Pattern Recognition 146 (2024) 110058

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

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

ARTICLE INFO

Keywords:
Image-to-image translation
Distribution modeling
Information entropy
Generative adversarial network

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 multi-modal (i.e. many-to-many mapping) due to variations of domain-specific information (e.g., the domain of house cat contains multiple sub-modes), which is usually addressed by predefined distribution sampling. For another, most existing multi-modal approaches have limits in handling more than two domains with one model, i.e. they have to independently build two distributions to capture variations for every pair of domains. To address these problems, we propose a Hierarchical Image-to-image Translation (HIT) method which jointly formulates the multi-domain and multi-modal problem in a semantic hierarchy structure by modeling a common and nested distribution space. Specifically, domains have inclusion relationships under a particular hierarchy structure. With the assumption of Gaussian prior for domains, distributions of domains at lower levels capture the local variations of their ancestors at higher levels, leading to the so-called nested distributions. To this end, we propose a nested distribution loss in light of the distribution divergence measurement and information entropy theory to characterize the aforementioned inclusion relations among domain distributions. Experiments on ImageNet, ShapeNet, and CelebA datasets validate the promising results of our HIT against state-of-the-arts, and as additional benefits of nested modeling, one can even control the uncertainty of multi-modal translations at different hierarchy levels.

1. Introduction

Image-to-image translation is the process of mapping images from one domain to another, during which change the domain-specific aspect and preserve the domain-irrelevant information [1]. It has wide applications in computer vision and computer graphics [2–9] such as mapping photographs to edges/segments, colorization, super-resolution, inpainting, attribute/category transfer, style transfer, etc. In this work, we focus on the task of category transfer [4,5,10], i.e. images sharing the same category label belong to one domain.

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) [11,12]. Representative methods include pix2pix [2], UNIT [13], CycleGAN [4], DiscoGAN [14], DualGAN [14] and DTN [15]. More recently the study of this task mainly focuses on two challenges. The first is the ability of involving translation among multiple (more than just two) domains into one model. It is quite a practical need for users. Most existing works have to train a separate model for each pair of domains, which is obviously inefficient. To deal with such problem, StarGAN [16] and AttGAN [17] leverage 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 multi-modal problem, which is early addressed by BicycleGAN [18]. Most techniques including the recent StarGAN can only yield a single determinate output in the target domain given a source image as input. However, for many translation tasks, mappings are naturally multi-modal (i.e. many-to-many). As shown in Fig. 1, when translating a *cat* (i.e. the source domain) to the dog category (i.e. the target domain), the target output actually could have many possible appearances, such as becoming a *Husky*, a *Samoyed*, or other specific dog breeds. To address this issue, most recent works including BicycleGAN [18], MUNIT [5] and DRIT [10] model a continuous and multivariant distribution independently for each domain to represent the variations of domain-specific information, and

* Corresponding author at: Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China.
E-mail address: wangruiping@ict.ac.cn (R. Wang).

https://doi.org/10.1016/j.patcog.2023.110058
Received 8 April 2022; Received in revised form 12 October 2023; Accepted 13 October 2023
Available online 16 October 2023
0031-3203/

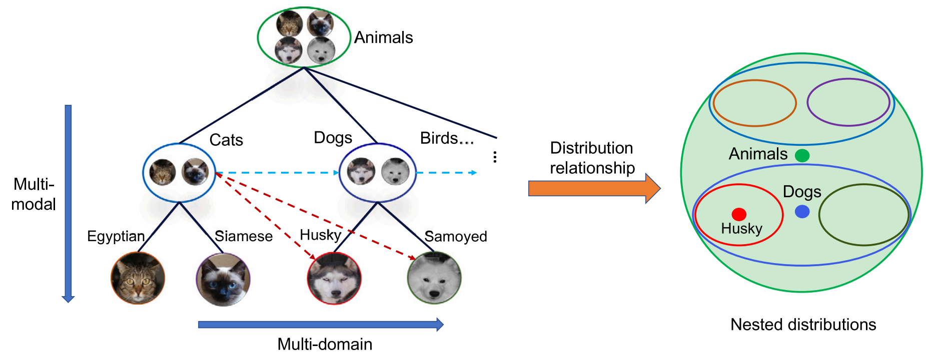


Fig. 1. An illustration of a hierarchy structure and the distribution relationship of categories in a 2D space. The multi-domain issue is shown in the horizontal direction (blue dashed arrow) while the multi-modal issue is indicated in the vertical direction (red dashed arrow). Since one child category is a special case of its parent, in the distribution space it is a conditional distribution of its parent, leading to the nested relationship. Best viewed in colors. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

have achieved diverse and high-quality results for several two-domain translation tasks, yet leaving space to multi-domain translation.

In this paper, we aim at involving the abilities of both multi-domain and multi-modal translation into one model. As shown in Fig. 1, it is noted that categories have 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 *husky* and some of them are called *samoyed* due to the appearance variations of being *dog*. Of course, one can continue to divide *husky* to be finer-grained categories based on the variations of certain visual attributes. Such hierarchical relationships widely exist among categories in the real world since it is a natural way for our human to understand objects according to our needs [19–22]. With such findings, when we review again the image translation task, the multi-domain and multi-modal issues can be understood from two orthogonal views. From the horizontal view as indicated by the blue dashed arrow, multi-domain translation is the mapping among domain-specific variations of categories. From the vertical view (the red dashed arrow), multi-modal translation further divides such domain-specific variations into some more specific and local sub-modes within each category domain. Such sub-modes at low levels are the local subspaces of the holistic variation space at high levels.

Inspired by the above observations, we propose a Hierarchical Image-to-image Translation (HIT) method which jointly formulates the multi-domain and multi-modal categorical translation problem in a semantic hierarchy structure. Specifically, our method models domain-specific variations of categories in the form of multiple continuous and multivariant Gaussian distributions in a common space. Such distribution modeling is inherently different from previous methods whose domain distributions are the same Gaussian. Therefore, their frameworks either needs multiple encoder–decoder pairs [5,10,18], or other elaborately designed auxiliary network modules [23,24] to realize the translation to multiple domains, which is memory inefficient when deployed. As for our method, only using one encoder–decoder can achieve this just by sampling from different categorical distributions in the common space. To further ensure the diversity (i.e. the multi-modal goal) of each category distribution in the common space, we consider the hierarchical inclusion relationship among categories, i.e. explicitly divide the distribution of one category into several more specific and local sub-distributions, leading to the nested distributions as conceptually shown in the 2D illustration in Fig. 1. To this end, we propose a novel nested distribution loss by resorting to the theory of distribution divergence and information entropy. On one hand, the divergence of distributions with nested relation (e.g. *husky* and *dog*) should be smaller than a threshold while that between other pair of distributions should be larger than a margin. On the other hand, the uncertainty of semantics when sampling at higher hierarchy levels (i.e more global distribution space) should be larger than that at lower levels (i.e. more local and specific area in the space), which can be characterized by the information entropy measurement, (e.g. variations of the *dog* are larger than one of its children *husky*, result in larger entropy of the former than the latter when sampling). Combining the nested distributions modeling with the conditional GAN framework, our HIT achieves multi-domain, multi-modal, and even multi-granularity translation abilities. Experiments on challenging ImageNet, ShapeNet, and CelebA datasets validate the promising performance of our method.

The rest of this paper is organized as follows. Section 2 reviews and discusses the progress of relevant research directions. Section 3 details the proposed method, followed by the experiments and results in Section 4. The discussion is presented in Section 5. Finally, our conclusion is drawn in Section 6.

2. Related works

Conditional Generative Adversarial Networks. GAN [11] is probably one of the most creative frameworks in the last decades 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 [25–29], more reasonable distribution metrics [30–32], and normalization [33,34]. With these improvements, GANs have been applied to many conditional tasks [12], such as image generation given class labels [35] or styles of real images [36], super-resolution [3], image dehazing [37], text2image [38], 3D reconstruction from 2D input [39], image manupulation/editing [40,41] and image-to-image translation introduced below.

Image-to-image Translation. Pix2pix [2] 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 between two domains. To address this issue, unpaired methods are proposed including UNIT [13], DiscoGAN [14], DualGAN [42], CycleGAN [4], AsymGAN [43] and DTCDN [44]. UNIT combines the variational 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, CycleGAN, AsymGAN and DTCDN all leverage a cycle consistency loss which enforces that one can re-translate the target image back to the original image. More recently, TUNIT [45–47] address the complete unsupervised translation setting without domain labels by clustering or contrastive learning.

Recent works mainly focus on the issues of multi-domain and multi-modal. To deal with multi-domain translation in one generator, StarGAN [16] and AttGAN [17] take target label and input image as

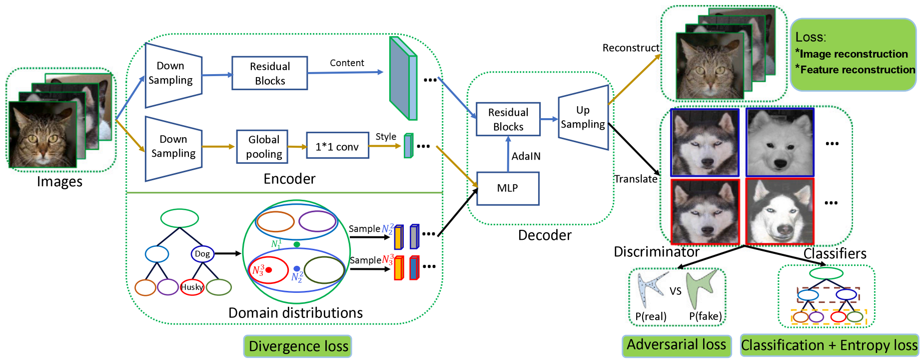


Fig. 2. Overview of our framework, which consists of five modules: an encoder, a distributions modeling module, a decoder, a discriminator, and a hierarchical classifier. Given images from different categories, the encoder extracts domain-irrelevant and domain-specific features respectively from the content and style branches. Then the decoder takes them as input to reconstruct the inputs supervised by the reconstruction losses. To realize the multi-modal and multi-domain translation, domain distributions are modeled in a common space based on the semantic hierarchy structure and elaborately designed nested loss including divergence and entropy constraints. Combining the domain-irrelevant features and sampled styles from the distribution (e.g., N_1^1, N_2^2 or N_3^3), the decoder translates them to the target domain, guided by the adversarial loss and hierarchical classification loss. Best viewed in colors and zoom-in. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

conditions, and uses an auxiliary classifier to classify translated image. As for the multi-modal issue, BicycleGAN [18] proposes to model continuous and multivariant distributions. However, it requires input–output pairwise annotations. To overcome this problem, MUNIT [5] and DRIT [10] adopt a disentangled representation for learning diverse translation results from unpaired training data. [48] proposes to interpolate the latent codes between input and referred image to realize diverse generations. [6] introduces a diversity objective by encouraging the distance among multiple outputs. DMIT [23], StarGAN v2 [24] and i-StyleGAN [49] combine the advantage of StarGAN and MUNIT, i.e. fusing the target label and styles sampled from a shared distribution to realize both multi-domain and multi-modal translation. They assume all domains share the same distribution of variations, which may not be reasonable especially for categories whose distribution structures are obviously different. To address such issue, GMM-UNIT [50] proposes to fit domains to a Gaussian mix distribution, where each component is associated to a domain. We also attempt to model multi-domain into one common distribution space, the key difference is that we leverage the natural hierarchy relationships to constrain the distributions space to be nested, resulting in both multi-domain, multi-modal, and granularity-controlled translations.

Hierarchy-regularized Learning. Hierarchical learning is a natural learning manner for humans and we describe objects in the world from abstract to detailed according to our needs. For machine learning and computer vision, such semantic hierarchies have been widely explored in object classification for accelerating recognition [19,51], obtaining multiple granularities of predictions [20,22], making use of category relation graphs [52,53], and improving recognition accuracy as additional supervision [21,54–58]. Apart from these discriminative tasks, [59,60] propose to use generative models to disentangle the factors from low-level to high-level representations that can construct an object. [61] uses an unsupervised generative framework to hierarchically disentangle the background, object shape, and appearance from an image. In natural language processing, [62] proposes a probabilistic word embedding method to capture the semantics described by the WordNet hierarchy. Our method first introduces such semantic hierarchy to tackle the challenging domain distributions modeling problem in the multi-modal and multi-domain image translation task with the novel nested distribution loss.

3. Approach

3.1. Problem formulation

Let x_i be a natural image from domain $\mathcal{X}_i$. The goal of translation between two category domains 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 $\mathcal{X}_j$. We assume that x_i can be disentangled by the encoder E into the content part $c \in C$ that is shared by all domains (i.e. domain-irrelevant information) and the style part $s_i \in \mathcal{S}_i$ that is specific to domain $\mathcal{X}_i$ (i.e. domain-specific variations). As discussed in Section 1, image-to-image translation is the mapping between domain-specific variations of categories, and such mapping is usually multi-modal (i.e. many-to-many mapping). By modeling $\mathcal{S}_j$ as a continuous and learnable Gaussian distribution 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.

In this paper, we aim to efficiently translate among multiple domains with only one pair of (E, G). To this end, we propose to model Gaussians of $\mathcal{S}$ for domains in a common space such that the single decoder G could generate a target image based on which Gaussian is sampled from. Furthermore, as introduced in Fig. 1, a coarse category domain can be divided into several fine-grained domains, leading to the hierarchy structure among domains and nested Gaussian distributions in the common space. Formally speaking, the concept of hierarchical level l is introduced for the domain $\mathcal{X}_i^l$ ($l = 1, 2, \dots, L$ and $i = 1, 2, \dots, C_l$, where C_l is the number of categories at the lth level). N_i^l denotes the Gaussian distribution for styles S_i^l of domain $\mathcal{X}_i^l$. With such hierarchical distributions, the goal of multi-modal translation could be more specific and granularity controlled, i.e. one can explicitly set which sub-mode and which hierarchical level of the target domain the input image will be translated to (e.g. to the *husky* or the *samoyed* instead of just the ambiguous *dog* domain).

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

3.2. Nested distribution loss

The common space of domain distributions is constrained in terms of two aspects. For one thing, the nested relationship is directly characterized by the distribution divergence. For another, the uncertainty of semantic information of generated images is aligned with the hierarchy levels of the target domains by the information entropy.

Divergence Loss. In math, the relation between a parent node u and a child node v in the hierarchy is called partial order relation [63], 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. 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 of violating 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. Eq. (1) describes the disjoint support set between f and g, given a density threshold η, i.e. how many regions with densities greater than η of f are not nested in those of g. Therefore, η indicates the nested degree required by us. Small value of η means high requirement for the overlap between f and g to satisfy $f \preceq g$ in Eq. (1).

We aim to end-to-end optimize the framework. Unfortunately, Eq. (1) is difficult to be computed with differentiable formulation for most distributions like widely used Gaussians. Inspired by the work in word embedding [62], we turn to use a thresholded divergence:

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

where $D(\cdot \parallel \cdot)$ is a divergence measurement between densities. We use the KL divergence which describes the loss of using g to fit f, considering its simple formulation for Gaussians. Such loss is a soft measure of violation of the nested relation. If and only if $f = g$, then $D(f \parallel g) = 0$. In case of $f \preceq g$, $D(f \parallel g)$ would be positive but not too larger than a threshold α. The threshold α is a necessary relax term. Assuming multiple f nested in g, directly minimizing $KL(f \parallel g)$ will lead to all f concentrating to the center of g, which is not desired.

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

$$\begin{aligned}\mathcal{L}_{dvg} =& \frac{1}{\mathcal{P}} \sum_{(N_i^l, N_j^k)\in\mathcal{P}} d_\alpha(N_i^l, N_j^k) \\ &+ \frac{1}{\mathcal{N}} \sum_{(N_i'^l, N_j'^k)\in\mathcal{N}} max\{0, m - d_\alpha(N_i'^l, N_j'^k)\}\end{aligned} \tag{3}$$

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

Entropy Loss. Under the hierarchy, a sample from one non-leaf distribution can be located in any one of its nested sub-distributions. To be more specific, a particular sample from N_j^k (e.g. the *dog*) is determinately located in one of its child domain (e.g. the *husky*), but plenty of such sampling would be uncertainly located in every child domain (e.g. half to half in the *husky* and *samoyed*). In this paper, we leverage the probability information entropy loss to capture such semantic certainty and uncertainty.

We introduce an auxiliary hierarchical classifier D_{cls} sharing the same backbone with the discriminator D_{dis}, which outputs normalized categorical predictions at different levels for a translated image. Assuming s_j^k of the generated image is sampled from a non-leaf target domain distribution N_j^k, and p_i^l is the prediction for such image on its encapsulated domain $\mathcal{X}_i^l$ $(N_i^l \preceq N_j^k, k < l <= L)$. The entropy loss can thus be formulated as:

$$\begin{aligned}\mathcal{L}_{ent} =& [\mathbb{E}_{s_j^k\sim N_j^k} \sum_{l=k+1}^{L} \sum_{i=1,N_i^l\preceq N_j^k}^{C_l} -p_i^l log^{p_i^l}] \\ &+ [\sum_{l=k+1}^{L} \sum_{i=1,N_i^l\preceq N_j^k}^{C_l} \bar{p}_i^l log^{\bar{p}_i^l}]\end{aligned} \tag{4}$$

where $\bar{p}_i^l$ is the average prediction for multiple translated images sampled from the same N_j^k. The first term in Eq. (4) constrains a particular sample from N_j^k can be definitely predicted as one of its children (small entropy), and the second means one cannot always predict every sample from N_j^k as its same child (large entropy). Combined with Eq. (3), the nested distribution loss is:

$$\mathcal{L}_{nest} = \mathcal{L}_{dvg} + \mathcal{L}_{ent} \tag{5}$$

3.3. Other translation loss functions

Apart from the proposed nested loss in Eqs. (3) and (4), 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_{dis} 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_{dis}, defined as:

$$\begin{aligned}\mathcal{L}_{GAN}(D_{dis}) =& \mathbb{E}_{c\sim p(E(x_i^l)),s_j^k\sim N_j^k}[log(D_{dis}(G(c,s_j^k)))] \\ &+ \mathbb{E}_{x_i^l\sim p(x)}[1 - log(D_{dis}(x_i^l))] \\ \mathcal{L}_{GAN}(E,N,G) =& \\ &\mathbb{E}_{c\sim p(E(x_i^l)),s_j^k\sim N_j^k}[log(1 - D_{dis}(G(c,s_j^k)))]\end{aligned} \tag{6}$$

Hierarchical Classification Loss. We impose hierarchical domain classification loss when optimizing G and D_{cls}, i.e. using real images to train D_{cls} and generated ones to optimize G. In general, the deeper of category levels in the hierarchy, the more difficult it to distinguish. To alleviate such problem, the loss is cumulative, i.e. classification loss of images at the kth level is the summation of losses of all levels above k (e.g. a *husky* should be classified as a *dog*, an *animal* at high levels). Note that this is different from Eq. (4) which is used for estimating the classification uncertainty below current levels while Eq. (7) is used for classification with real category labels above (and including) current levels.

$$\begin{aligned}&\mathcal{L}_{cls}(D_{cls}) = \mathbb{E}_{x_i^l\sim p(x)}[\sum_{l=1}^{L} -log(D_{cls}(y_i^l|x_i^l))] \\ &\mathcal{L}_{cls}(E,N,G) = \\ &\mathbb{E}_{c\sim p(E(x_i^l)),s_j^k\sim N_j^k}[\sum_{l=1}^{k} -log(D_{cls}(y_j^l|G(c,s_j^k)))]\end{aligned} \tag{7}$$

where y_i^l is the category label of x_i at the lth level.

Bidirectional Reconstruction Loss. To ensure meaningful information encoded and inverse between G and E, we encourage the net to reconstruct both images and latent features.

- **Image reconstruction loss:**

$$\mathcal{L}_{recon}^x = \mathbb{E}_{x_i^l\sim p(x)}[\|G(c,s_i^l) - x_i^l\|_1] \tag{8}$$

- **Feature reconstruction loss:**

$$\begin{aligned}\mathcal{L}_{recon}^c &= \mathbb{E}_{c\sim p(E(x_i^l)),s_j^k\sim N_j^k}[\|E(G(c,s_j^k)) - c\|_1] \\ \mathcal{L}_{recon}^s &= \mathbb{E}_{c\sim p(E(x_i^l)),s_j^k\sim N_j^k}[\|E(G(c,s_j^k)) - s_j^k\|_1]\end{aligned} \tag{9}$$

Full Objectives. To learn E, G and N, we need to optimize the following terms:

$$\begin{aligned}\mathcal{L}(E,G,N) =& \mathcal{L}_{GAN}(E,N,G) + \mathcal{L}_{cls}(E,N,G) + \lambda_1\mathcal{L}_{nest} \\ &+ \lambda_2\mathcal{L}_{recon}^x + \lambda_3(\mathcal{L}_{recon}^c + \mathcal{L}_{recon}^s)\end{aligned} \tag{10}$$

where λ_1, λ_2 and λ_3 are loss weights of different terms. D_{dis} and D_{cls} are updated with the following losses:

$$\mathcal{L}(D_{dis}, D_{cls}) = \mathcal{L}_{GAN}(D_{dis}) + \mathcal{L}_{cls}(D_{cls}) \tag{11}$$

3.4. Implementation details

Our HIT is implemented with Pytorch platform.[1] Images are resized to 128*128 resolution for all datasets. The design of the backbones follows recently proposed image generation [34] and translation works [5]. 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. We also equip the residual blocks of G with Adaptive Instance Normalization (AdaIN) whose parameters are dynamically generated by a multi-layer perception (MLP) from the encoded or sampled style code. More network details are given in the supplementary material.

We use Adam optimizer with $\beta_1 = 0.5$, $\beta_2 = 0.999$, and initial learning rate of 0.0001. We train HIT for 500K iterations and half decay the learning rate every 100K iterations. We set batch size to 8. In Eq. (4), 5 samples for each target domain are used to find the average prediction. The loss weights λ_1, λ_2 and λ_3 in Eq. (10) are set as 1, 10 and 1 respectively. α and m in Eq. (3) are empirically set as 50, 200 respectively. Random mirroring is applied during training.

4. Experiments

4.1. Datasets and evaluation metrics

Datasets. We conduct experiments on hierarchical annotated data of ImageNet [64] and ShapeNet [65]. Typical images are shown in the supplementary material. Following [5], we collect animal heads from 3 super domains including *house cat, dog* and *big cat* in ImageNet using the official train/test protocol. Each super domain contains 4 fine-grained categories, which thus construct a three-level hierarchy (root is animal). These images 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. The ratio of train/test split is 4:1.

Evaluation Metrics. Following [23], we use Fréchet Inception Distance (FID) [66] to evaluate the appearance quality of images, and Learned Perceptual Image Patch Similarity [67] (LPIPS) to measure the diversity of visual modes. We also employ 30 users to choose the best translated images from different methods in terms of semantic matching degree with target domains and image quality. The percentage of human preference for each method is reported (more details about the user study can be found in the supplementary materials). Besides, to quantitatively and automatically evaluate the semantic matching degree with target domains, we also finetune the AlexNet classifiers on ImageNet and ShapeNet datasets, and compute the top-1 classification accuracy of the translated images for compared methods. Last but not the least, the number of network parameters of each method is reported to evaluate the memory efficiency.

4.2. Comparisons with state-of-the-arts

We mainly compared methods proposed for the objectives of either multi-domain or multi-modal translation (or both). Considering the unpaired training settings, the multi-domain method StarGAN [16], multi-modal method MUNIT [5], and multi-mapping method DMIT [23] and StarGAN v2 [24] are compared. Since MUNIT needs to train a model for each pair of domains, it is trained for domain pairs of *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 is reported. As for StarGAN, DMIT, and StarGAN v2, translations among *house cat, dog* and *big cat* domains on ImageNet, and between *sofa* and *table* domains on ShapeNet are learned. Important parameters of methods are tuned according to the recommendations by their authors, original references, as well as their opened codes. As comparison, results of our HIT in corresponding domain levels are reported.

Figs. 3 and 4 show qualitative results of translations on ImageNet and ShapeNet respectively. We can draw three main conclusions. (1). Existing multi-domain methods including StarGAN and DMIT do not perform well on the challenging category domain translation task. This is mainly limited by the flat auxiliary classifiers they used to distinguish domains of images or styles. Such supervisions do well in the fine-grained domain transfer tasks such as face attributes editing [16] (results on CelebA [68] dataset are shown in the supplementary materials.), fine-grained text-to-image generation and scene style transfer [23]. When it comes to the categorical translation which requires large variations on object appearance and shapes, the flat domain classifiers may not capture the full semantic difference among categories (i.e. over-fitting to finite category annotations) and thus these methods only make slight textures or colors change to the inputs. (2). As comparison, the two-domain adversarial learning method MUNIT and multi-task adversarial learning method StarGAN v2 can better capture the global distributions of categories, resulting in reasonable semantic changes of objects. (3). Our HIT performs well on this task, though we also adopt a domain classifier. Differently, our classifier is hierarchical which fully leverages the supervisions at different semantic levels and can thus capture domain difference as much as possible. Besides, the proposed entropy loss in Eq. (4) can be regarded as a kind of domain adversarial learning to some extent, i.e. a particular sample from a target domain at a high level should be certainly classified among its children (small entropy) while plenty of sampling from the same domain should distribute evenly among every child (large entropy). In other words, the hierarchical and entropy-based adversarial classification together make our HIT better capture the semantics of categories.

Table 1 shows the quantitative evaluations of image quality (FID), diversity (LPIPS), semantic matching with the target domains (Human and Accuracy), and the parameters scale (#Parameters) of each method. It is observed that the multi-domain methods StarGAN and DMIT generate images with high quality in terms of FID. However, as we discussed above, the classifiers of StarGAN and DMIT are over-fitted, and their generators fool the classifiers by slightly changing textures or colors of the inputs (low LPIPS on DMIT also verifies such observations). To avoid the limits of FID, on one hand, we ask users to make decisions about which images from compared methods best match the target categories and also have high appearance quality. On the other hand, we use the AlexNet classifier introduced in Section 4.1 to automatically measure the semantic matching degree. Human preference results and the classification accuracy of generated images from Table 1 validate our discussions about StarGAN and DMIT. Differently, another multi-domain method StarGAN v2 achieves outstanding results on most measurements, owing to its elaborately designed multi-task discriminator. However, it has some limitations. For one thing, the diversity is poor in terms of the LPIPS measurement in Table 1. In Figs. 3 and 4, the results are almost the same when translating to the original domain of the inputs. For another, it needs one target class label to index the channel-wise style features (one channel for each class), and then inputs such indexed fix-length features to the generator. In other words, the framework requires the domain labels are mutually exclusive (i.e. one-hot vectors). Therefore, it would be difficult for StarGAN v2 to handle the multi-label translation settings, e.g. to the domain of young women with black hair.

Our method achieves significant advantages over StarGAN and DMIT, and comparable with two-domain-based MUNIT in terms of human preference and accuracy. Though inferior to StarGAN v2, our HIT is more general without the aforementioned drawbacks of StarGAN v2. Besides, HIT is efficient in handling both multi-domain and multi-modal, and even multi-granularity task. As shown in Figs. 5 and 6, given a source image, one can not only translate it to different

[1] The source codes are released at https://github.com/ssqiao/HIT.

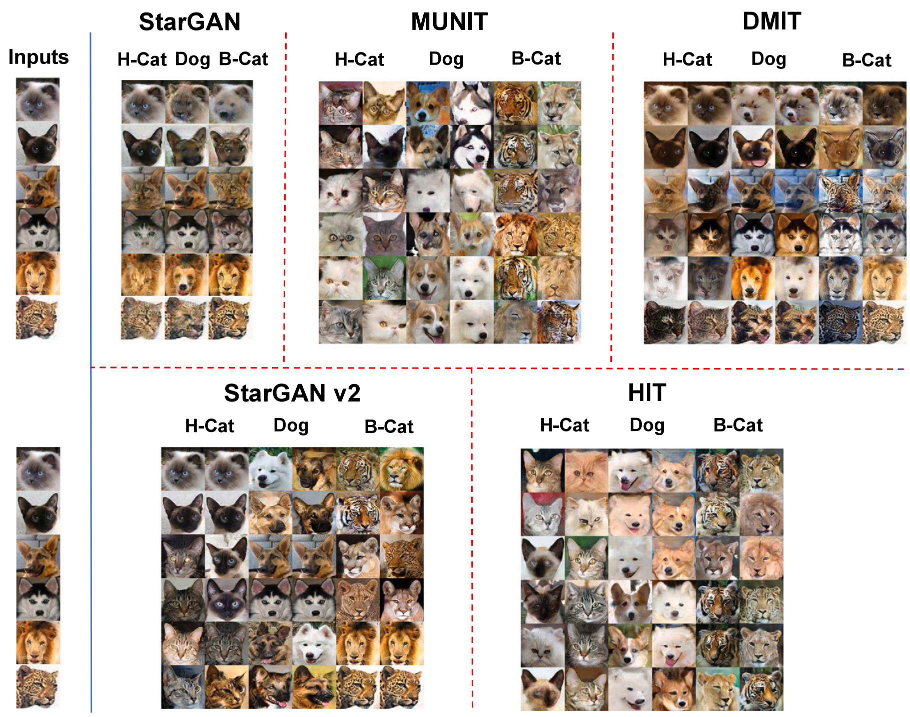


Fig. 3. Qualitative comparison on ImageNet. The inputs are translated to three super domains (H-Cat and B-Cat denote *House cat* and *Big cat* respectively, and the same meaning in the following). Two outputs (every 2 columns) for each input are sampled from predefined (MUNIT, DMIT, and StarGAN v2) or dynamically learned (our HIT) category distributions. Best viewed in colors and zoom-in. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

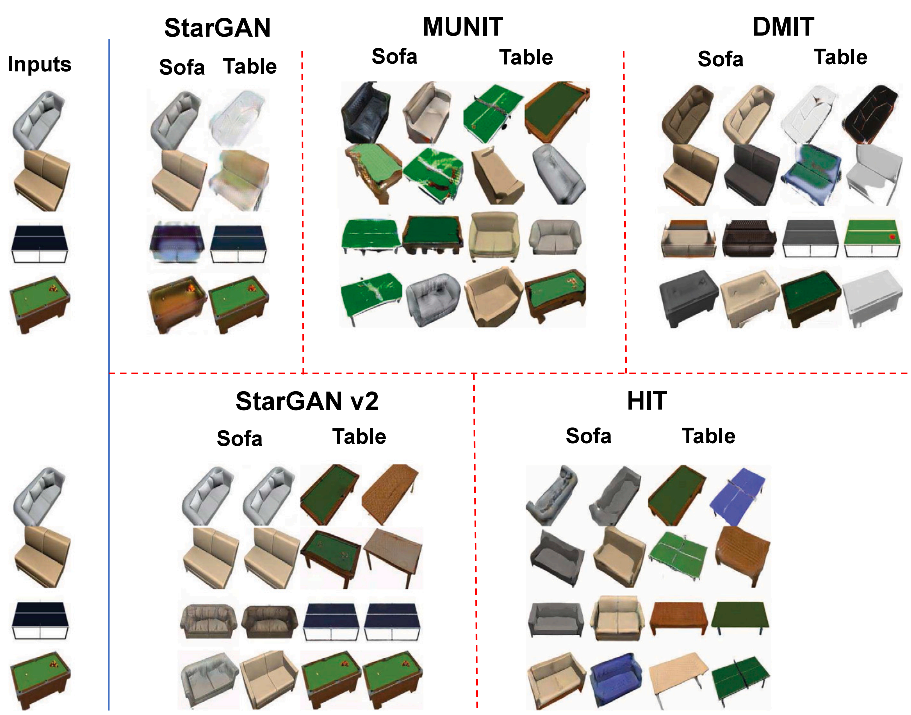


Fig. 4. Qualitative comparison on ShapeNet. The inputs are translated to *sofa* or *table* domains. Two outputs (every 2 columns) for each input are sampled from predefined (MUNIT, DMIT, and StarGAN v2) or dynamically learned (our HIT) category distributions. Best viewed in colors and zoom-in. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

categories (including the one it belongs to) with diverse outputs, but also control the semantic granularity of target categories, befitting from the nested distributions modeling. To further study whether the distributions are learned well (i.e. nested), using the UMAP [69] dimension reduction technique, we make a 2D visualization of learned Gaussians of some categories at different hierarchy levels. Specifically, 1000 points are randomly sampled from each Gaussian, and then projected to 2D space and fitted for an ellipse. From Fig. 7(a) (and results on ShapeNet in the supplementary materials), it can be seen that the child categories have a large overlap with their ancestors, e.g. the *Persian-House cats-Animal*, demonstrating the effectiveness of proposed

Table 1
Quantitative evaluation of images from different methods. Up-arrow/down-arrow means higher/lower result is better.

	ImageNet				ShapeNet				# Parameters (M)↓
	FID↓	LPIPS↑	Human↑	Accuracy↑	FID↓	LPIPS↑	Human↑	Accuracy↑	
StarGAN [16]	73.80	–	0.61%	0.8973	83.17	–	0.27%	0.7567	53
MUNIT [5]	77.73	0.491	20.38%	0.9643	167.20	0.392	23.97%	0.9901	47*N
DMIT [23]	49.64	0.199	0.69%	0.9197	76.76	0.256	0.89%	0.8457	34
StarGAN v2 [24]	29.53	0.275	68.31%	0.9927	56.88	0.153	56.15%	0.9670	60
HIT w/o $\mathcal{L}_{nest}$	128.20	0.420	–	0.5765	116.61	0.028	–	0.8814	21
HIT w/o $\mathcal{L}_{ent}$	83.92	0.419	–	0.9048	134.27	0.270	–	0.8942	21
HIT	62.40	0.458	10.01%	0.9979	107.39	0.320	18.72%	0.9638	21
Real	0	0.561	–	0.9920	0	0.583	–	0.9939	–

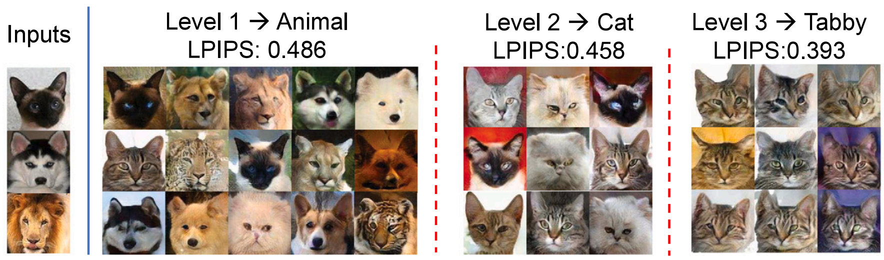
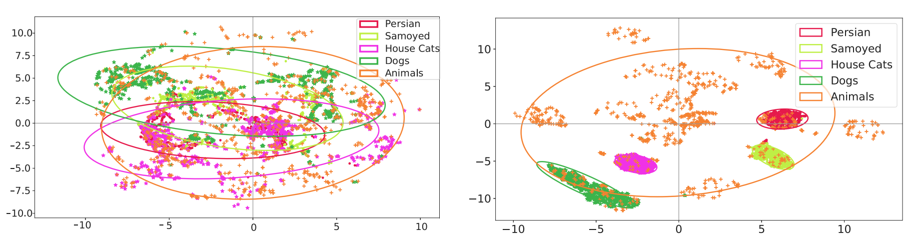


Fig. 5. Examples of multi-granularity translation on ImageNet. For a target domain (*animal, cat* and *tabby* in this case) at a particular hierarchy level, styles are sampled from its distribution. With the level becoming deeper, translations become more specific. The average LPIPS of translated images at corresponding levels is shown. Best viewed in colors and zoom-in. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Fig. 6. Examples of multi-granularity translation on ShapeNet. For a target domain (*furniture, table* and *billiard* in this case) at a particular hierarchy level, styles are sampled from its distribution. With the level becoming deeper, translations become more specific. The average LPIPS of translated images at corresponding levels is shown. Best viewed in colors and zoom-in. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



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

Fig. 7. 2D UMAP visualization of learned Gaussian distributions of domains in different hierarchy levels on ImageNet for (a) Full HIT and (b) HIT w/o $\mathcal{L}_{nest}$. For each domain, 1000 points are sampled and fitted for a Gaussian ellipse. Best viewed in colors and zoom-in. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

nested loss. Finally, from the metric of network parameters scale, we can find our HIT is more memory efficient than other state-of-the-arts.

Apart from comparisons with the most representative works, we further evaluate the performance of more recently proposed translation works on such categorical image translation tasks. Specifically, the recent competitive works TUNIT [45], StyleDis [47] and i-StyleGAN [49] are trained and test. The first two works are reference-guided and proposed to address the problem of truly unsupervised training in image translation problem, and the TUNIT supports both supervised and unsupervised training in their public released codes. Therefore, we trained two kinds of models for TUNIT, i.e., TUNIT-sup and TUNIT-unsup, and the unsupervised model for StyleDis on our collected ImageNet and ShapeNet datasets. All the hyper-parameter settings referred to the recommendation of their released codes or papers. To quantitatively compute the FID, LPIPS and semantic accuracy, for each test image as the source (i.e., content), we randomly select 38 images from every category as the target (i.e., style) for the reference-guided translations. The results are show in Table 2, Figs. 8 and 9. It can be seen that these more recent SOTAs achieve outstanding generated image quality in terms of the FID score and visual perception in the figures, which are

Table 2
Quantitative evaluation of more recent translation methods. Up-arrow/down-arrow means a higher/lower result is better.

	ImageNet			ShapeNet			# Parameters (M)↓
	FID↓	LPIPS↑	Accuracy↑	FID↓	LPIPS↑	Accuracy↑	
TUNIT-unsup [45]	31.67	0.392	0.8465	77.39	0.254	0.6907	129
TUNIT-sup [45]	31.78	0.366	0.9621	39.81	0.094	0.8053	129
StyleDis [47]	26.76	0.364	0.9350	63.54	0.410	0.8946	117
i-StyleGAN [49]	33.39	0.357	0.9861	63.38	0.144	0.9736	60
HIT	62.40	0.458	0.9979	107.39	0.320	0.9638	21
Real	0	0.561	0.9920	0	0.583	0.9939	–

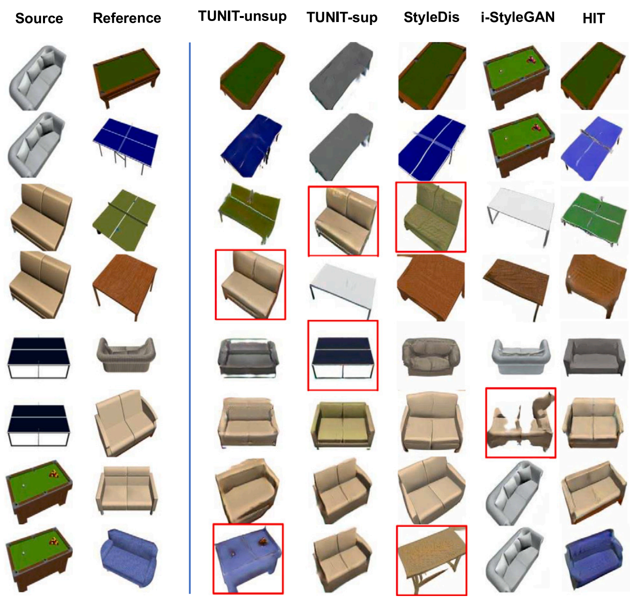


Fig. 8. Qualitative comparison with more recent translation methods on ShapeNet. Red rectangles indicate failed results. TUNIT and StyleDis are reference-guided, and the others are sampling-based. Best viewed in colors and zoom-in. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

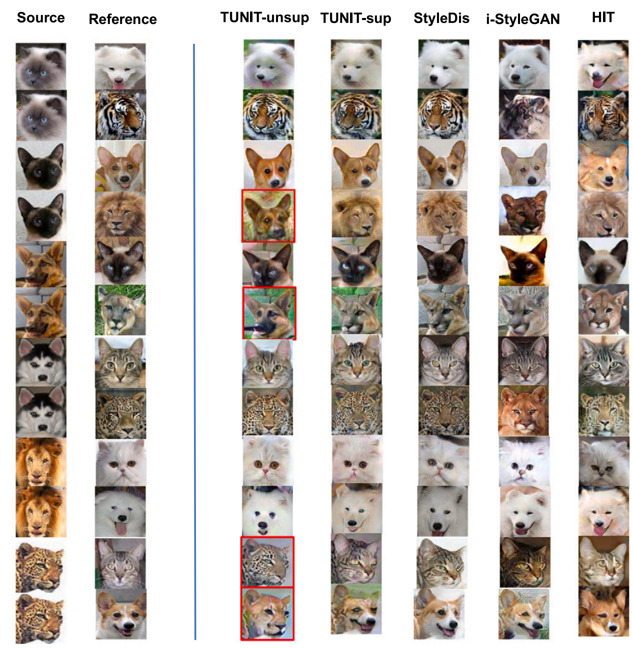


Fig. 9. Qualitative comparison with more recent translation methods on ImageNet. Red rectangles indicate failed results. TUNIT and StyleDis are reference-guided, and the others are sampling-based. Best viewed in colors and zoom-in. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

comparable with and enen surpass the most competitive work StarGAN v2 in Table 1. This may be owing to their elaborately designed network modules, the parameters of which are three or even six times larger than ours. However, the semantic accuracy of translated images is not good enough, especially for the two unsupervised models (i.e., TUNIT-unsup and StyleDis), which is mainly due to the lack of domain labels for supervision. On the more challenging ShapeNet dataset, we find that TUNIT and i-StyleGAN cannot well handle the translation task, i.e., they either fail to change the categories (i.e., low accuracy of TUNIT) or suffer from mode collapse (i.e., low LPIPS score of both TUNIT and i-StyleGAN). Another general drawback of these works is the relatively lower diversity (LPIPS score) of generations, which might be due to the relatively smaller style space modeled by these methods. In contrast, our HIT usually has satisfactory generation diversity as it obtains styles from the explicitly divided hierarchical distribution space.

4.3. Model analysis

In this section, we study the impacts of the proposed nested distributions loss in Eq. (5). Table 1 shows quantitative comparisons. Figs. 10 and 11 give qualitative results of baselines without $\mathcal{L}_{nest}$ or $\mathcal{L}_{ent}$ on different datasets. We can see that by completely dropping the nested loss with only classifier and discriminator left for domain classification and distribution modeling, the quality of image appearance is poor (high FID), and even leads to mode collapse on ShapeNet (low LPIPS). Fig. 7(b) shows that without the nested loss, distributions of parents and children are separated and the distribution space is holistically sparse, which is not semantically reasonable and would easily lead to unavailable sampling in such sparse space. It verifies that directly learning distributions of multiple domains in a common space is a quite challenging task. Further adding the divergence loss (i.e. w/o entropy loss), the quality and diversity are improved, but the semantics of some cases are still not satisfactory. Finally, with the entropy loss added, the quality, diversity, and semantics of generated images are all improved.

It is noted that the nested distributions loss contains two sub-terms, we further conduct the studies of the weight settings of the two sub-terms in Eq. (5), i.e., the divergence loss L_{dvg} and entropy loss L_{ent}. Specifically, such experiments are conducted on ImageNet by fixing the original weight settings of the other loss terms, and varying the settings of either L_{dvg} or L_{ent}. The evaluation results are shown in the Table 3. As for the weight settings of the divergence loss, it can be seen that with too large or too small weight settings, the performance in terms of image quality (FID) and semantic accuracy will be degraded compared with the default setting. As for the weight settings of the entropy loss, the impact on the performance of different loss weight settings is more heavily. For instance, under the quite large weight setting of 100.0, the model fails to generate satisfactory images in terms of the quite large FID, unreasonable LPIPS and randomly predicted category domains, all of which are noisy pixels on the generations. Therefore, the default settings of 1.0 for both terms are optimal.

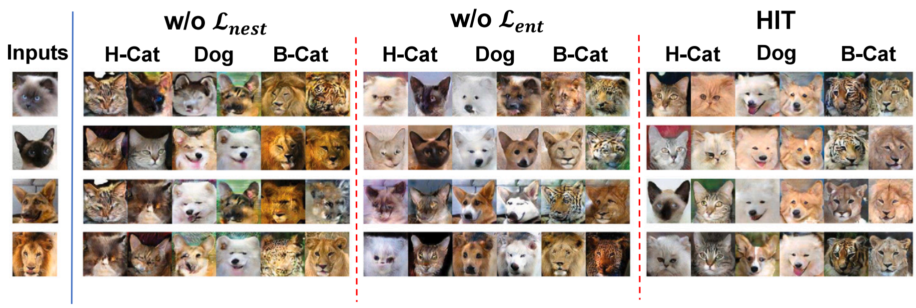


Fig. 10. Qualitative comparison with baselines of our method on ImageNet, including w/o the whole nested distributions loss and w/o the entropy loss. Two translated images to each domain of an input are shown in adjacent columns. Best viewed in colors and zoom-in. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

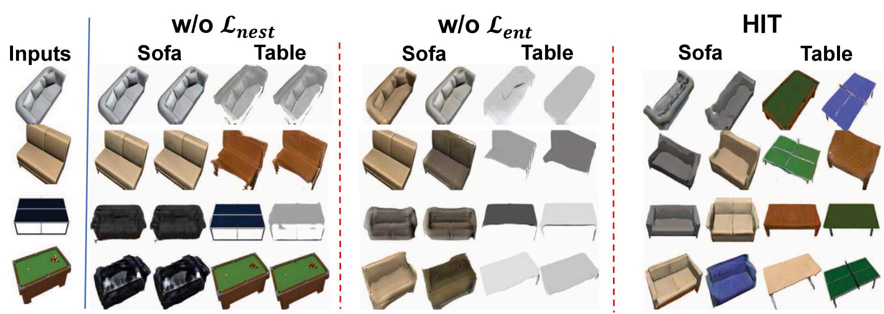
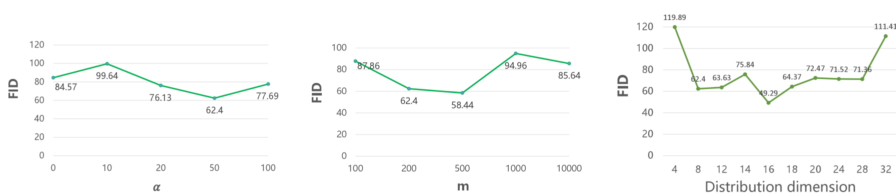


Fig. 11. Qualitative comparison with baselines of our method on ShapeNet, including w/o the whole nested distributions loss and w/o the entropy loss. Two translated images to each domain of an input are shown in adjacent columns. Best viewed in colors and zoom-in. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



(a) Threshold α (b) Margin m (c) Distribution dimension

Fig. 12. FID of translated images on ImageNet with different hyper-parameters settings. (a). Fix $m = 200$, distribution dimension as 8, and change the threshold α (b). Fix $\alpha = 50$, distribution dimension as 8, and change the margin m. (c) Fix $m = 200, \alpha = 50$, and change the distribution dimension.

Table 3
Quantitative evaluation of generated images by setting different weights of the divergence loss (left half panel) and entropy loss (right half panel). Up-arrow/down-arrow means a higher/lower result is better.

	Weights of divergence loss					Weights of entropy loss				
	0	0.1	1.0	10.0	100.0	0	0.1	1.0	10.0	100.0
FID↓	161.89	96.04	62.40	80.00	86.19	83.92	110.12	62.40	115.78	321.84
LPIPS↑	0.405	0.469	0.458	0.475	0.459	0.419	0.439	0.458	0.425	0.558
Accuracy↑	0.7675	0.9443	0.9979	0.9907	0.9013	0.9048	0.8139	0.9979	0.8725	0.3339

The proposed divergence loss in Eq. (3) contains two hyper-parameters, i.e. nested threshold α and margin m. Besides, the dimension of the distribution space is also one significant hyper-parameter that has impacts on the sparseness and capacity of the learned space. We conduct parameters analysis on ImageNet by fixing one parameter and varying others. Fig. 12 shows the impacts on image quality (FID) with different settings. First, with too large settings of m, distributions which do not have nested relationship would be pushed too far away, leading to sparse space. Sampling in such space would make the learning of the generator quite difficult. In contrast, with too small settings of m, the discriminabilities of distributions may be poor. Second, as for nested threshold α, a large setting of α would relax the nested constraint too much, resulting in a small overlap between parent and children. When α is set as 0, it means parent and children are completely overlapped, which would lead to concentration of all its children. Third, similar to the impact of m, with too high dimension distribution space but limited training data, the learned distributions would be quite sparse. Similarly, with lower dimension settings, the discriminabilities of sampled styles may be poor. Therefore, a trade-off value of 200 for m, 50 for α, and 8 for distribution dimension is generally set for all datasets. Please note that on a specific dataset, one may obtain better results by elaborately tuning each hyper-parameter, e.g. FID of 49.29 by setting distribution dimension as 16 is better in Fig. 12(c).

Nested distributions modeling is the core characteristic of the proposed method, which relies on the construction of a hierarchy. In

Table 4
Quantitative evaluation of generated images by setting different numbers of leaf-level child classes (left half panel) or hierarchical levels (right half panel). Up-arrow/down-arrow means a higher/lower result is better.

	# Child classes					# Hierarchical levels			
	2	3	4	5	6	L2	L1+L2	L2+L3	L1+L2+L3
FID↓	86.97	84.50	62.40	66.33	76.51	67.70	66.89	73.40	62.40
LPIPS↑	0.233	0.462	0.458	0.461	0.476	0.009	0.132	0.450	0.458
Accuracy↑	0.8093	0.9107	0.9979	0.9496	0.9736	0.6187	0.9285	0.9853	0.9979

this part, we conduct study in terms of the number of child classes and the number of hierarchical levels. Specifically, in our collected 3 super categories of animal heads data of the ImageNet, there exists at most 6 child classes for each super one. To investigate the impact of the number of child classes, we respectively set the number of child nodes at the leaf level as $2, 3, 4, 5, 6$, and evaluate the translation performance at the super category level as done in Section 4.2. From the results in Table 4, it is observed that too fewer child classes (e.g., 2 or 3) are not sufficient to fit the nested distribution space, leading to poor FID score and accuracy of the translated images. Given only 2 child classes for each super domain, the LPIPS is quite small, which reflects the poor diversity of generated modes. With the number of child classes increasing, the overall translation performance become better and stable. However, if more child classes are added in the hierarchy, the FID score tends to gradually decrease, which may be due to the increasing difficulty of optimizing the nested distribution loss. To be more specific, inserting more sub-distributions into one parent needs more metric learning efforts for the model. Besides, the computation of the average prediction on the child classes in the entropy loss (i.e., Eq. (4)) would also be less accurate in statistic when the number of child classes becomes larger than the sampling number (i.e., set as 5 due to GPU memory limit) for each input image. Therefore, 4 or 5 child classes seems an optimal choice for the hierarchy construction in our model.

Furthermore, we study the impact of number of levels for training our HIT. We respectively train our HIT using only one level (i.e., the second level L2 that the super domains belong to), the first two levels (i.e., L1 and L2), the last two levels (i.e., L2 and L3), and all the three levels. By evaluating the categorical translation performance at L2 level in Table 4, we can find that using all the three levels perform best in terms of all metrics. Besides, using auxiliary levels (i.e., the non-target levels, L1 and L3), either using only one or all, is beneficial to the translation performance at the target categorical level, which demonstrates the effectiveness of the nested distribution modeling and designed relevant constraints. Instead, directly training in one flat level leads to mode collapse (quite poor LPIPS score and semantic accuracy). Another interesting observation is that using the bottom two levels overall performs better than using the upper two levels, which means that dividing the super domains into finer-grained modes may better benefit the distribution fit at the target level.

5. Discussions

About the continue learning in categorical translation. With the notion of hierarchy, one of the advantages for the proposed model is that newer sub-domains could be introduced on top of learning the known ones. For example, the *poodle-dog* which was not initially known by the model. We think adding new domains on known ones in our method can be discussed two situations. The **first** is that the data of newer sub-domains are included in the training stage, i.e., they participated in the training as the roles of their parent categories (e.g., the *poodle-dog* as the *dog*) . In such case, since all their ancestor nodes have been included in the model and distributions of which were learned, we can directly finetune the original model to further only learn the distributions of the newer sub-domains by adding such distribution nodes at the leaf level in the nested distribution space, supervised by the same devised losses. The distributions of the known ones can be fixed. The **second** is that the data of newer sub-domains are unseen in the training stage. In such cases, all relevant distribution nodes (i.e., all the ancestors in the hierarchy of the newer sub-domains) should be finetuned, and the other nodes are unchanged. As for the competitive StarGAN v2, since it needs one channel for each domain as the output style vector, added newer sub-domains would change the network architecture. Therefore, it may need to train the whole framework from scratch. Since our method is built on distribution sampling, the dimension of the style vector is fixed and thus the architectures of generator and discriminator are not affected. Directly finetuning the distribution space would be feasible. Besides, the generator in our method have learnt the semantic of the newer sub-domains at their parent level, which would lead to good generalization ability on the unseen sub-domain data in the second situation.

As for the current inferior FID score and poor human evaluation score. We make analysis of the possible reasons for inferior performance of our method, especially compared to the StarGAN v2. **First**, the assumption of a single Gaussian for each category domain is not the optimal scheme to realize the nested distribution modeling. As shown and discussed in the section of failure case analysis in the supplementary materials, such issue would lead to sparse sampling around the centers of parent distributions and poor generated results sometimes. **Second**, the model capacity of our method is limited (21M parameters in Table 1) compared to the other SOTAs, especially for the deeper and larger StarGAN v2 (60M). To be honest, categorical translation is a quite challenge task. Both the textures and shapes need to be changed and rendered, which has high request on the network capacity. In our framework, to compute the entropy loss in Eq. (4), we need to sample at least 5 styles from the target domain for each input image, which leads to high burden on the GPU memory usage. Therefore, in our current implementation, we design a relative smaller network, which may be one of the reasons for the inferior performance. Third, it is noted that most of recent multi-domain translation methods including StarGAN v2 leverage the multi-task discriminator instead of the traditional multi-category classifier (the hierarchical classifier in our method belongs to the multi-category classifier), i.e., one adversarial discriminator branch for each domain, which have been verified to be more effective on the multi-domain categorical translation task. Currently, we think the three reasons mentioned above are the main obstacles for our method to achieve better FID and human evaluation performance. Investigation of combining the proposed entropy loss with the multi-task discriminator or attempt of the other derivable nested distribution modeling may help to improve our method.

6. Conclusions

In this paper, we propose a Hierarchical Image-to-image Translation (HIT) method which incorporates multi-domain and multi-modal translation into one model. Experiments on challenging object datasets show that the proposed method can well achieve such two goals and additional granularity controlled translations owing to the nested distributions modeling. However, current work has a limitation, i.e. the assumption of a single Gaussian for each category domain. The parent distributions should be the 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 generated results sometimes. Despite such limits, we believe modeling multiple domains in a common space is a promising way to realize efficient multi-domain and multi-modal translation tasks, and a better assumption to realize the nested relationships among distributions is one of our future research directions.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

Most of this work was done at the Institute of Computing Technology, Chinese Academy of Sciences, where Shishi Qiao pursued his PhD degree. This work is partially supported by National Key R&D Program of China No. 2021ZD0111901, Natural Science Foundation of China under contracts Nos. U21B2025, U19B2036, 62206260.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.patcog.2023.110058.

References

[1] Y. Pang, J. Lin, T. Qin, Z. Chen, Image-to-image translation: Methods and applications, 2021, CoRR abs/2101.08629.
[2] P. Isola, J. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: IEEE,CVPR, 2017, pp. 5967–5976.
[3] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A.P. Aitken, A. Tejani, J. Totz, Z. Wang, W. Shi, Photo-realistic single image super-resolution using a generative adversarial network, in: IEEE,CVPR, 2017, pp. 105–114.
[4] J. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: IEEE , ICCV, 2017, pp. 2242–2251.
[5] X. Huang, M. Liu, S.J. Belongie, J. Kautz, Multimodal unsupervised image-to-image translation, in: ECCV, 2018, pp. 179–196.
[6] Z. Yang, H. Liu, D. Cai, On the diversity of conditional image synthesis with semantic layouts, IEEE Trans. Image Process. 28 (6) (2019) 2898–2907.
[7] L. Yan, W. Zheng, C. Gou, F.-Y. Wang, IsGAN: Identity-sensitive generative adversarial network for face photo-sketch synthesis, Pattern Recognit. 119 (2021) 108077.
[8] X. Yang, J. Zhao, Z. Wei, N. Wang, X. Gao, SAR-to-optical image translation based on improved CGAN, Pattern Recognit. 121 (2022) 108208.
[9] T. Wang, L. Wu, C. Sun, A coarse-to-fine approach for dynamic-to-static image translation, Pattern Recognit. 123 (2022) 108373.
[10] H. Lee, H. Tseng, J. Huang, M. Singh, M. Yang, Diverse image-to-image translation via disentangled representations, in: ECCV, 2018, pp. 36–52.
[11] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A.C. Courville, Y. Bengio, Generative adversarial nets, in: NIPS, 2014, pp. 2672–2680.
[12] M. Mirza, S. Osindero, Conditional generative adversarial nets, 2014, CoRR abs/1411.1784.
[13] M. Liu, T. Breuel, J. Kautz, Unsupervised image-to-image translation networks, in: NIPS, 2017, pp. 700–708.
[14] T. Kim, M. Cha, H. Kim, J.K. Lee, J. Kim, Learning to discover cross-domain relations with generative adversarial networks, in: ICML, 2017, pp. 1857–1865.
[15] Y. Taigman, A. Polyak, L. Wolf, Unsupervised cross-domain image generation, in: ICLR, 2017.
[16] Y. Choi, M. Choi, M. Kim, J. Ha, S. Kim, J. Choo, StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation, in: IEEE,CVPR, 2018, pp. 8789–8797.
[17] Z. He, W. Zuo, M. Kan, S. Shan, X. Chen, AttGAN: Facial attribute editing by only changing what you want, IEEE Trans. Image Process. 28 (11) (2019) 5464–5478.
[18] J. Zhu, R. Zhang, D. Pathak, T. Darrell, A.A. Efros, O. Wang, E. Shechtman, Toward multimodal image-to-image translation, in: NIPS, 2017, pp. 465–476.
[19] G. Griffin, P. Perona, Learning and using taxonomies for fast visual categorization, in: IEEE,CVPR, 2008.
[20] J. Deng, J. Krause, A.C. Berg, F. Li, Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition, in: IEEE,CVPR, 2012, pp. 3450–3457.
[21] B. Zhao, F. Li, E.P. Xing, Large-scale category structure aware image categorization, in: NIPS, 2011, pp. 1251–1259.
[22] V. Ordonez, J. Deng, Y. Choi, A.C. Berg, T.L. Berg, From large scale image categorization to entry-level categories, in: IEEE, ICCV, 2013, pp. 2768–2775.
[23] X. Yu, Y. Chen, S. Liu, T.H. Li, G. Li, Multi-mapping image-to-image translation via learning disentanglement, in: NIPS, 2019, pp. 2990–2999.
[24] Y. Choi, Y. Uh, J. Yoo, J. Ha, Stargan v2: Diverse image synthesis for multiple domains, in: IEEE,CVPR, 2020, pp. 8185–8194.
[25] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, in: ICLR, 2016.
[26] E.L. Denton, S. Chintala, A. Szlam, R. Fergus, Deep generative image models using a Laplacian pyramid of adversarial networks, in: NIPS, 2015, pp. 1486–1494.
[27] H. Zhang, T. Xu, H. Li, StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks, in: IEEE, ICCV, 2017, pp. 5908–5916.
[28] T. Karras, T. Aila, S. Laine, J. Lehtinen, Progressive growing of GANs for improved quality, stability, and variation, 2017, CoRR abs/1710.10196.
[29] A. Brock, J. Donahue, K. Simonyan, Large scale GAN training for high fidelity natural image synthesis, in: ICLR, 2019.
[30] 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.
[31] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein GAN, 2017, CoRR abs/1701.07875.
[32] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A.C. Courville, Improved training of wasserstein GANs, in: NIPS, 2017, pp. 5767–5777.
[33] T. Miyato, T. Kataoka, M. Koyama, Y. Yoshida, Spectral normalization for generative adversarial networks, in: ICLR, 2018.
[34] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, 2018, CoRR abs/1812.04948.
[35] A. Odena, C. Olah, J. Shlens, Conditional image synthesis with auxiliary classifier GANs, in: ICML, 2017, pp. 2642–2651.
[36] G. Kwon, J.C. Ye, Diagonal attention and style-based GAN for content-style disentanglement in image generation and translation, in: IEEE, ICCV, 2021, pp. 13960–13969.
[37] J. Park, D.K. Han, H. Ko, Fusion of heterogeneous adversarial networks for single image dehazing, IEEE Trans. Image Process. 29 (2020) 4721–4732.
[38] S.E. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, H. Lee, Generative adversarial text to image synthesis, in: ICML, 2016, pp. 1060–1069.
[39] J. Wu, C. Zhang, T. Xue, B. Freeman, J. Tenenbaum, Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling, in: NIPS, 2016, pp. 82–90.
[40] T. Park, J. Zhu, O. Wang, J. Lu, E. Shechtman, A.A. Efros, R. Zhang, Swapping autoencoder for deep image manipulation, in: NIPS, 2020.
[41] H. Kim, Y. Choi, J. Kim, S. Yoo, Y. Uh, Exploiting spatial dimensions of latent in GAN for real-time image editing, in: IEEE, CVPR, 2021, pp. 852–861.
[42] Z. Yi, H.R. Zhang, P. Tan, M. Gong, DualGAN: Unsupervised dual learning for image-to-image translation, in: IEEE , ICCV, 2017, pp. 2868–2876.
[43] Y. Li, S. Tang, R. Zhang, Y. Zhang, J. Li, S. Yan, Asymmetric GAN for unpaired image-to-image translation, IEEE Trans. Image Process. 28 (12) (2019) 5881–5896.
[44] X. Li, Z. Du, Y. Huang, Z. Tan, A deep translation (GAN) based change detection network for optical and SAR remote sensing images, ISPRS J. Photogramm. Remote Sens. 179 (2021) 14–34.
[45] K. Baek, Y. Choi, Y. Uh, J. Yoo, H. Shim, Rethinking the truly unsupervised image-to-image translation, in: IEEE, ICCV, 2021, pp. 14134–14143.
[46] H. Lee, J. Seol, S. Lee, Contrastive learning for unsupervised image-to-image translation, 2021, CoRR abs/2105.03117.
[47] K. Kim, S. Park, E. Jeon, T. Kim, D. Kim, A style-aware discriminator for controllable image translation, in: IEEE,CVPR, 2022, pp. 18218–18227.
[48] 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.
[49] S. Xie, L. Kong, M. Gong, K. Zhang, Multi-domain image generation and translation with identifiability guarantees, in: ICLR, 2023.
[50] Y. Liu, M.D. Nadai, J. Yao, N. Sebe, B. Lepri, X. Alameda-Pineda, GMM-UNIT: Unsupervised multi-domain and multi-modal image-to-image translation via attribute Gaussian mixture modeling, 2020, CoRR abs/2003.06788.
[51] M. Marszalek, C. Schmid, Constructing category hierarchies for visual recognition, in: ECCV, 2008, pp. 479–491.
[52] J. Deng, N. Ding, Y. Jia, A. Frome, K. Murphy, S. Bengio, Y. Li, H. Neven, H. Adam, Large-scale object classification using label relation graphs, in: ECCV, 2014, pp. 48–64.
[53] N. Ding, J. Deng, K.P. Murphy, H. Neven, Probabilistic label relation graphs with ising models, in: IEEE, ICCV, 2015, pp. 1161–1169.
[54] N. Srivastava, R. Salakhutdinov, Discriminative transfer learning with tree-based priors, in: NIPS, 2013, pp. 2094–2102.
[55] S.J. Hwang, L. Sigal, A unified semantic embedding: Relating taxonomies and attributes, in: NIPS, 2014, pp. 271–279.

[56] Z. Yan, H. Zhang, R. Piramuthu, V. Jagadeesh, D. DeCoste, W. Di, Y. Yu, HD-CNN: Hierarchical deep convolutional neural networks for large scale visual recognition, in: IEEE, ICCV, 2015, pp. 2740–2748.
[57] W. Goo, J. Kim, G. Kim, S.J. Hwang, Taxonomy-regularized semantic deep convolutional neural networks, in: ECCV, 2016, pp. 86–101.
[58] K. Ahmed, M.H. Baig, L. Torresani, Network of experts for large-scale image categorization, in: ECCV, 2016, pp. 516–532.
[59] J. Xie, Y. Xu, E. Nijkamp, Y.N. Wu, S. Zhu, Generative hierarchical learning of sparse FRAME models, in: IEEE, CVPR, 2017, pp. 1933–1941.
[60] S. Zhao, J. Song, S. Ermon, Learning hierarchical features from deep generative models, in: ICML, 2017, pp. 4091–4099.
[61] K.K. Singh, U. Ojha, Y.J. Lee, Finegan: Unsupervised hierarchical disentanglement for fine-grained object generation and discovery, in: IEEE, CVPR, 2019, pp. 6490–6499.
[62] B. Athiwaratkun, A.G. Wilson, Hierarchical density order embeddings, in: ICLR, 2018.
[63] I. Vendrov, R. Kiros, S. Fidler, R. Urtasun, Order-embeddings of images and language, in: ICLR, 2016.
[64] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M.S. Bernstein, A.C. Berg, F. Li, ImageNet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (3) (2015) 211–252.
[65] A.X. Chang, T.A. Funkhouser, L.J. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, F. Yu, ShapeNet: An information-rich 3D model repository, 2015, CoRR abs/1512.03012.
[66] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, GANs trained by a two time-scale update rule converge to a local Nash equilibrium, in: NIPS, 2017, pp. 6626–6637.
[67] R. Zhang, P. Isola, A.A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: IEEE,CVPR, 2018, pp. 586–595.
[68] Z. Liu, P. Luo, X. Wang, X. Tang, Deep learning face attributes in the wild, in: IEEE ,ICCV, 2015, pp. 3730–3738.
[69] L. McInnes, J. Healy, N. Saul, L. Großberger, UMAP: Uniform manifold approximation and projection, J. Open Source Software 3 (29) (2018) 861.

Shishi Qiao received the B.S. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 2014, and the Ph.D. degree in computer science from the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing, China, in 2021. In 2021, he joined the Faculty of the Department of Information Science and Engineering, Ocean University of China (OUC), Qingdao, China, where he has been an Assistant Professor. His research interests mainly include computer vision, pattern recognition, machine learning and, in particular, video face recognition, multimedia retrieval, object and scene understanding with deep generative models.

Ruiping Wang (S'08-M'11) received the B.S. degree in applied mathematics from Beijing Jiaotong University, Beijing, China, in 2003, and the Ph.D. degree in computer science from the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing, in 2010. He was a Post Doctoral Researcher with the Department of Automation, Tsinghua University, Beijing, from 2010 to 2012. He also spent one year as a Research Associate with the Computer Vision Laboratory, Institute for Advanced Computer Studies, University of Maryland at College Park, College Park, from 2010 to 2011. In 2012, he joined the Faculty of the Institute of Computing Technology, Chinese Academy of Sciences, where he has been a Professor since 2017. His research interests include computer vision, pattern recognition, and machine learning.

Shiguang Shan (M'04-SM'15-F'21) received the M.S. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 1999, and the Ph.D. degree in computer science from the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing, China, in 2004. In 2002, he joined ICT, CAS, where he has been a Professor since 2010. He is currently the Deputy Director of the Key Laboratory of Intelligent Information Processing, CAS. He has authored over 200 papers in refereed journals and proceedings in computer vision and pattern recognition. His research interests include computer vision, pattern recognition, and machine learning. He especially focuses on face recognition related research topics. He was a recipient of the Chinas State Natural Science Award in 2015 and the Chinas State S&T Progress Award in 2005 for his research work. He is an Associate Editor of several journals, including the IEEE TRANSACTIONS ON IMAGE PROCESSING, the Computer Vision and Image Understanding, the Neurocomputing, and the Pattern Recognition Letters. He has served as the Area Chair for some international conferences, including ICCV11, ICPR12/14/20, ACCV12/16/18, FG13/18/20, ICASSP14, BTAS18, and CVPR19/20.

Xilin Chen (M'00-SM'09-F'16) is a professor with the Institute of Computing Technology, Chinese Academy of Sciences (CAS). He has authored one book and more than 300 papers in refereed journals and proceedings in the areas of computer vision, pattern recognition, image processing, and multimodal interfaces. He is currently an associate editor of the IEEE Transactions on Multimedia, and a Senior Editor of the Journal of Visual Communication and Image Representation, a leading editor of the Journal of Computer Science and Technology, and an associate editor-in-chief of the Chinese Journal of Computers, and Chinese Journal of Pattern Recognition and Artificial Intelligence. He served as an Organizing Committee member for many conferences, including general co-chair of FG13 / FG18, program co-chair of ICMI 2010. He is / was an area chair of CVPR 2017 / 2019 / 2020, and ICCV 2019. He is a fellow of the IEEE, IAPR, and CCF.