, dogs and big cats of ImageNet. Each super domain contains 4 fine-grained categories, which thus construct in a three-level hierarchy (root is animal). All images split by official train/test protocol are processed by a pre-trained faster-rcnn head detector and then cropped as the inputs for translation.

ShapeNet is constitutive of 51,300 3D models covering 55 common and 205 finer-grained categories. 12 2D images with different poses are obtained for each 3D model. A three-level hierarchy of furniture containing different kinds of tables and sofas are defined. Ratio of train/test split is 4:1.

4.3 EVALUATION METRICS

Frankly speaking, quantitatively evaluating the quality of generated images is not easy. Recent proposed metrics may be fooled by artifacts in some extent. In this paper, we use the Inception Score (IS) Salimans et al. (2016) and Frchet Inception Distance (FID) Heusel et al. (2017) to evaluate the semantics of generated images, and leverage the Learned Perceptual Image Patch Similarity Zhang et al. (2018) (LPIPS) to measure the semantic diversity of generated visual modes.

4.4 COMPARED BASELINES

We mainly compared methods proposed for the objectives of either multi-domain or multimodal translation. Considering the unpaired training settings, the multi-domain method StarGAN Choi et al. (2018) and multimodal method MUNIT Huang et al. (2018) are compared in this paper. Since MUNIT needs to train a model for each pair of domains, it is trained for domain pairs of male/female, young/old and black/golden hair on CelebA, house cat/dog, house cat/big cat and big cat/dog on ImageNet, and sofa/table on ShapeNet, respectively. The average of evaluations on all domain pairs for each dataset is reported. As for StarGAN, it is trained on CelebA as done in its opened source codes. Translations among house cat, dog and big cat domains on ImageNet, and between sofa and table domains on ShapeNet are learned for StarGAN. As comparison, results of our HIT in corresponding domain levels for each dataset are reported.

4.5 RESULTS

Table.1 shows the quantitative comparisons of the proposed HIT with the baselines. Fig.3 shows qualitative results on CelebA. It is observed that StarGAN achieves outstanding image quality espe-

[2]The source codes will be released to the public.

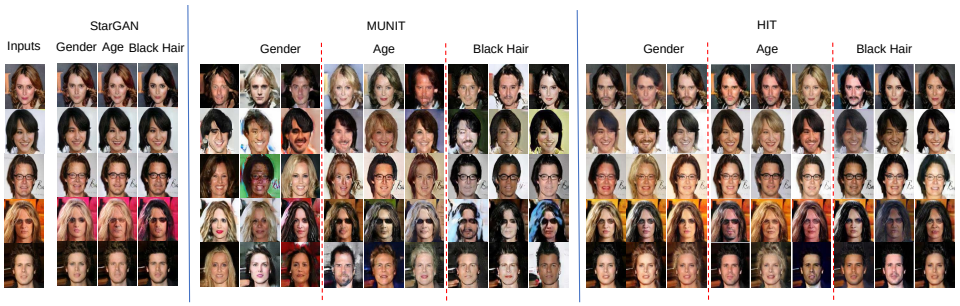


Figure 3: Qualitative comparison on CelebA. The inputs are translated to their reversed value for gender and age attributes, and to black for hair color. StarGAN learns one-to-one mapping. MUNIT and our HIT can generate multimodal results (3 outputs for each input are randomly sampled).

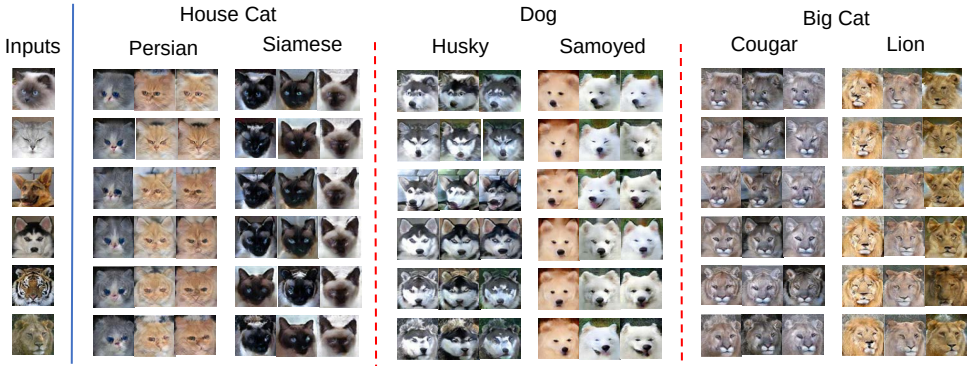


Figure 4: Example results of HIT on ImageNet. For each input, 3 fixed styles are sampled from learned distribution of each category domain.

cially on the fine-grained translations among attribute domains, while the advantages of multimodal methods are generating multiple translations with intra-domain mode variations at the cost of image quality. The image quality of MUNIT is not satisfactory on CelebA both in quantitatively in Table.1 and in qualitatively in Fig.3. The reasons for this may be that using only the adversarial learning to find fine-grained attribute differences between domains is not stable while multi-domain classifier is good at such task. Besides MUNIT obtains the best diversity performance. It is reasonable as it only involves two domains in one model and equips a triplet of backbone including encoder, decoder and discriminator for each domain. Our method considers both multimodal and multi-domain translation within only one triplet of such backbone, which has high requirement for capacity of networks. It performs in trade-off between image quality and diversity. From Fig.3, artifacts accompanying the generated faces for MUNIT may overestimate the LPIPS on CelebA.

Fig.4 and Fig.5 further shows the qualitative results of our HIT on ImageNet and ShapeNet datasets respectively. Generally speaking, translation among such categories with large variations is much more challenging than that for face data (several times of increase of the FID in Table.1 can be found). Even so, our HIT achieves promising qualitative results. Besides, using the fixed styles from a particular category distribution (same columns in Fig.4 and Fig.5), the generated images indeed have similar styles of that category and dissimilar content appearances (e.g. pose, expression), demonstrating good disentanglement of content and style of images.

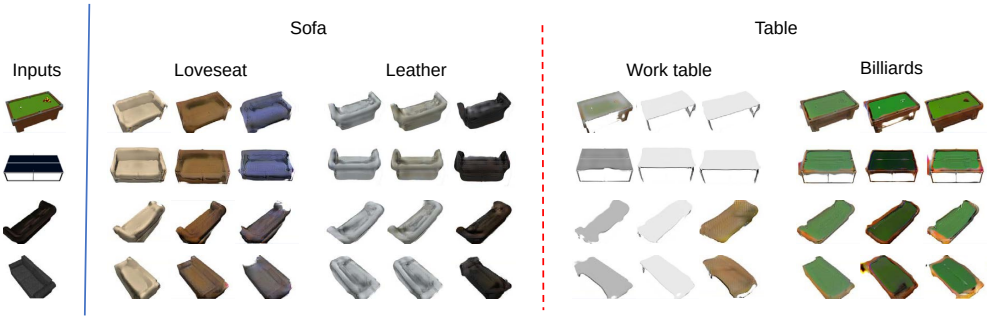


Figure 5: Example results of HIT on ShapeNet. For each input, 3 fixed styles are sampled from learned distribution of each category domain.

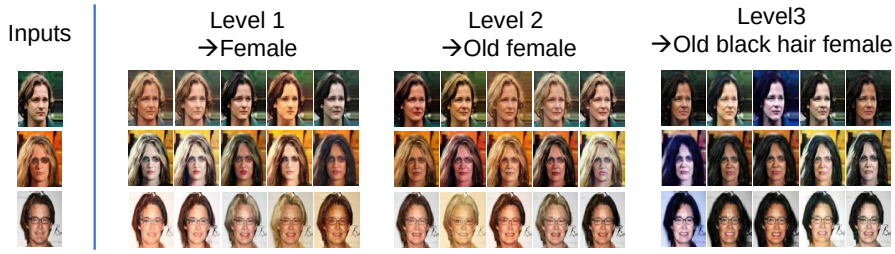
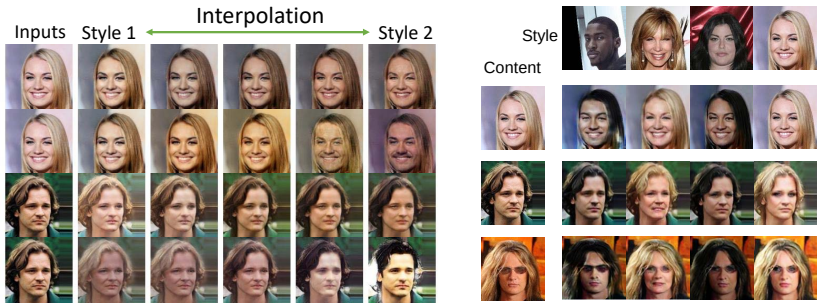


Figure 6: Examples of hierarchical translation. For a target domain in a particular level, 5 styles are sampled from its distribution. With level becoming deeper, translations become more specific.



(a) Style interpolation (b) Style transfer

Figure 7: Translations using interpolations of sampled styles from different domain distributions (a) and style transfer between two real images (b).

Fig.6 shows examples of translations in different hierarchy levels. In the first level, images are divided into male and female domains. Sampling styles from female distribution, translated images may contain the mode variations in the second and third levels (i.e. age and hair color variations). Walking in the path towards leaf-level, translated images would have fewer variations with more conditions being specified by the categories in high levels. In other words, distributions in low levels are local modes of its ancestor domains in high levels, leading to the nested relationship. Results in Fig.6 validate the learned distributions of styles in different levels are exactly nested. In the Appendix, we give an experimental parameter-sensitiveness analysis of m and α which constrain the nested relationships among distributions.

In Fig.7(a), we further study the smoothness of learned distributions. It is observed one can conduct smooth translation via interpolations between styles from different attribute domains. Besides, with the help of additional style adversarial loss introduced in Sec.4.1, our method can provide users more controlled translation as done in Huang et al. (2018), i.e. use the styles of referenced real images instead of sampling them from distributions. Fig.7(b) shows some example results. We can find that the semantics of gender, age and hair color are all correctly transferred to the input images.

5 Discussions

In this paper we propose the Hierarchical Image-to-image Translation (HIT) method which incorporates multi-domain and multimodal translation into one model. Experiments on three datasets especially on CelebA show that the proposed method can well achieve such granularity controlled translation objectives, i.e. the variation modes of outputs can be specified owe to the nested distributions. However, current work has a limitation, i.e. the assumption of single Gaussian for each category domain. On one hand, though Gaussian distribution prior is a good approximation for many data, it may not be applicable when scale of available training data is small but variations within domain are large such as the used hierarchical data on ImageNet and ShapeNet in this paper. On the other hand, the parent distributions should be mixture of Gaussians given multiple single Gaussians of its children. This issue would lead to sparse sampling around the centers of parent distributions and poor nested results if samples are not enough to fulfill the whole space. We have made efforts to the idea of mixture of Gaussians and 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 mixture of Gaussians. A better assumption to realize the nested relationships among parent-children distributions is a promising direction for our future research.

REFERENCES

Karim Ahmed, Mohammad Haris Baig, and Lorenzo Torresani. Network of experts for large-scale image categorization. In *ECCV*, pp. 516–532, 2016.

Martín Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. *CoRR*, abs/1701.07875, 2017.

Ben Athiwaratkun and Andrew Gordon Wilson. Hierarchical density order embeddings. In *ICLR*, 2018.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *ICLR*, 2019.

Angel X. Chang, Thomas A. Funkhouser, Leonidas J. Guibas, Pat Hanrahan, Qi-Xing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. Shapenet: An information-rich 3d model repository. *CoRR*, abs/1512.03012, 2015.

Ying-Cong Chen, Xiaogang Xu, Zhuotao Tian, and Jiaya Jia. Homomorphic latent space interpolation for unpaired image-to-image translation. In *CVPR*, 2019.

Yunjey Choi, Min-Je Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. S-targan: Unified generative adversarial networks for multi-domain image-to-image translation. In *IEEE,CVPR*, pp. 8789–8797, 2018.

Jia Deng, Jonathan Krause, Alexander C. Berg, and Fei-Fei Li. Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition. In *IEEE,CVPR*, pp. 3450–3457, 2012.

Jia Deng, Nan Ding, Yangqing Jia, Andrea Frome, Kevin Murphy, Samy Bengio, Yuan Li, Hartmut Neven, and Hartwig Adam. Large-scale object classification using label relation graphs. In *ECCV*, pp. 48–64, 2014.

Emily L. Denton, Soumith Chintala, Arthur Szlam, and Rob Fergus. Deep generative image models using a laplacian pyramid of adversarial networks. In *NIPS*, pp. 1486–1494, 2015.

Nan Ding, Jia Deng, Kevin P. Murphy, and Hartmut Neven. Probabilistic label relation graphs with ising models. In *IEEE, ICCV*, pp. 1161–1169, 2015.

Wonjoon Goo, Juyong Kim, Gunhee Kim, and Sung Ju Hwang. Taxonomy-regularized semantic deep convolutional neural networks. In *ECCV*, pp. 86–101, 2016.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, pp. 2672–2680, 2014.

Gregory Griffin and Pietro Perona. Learning and using taxonomies for fast visual categorization. In *IEEE,CVPR*, 2008.

Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C. Courville. Improved training of wasserstein gans. In *NIPS*, pp. 5767–5777, 2017.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, pp. 6626–6637, 2017.

Xun Huang, Ming-Yu Liu, Serge J. Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, pp. 179–196, 2018.

Sung Ju Hwang and Leonid Sigal. A unified semantic embedding: Relating taxonomies and attributes. In *NIPS*, pp. 271–279, 2014.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *IEEE,CVPR*, pp. 5967–5976, 2017.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *CoRR*, abs/1710.10196, 2017.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *CoRR*, abs/1812.04948, 2018.

Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *ICML*, pp. 1857–1865, 2017.

Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew P. Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *IEEE,CVPR*, pp. 105–114, 2017.

Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *ECCV*, pp. 36–52, 2018.

Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *NIPS*, pp. 700–708, 2017.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *IEEE ,ICCV*, pp. 3730–3738, 2015.

Xudong Mao, Qing Li, Haoran Xie, Raymond Y. K. Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *IEEE , ICCV*, pp. 2813–2821, 2017.

Marcin Marszalek and Cordelia Schmid. Constructing category hierarchies for visual recognition. In *ECCV*, pp. 479–491, 2008.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *ICLR*, 2018.

Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *ICML*, pp. 2642–2651, 2017.

Vicente Ordonez, Jia Deng, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. From large scale image categorization to entry-level categories. In *IEEE, ICCV*, pp. 2768–2775, 2013.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016.

Scott E. Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *ICML*, pp. 1060–1069, 2016.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NIPS*, pp. 2226–2234, 2016.

Krishna Kumar Singh, Utkarsh Ojha, and Yong Jae Lee. Finegan: Unsupervised hierarchical disentanglement for fine-grained object generation and discovery. In *CVPR*, 2019.

Nitish Srivastava and Ruslan Salakhutdinov. Discriminative transfer learning with tree-based priors. In *NIPS*, pp. 2094–2102, 2013.

Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. In *ICLR*, 2017.

Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. In *ICLR*, 2016.

Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *NIPS*, pp. 82–90, 2016.

Jianwen Xie, Yifei Xu, Erik Nijkamp, Ying Nian Wu, and Song-Chun Zhu. Generative hierarchical learning of sparse FRAME models. In *IEEE, CVPR*, pp. 1933–1941, 2017.

Zhicheng Yan, Hao Zhang, Robinson Piramuthu, Vignesh Jagadeesh, Dennis DeCoste, Wei Di, and Yizhou Yu. HD-CNN: hierarchical deep convolutional neural networks for large scale visual recognition. In *IEEE, ICCV*, pp. 2740–2748, 2015.

Zili Yi, Hao (Richard) Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *IEEE , ICCV*, pp. 2868–2876, 2017.

Han Zhang, Tao Xu, and Hongsheng Li. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *IEEE, ICCV*, pp. 5908–5916, 2017.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE,CVPR*, pp. 586–595, 2018.

Bin Zhao, Fei-Fei Li, and Eric P. Xing. Large-scale category structure aware image categorization. In *NIPS*, pp. 1251–1259, 2011.

Shengjia Zhao, Jiaming Song, and Stefano Ermon. Learning hierarchical features from deep generative models. In *ICML*, pp. 4091–4099, 2017.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE , ICCV*, pp. 2242–2251, 2017a.

Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A. Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *NIPS*, pp. 465–476, 2017b.

A Appendix

A.1 Network architectures and training details

A.1.1 Network Architectures

Following the backbone designs in Huang et al. (2018) 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 covariance matrix for each Gaussian are also parameterized with 8 dimensions to be optimized with 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 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 Mao et al. (2017) 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

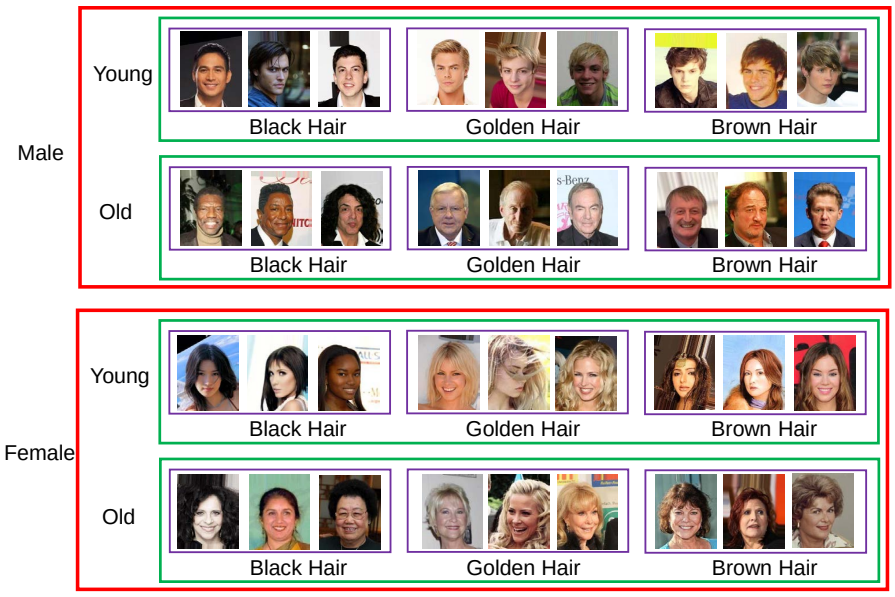


Figure 8: Typical samples of hierarchical data on CelebA. Images within a purple rectangular box are some instances of a leaf-level category. Categories within a green rectangular box belong to one common super-category. The super-categories within a red rectangular box share one common ancestor.

A.1.2 Training Hyperparameters

We use the Adam optimizer with $\beta_1 = 0.5$, $\beta_2 = 0.999$, and initial learning rate of 0.0001. We train HIT on all datasets for 300K iterations and half decay the learning rate every 100K iterations. We set batch size to 16. The loss weights λ_1, λ_2 andλ_3 in Eqn.(8) are set as 1, 10 and 1 respectively. α and m in Eqn.(3) are empirically set as 50, 200 respectively. Random mirroring is applied during training.

A.2 Hierarchical data construction

In this section, Fig.8, Fig.9 and Fig.10 provide leaf-level examples for better understanding the nested relationships among categories in different hierarchy levels. Take the CelebA for example, the root category *face* has two children distinguished by gender attribute. For each of the two super categories, it includes two finer granular children which are further divided by the age attribute (young/old). Finally, in the leaf-level, each local branch are classified according to their hair colors, i.e. black, golden and brown hair. Within each leaf-level category, samples mainly contain intra-class variations caused by identities, expressions, poses, etc.

A.3 Parameters sensitiveness analysis

The impacts of hyper-parameters in the nested distribution learning on word embedding task have been studied in Athiwaratkun & Wilson (2018). In this section, we further make an analysis of them in current image generation task. Fig.11 and Fig.12 show the impacts of hyper-parameters m and α in the nested loss of Eqn.(3). It is observed that distribution margin m has larger impact than nested threshold α. With too large settings of m, distributions which do not have nested relationship would be pushed far away, leading to sparse space. Sampling in such space would make the learning of generator quite difficult. In contrast, with too small settings of m, the discriminabilities of distributions may be poor. Therefore, a trade-off value 200 is set for m in this paper. As for nested threshold α, a relative small or large value performs well in terms of the image quality metric. However, in theory, large value setting of α would relax the nested constraint too much, result in small overlap between parent and children distributions. Therefore, we recommend to set α in the left half axis of

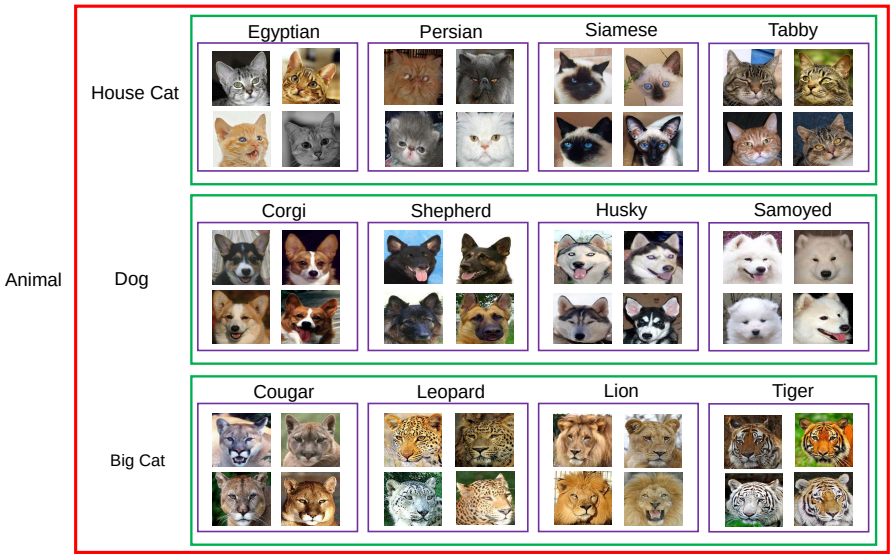


Figure 9: Typical samples of hierarchical data on ImageNet. Images within a purple rectangular box are some instances of a leaf-level category. Categories within a green rectangular box belong to one common super-category. The super-categories within a red rectangular box share one common ancestor.

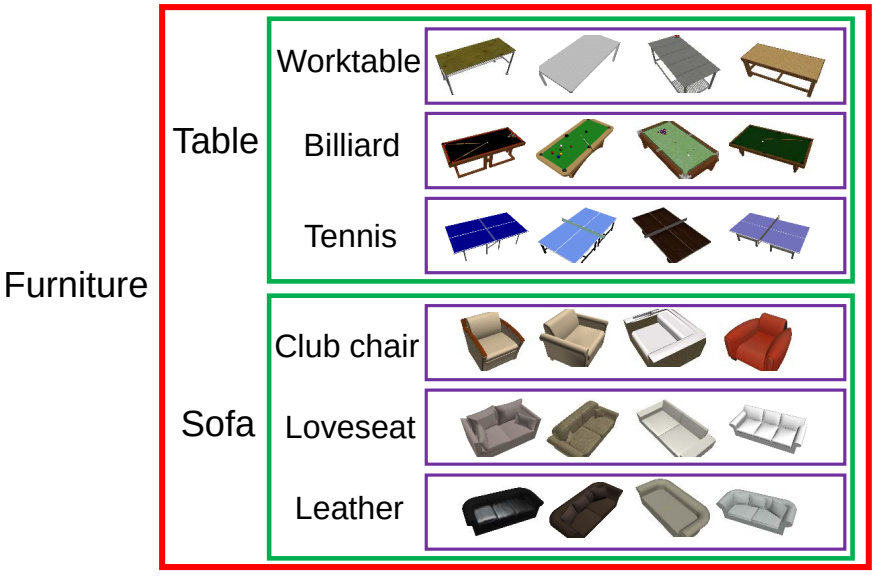


Figure 10: Typical samples of hierarchical data on ShapeNet. Images within a purple rectangular box are some instances of a leaf-level category. Categories within a green rectangular box belong to one common super-category. The super-categories within a red rectangular box share one common ancestor.

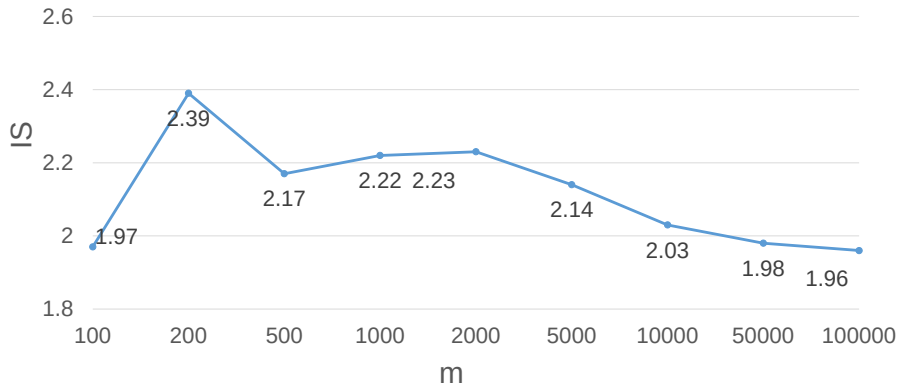


Figure 11: The Inception Score (IS) of translated images in leaf-level on CelebA with different distribution margin m and fixed threshold $\alpha = 50$.

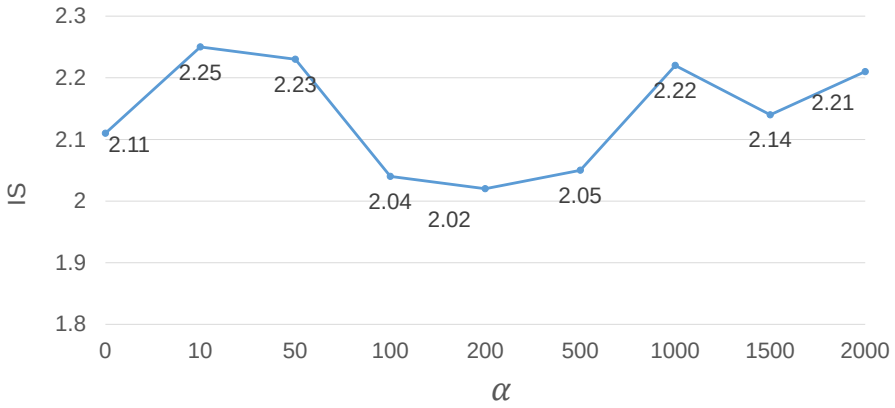


Figure 12: The Inception Score (IS) of translated images in leaf-level on CelebA with different nested threshold α and fixed distribution margin $m = 2000$.

α. When α is set as 0, it means the parent and children distributions are all overlapped, which is too strict to learn. Finally, we set α as 50, and the ratio of 1:4 between α and m is consistent with the observation in Athiwaratkun & Wilson (2018).