TransGaGa: Geometry-Aware Unsupervised Image-to-Image Translation

Wayne Wu[1] Kaidi Cao[2] Cheng Li[1] Chen Qian[1] Chen Change Loy[3]

[1]SenseTime Research [2]Stanford University
[3]Nanyang Technological University

{wuwenyan, chengli, qianchen}@sensetime.com kaidicao@cs.stanford.edu
ccloy@ntu.edu.sg

Figure 1: We propose a geometry-aware framework for *unsupervised* image-to-image translation, which is robust to arbitrary shape variations between domains. We show the results of both near-rigid and non-rigid objects. *(left)* Cow and cheetah rendered from CAD models. *(center)* Cat and human face from in-the-wild datasets. *(right)* Horse and giraffe from Flickr.

Abstract

Unsupervised image-to-image translation aims at learning a mapping between two visual domains. However, learning a translation across large geometry variations always ends up with failure. In this work, we present a novel disentangle-and-translate framework to tackle the complex objects image-to-image translation task. Instead of learning the mapping on the image space directly, we disentangle image space into a Cartesian product of the appearance and the geometry latent spaces. Specifically, we first introduce a geometry prior loss and a conditional VAE loss to encourage the network to learn independent but complementary representations. The translation is then built on appearance and geometry space separately. Extensive experiments demonstrate the superior performance of our method to other state-of-the-art approaches, especially in the challenging near-rigid and non-rigid objects translation tasks. In addition, by taking different exemplars as the appearance references, our method also supports multimodal translation. Project page: https://wywu.github.io/projects/TGaGa/TGaGa.html

1. Introduction

> I will be your mirror. Reflect what you are, in case you don't know. I will be the wind, the rain and the sunset. The light on your door to show that you are home.
>
> *Lou Reed*

Unsupervised image-to-image translation aims at learning tahe translation between two different image domains without any pairwise supervision. The notion of image translation has been widely applied in colorization [47], super-resolution [22, 43] and style transfer [9].

Early works demonstrated the effectiveness of deep neural network in transferring local textures, demonstrating successful cases on seasonal scene shifting [52, 27] and painting style transfer [23]. However, researchers soon realized its limitation on the more complicated cases, *i.e.*, translation between two domains with large geometry variations [52, 10]. To handle more complex cases, one has to establish the translation on the higher semantic level. For example, based on the understanding of the components of neck, body and leg of a horse, we may imagine a giraffe with the same posture. However, one can hardly implement this translation by replacing the local texture due to the large geometry variations between the two domains.

Performing a translation on the higher semantic level is non-trivial. Geometry information plays a critical role here but, often, there is a significant geometry gap between two image domains, *e.g.*, cat to human-face and horse to giraffe. Although containing the same corresponding components with the similar semantic meaning between the two domains, their spatial distributions are rather different.

In this paper, we propose a novel geometry-aware framework for unsupervised image-to-image translation. Instead of directly translating on the image space, we first map the image into the Cartesian product of geometry and appearance spaces and then perform the translation in each latent space. To encourage the disentanglement of two spaces, we propose an unsupervised conditional variational AutoEncoder framework, in which a Kullback-Leibler (KL) divergence loss and skip-connection design are introduced to encourage the network to learn a complementary representation of geometry and appearance. Then we build the translation between two domains based on their bottleneck representation. Extensive experiments show the effectiveness of our framework in establishing translation between objects both on synthesis and real-world datasets. Our method achieves superior performance to state-of-the-art methods in both qualitative and quantitative experiments.

We summarize the contributions of this work as follows: 1) We propose a novel framework for unsupervised image-to-image translation. Instead of directly translating on the image space, we build the mapping between two domains on their disentangled latent appearance-geometry space. Our framework extends the ability of CycleGAN on more complicated objects like animals. 2) Fine-disentangled latent space naturally endows our model with the ability of diverse and exemplar-guided generation, which is a challenging and ill-posed multimodal problem in unsupervised image-to-image translation.

2. Related Work

Image-to-Image Translation. The goal of image-to-image translation is to learn a mapping from a source image domain to a target image domain. Pix2Pix [15] proposes a unified framework for image-to-image translation first time based on conditional GANs. Several works [41, 40] extend it to deal with high-resolution or video synthesis. Although appealing results have been shown, these methods need paired data for training. For unsupervised image-to-image translation with unpaired training data, CycleGAN [52], DiscoGAN [20], DualGAN [46] and UNIT [27] are proposed based on the idea of cycle-consistency. GANimorph [10] introduce a discriminator with dilated convolutions to get a more context-aware generator. However, without paired training data, the translation problem is inherently ill-posed because of the infinite existing mappings between two domains. Recent studies have attempted to solve this problem for multimodal generations. CIIT [24], MUNIT [14], DRIT [23] and EG-UNIT [29] decompose the latent space of images into a domain-invariant content space and a domain-specific style space to get diverse outputs. However, once the cross-domain structure variation becoming large, the assumption of domain-invariant content space is violated. Even though it is intuitive to share the latent space of content across domains in style transfer tasks, it is hard to embed the complex geometry cues of different domains with one shared distribution. The performance of all existing methods degrades dramatically in the translation with large cross-domain geometry variations.

Structural Representation Learning. To model visual content, several unsupervised techniques have been proposed including VAE [21], GANs [11] and ARNs [32, 39]. Recently, many literature focus on unsupervised landmark discovery [38, 37, 49, 16, 6] for structural representation learning. Since landmark is an explicit representation for the structure of objects, it can better capture the intrinsic shape of object than other representations. Inspired by the recent development of unsupervised landmark discovery, a heatmap-stack of landmarks are learned in this work for explicit structure representation.

Disentanglement of Representation. Disentanglement is important for the control over structure and appearance. There exist a number of studies on face and person image generation [1, 8, 30, 42]. Although enjoying the advantage of well pose-guided synthesis, these methods require predefined annotations for supervised learning. Several works for unsupervised disentanglement have been proposed, *e.g.*, InfoGAN [5] and β-VAE [13]. However, these methods suffer from the lack of interpretability, and the meaning of each learned factor is uncontrollable. Instead, our method is able to obtain a controllable disentanglement of structure and appearance in a completely unsupervised manner.

3. Methodology

Given two image domains X and Y. The goal of our work is to learn a pair of mapping $\Phi_{X\to Y}$ and $\Phi_{Y\to X}$ that could transfer an input $x \in X$ to a sample $y = \Phi_{X\to Y}(x)$, $y \in Y$, and vice versa. This problem formulation is a typical unpaired cross-domain image translation task, where the biggest challenge lies in tasks that require geometric changes [52, 10]. Most existing frameworks try to parameterize these pairs of mapping through two neural networks, *e.g.*, ResNet [12] or HourGlass [31].], of which the optimization is hard under complicated scenarios. In this study, we assume each domain can be disentangled into a Cartesian of structure space $G.$ and appearance space $A.$. Then on each space, we build a transition between the two domains, *i.e.*, *geometry transformer* $\Phi^g_{X\to Y}$ and $\Phi^g_{Y\to X}$ for geometry space and *appearance transformer* $\Phi^a_{X\to Y}$ and $\Phi^a_{Y\to X}$ for appearance space. Figure 2 illustrates the frame-

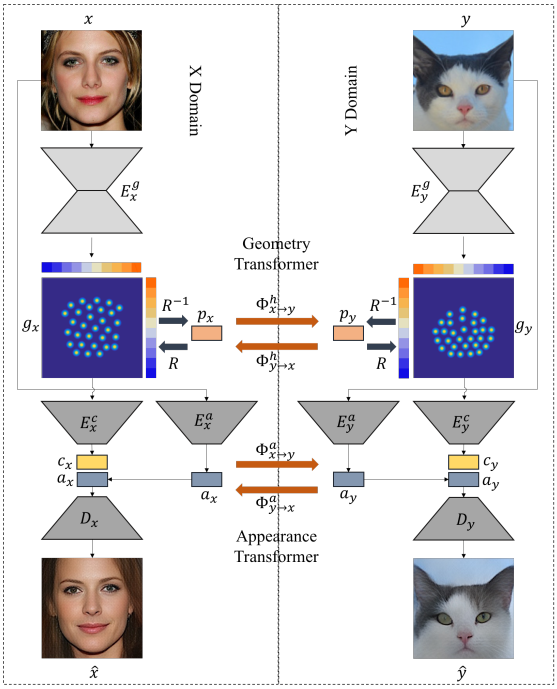

Figure 2: **Architecture.** Our framework consists of four main components: two auto-encoders (X/Y domain) and two transformers (geometry/appearance). *Auto-Encoder*: Taking X domain for example. For the input x, we use an encoder E_x^g to obtain the geometry representation g_x, which is a 30-channel point-heatmap with the same resolution as x. We project all channels of g_x together for visualisation. Then, g_x is embedded again to get the geometry code c_x. At the same time, x is also embedded by appearance encoder E_x^a to get the appearance code a_x. Finally, a_x and c_x are concatenated together to generate $\hat{x}$ with D_x. *Transformer*: For cross-domain translation, geometry ($g_x \leftrightarrow g_y$) and appearance ($a_x \leftrightarrow a_y$) transformation are performed separately.

work of our proposed approach.

3.1. Learning Disentangled Structure and Style Encoders

Unlike previous works that employ an encoder-decoder structure aiming at encoding all the information using one convolutional network [52, 50], our approach tries to encode geometry structure and the appearance style separately. To achieve this, we apply a conditional variational autoencoder in each domain. The conditional VAE system consists of an unsupervised geometry estimator $E_\cdot^g(;\pi)$, a geometry encoder $E_\cdot^c(;\theta)$ which embeds the heatmap structure into the latent space $C_\cdot$, an appearance encoder $E_\cdot^a(;\phi)$ which embeds the appearance information into the latent space $A_\cdot$, and a decoder $D_\cdot(;\omega) : C_\cdot \times A_\cdot \rightarrow X/Y$, which maps the latent space back to the image space. To disentangle two representations in an unsupervised manner, we formulate our loss as the combination of a conditional VAE loss and a prior loss for geometry estimation, which is

$$\mathcal{L}_{\text{disentangle}} = \mathcal{L}_{\text{CVAE}} + \mathcal{L}_{\text{prior}}. \tag{1}$$

Inspired by previous literature [21, 36, 8], we implement the conditional VAE loss as:

$$\begin{aligned}\mathcal{L}_{\text{CVAE}}(\pi,\theta,\phi,\omega) = &-KL(q_\phi(c|x,g)||p(a|x)) \\ &+\|x - D(E^c(E^g(x)), E^a(x))\|,\end{aligned} \tag{2}$$

where the first term is the KL-divergence loss between two parametric Gaussian distributions and the second term is a reconstruction loss. Here we replace it with the perceptual loss of a VGG-16 [35] network. In the supervised manner, $\mathcal{L}_{\text{CVAE}}$ can facilitate the learning of a complementary representation of geometry and appearance as described in [8]. However, in our unsupervised scenario, one cannot guarantee any branch of encoders to learn the geometry information without the supervision of geometry map $g_\cdot$. Next we will introduce our prior loss to constrain the geometry estimator.

3.2. Prior Loss for Geometry Estimator

Contrary to existing literature that use a content encoder to embed all of the detailed contents [27, 23], our geometry estimator $E_\cdot^g$ tries to distil pure geometry structure information as a stack of landmark heatmap. To achieve this, we rely on prior knowledge of how object landmarks should distribute to constrain the learning of our structure estimator E_x^g and E_y^g as described in [49, 16]. These previous work has shown that it is possible when given appropriate prior losses and learning architecture.

We now introduce the set of prior losses we used:

$$\mathcal{L}\text{prior} = \sum_{i \neq j} \exp\left(-\frac{||g^i - g^j||^2}{2\sigma^2}\right) + \text{Var}(g) \tag{3}$$

The first term is a *Separation Loss*. Similar to the difficulty described in [49], we find that training the structure branch with general random initialization tend to locate all structural points around the mean location at the center of the image. This could lead to a local minimum from which optimizer might not escape. As such, we introduce the separation loss to encourage each heatmap to sufficiently cover the object of interest. This is achieved by the first part in Eq. 3, where we encourage each pair of i^{th} and j^{th} heatmaps to share different activations. σ can be regarded as a normalization factor here. The second term is a *Concentration Loss*, which we introduce to encourage the variance of activations g to be small so that it could concentrate at a single location. This corresponds to the second term in Eq. 3.

The geometry prior, which is an explicit presentation of object shape, is important for a fine disentanglement of appearance and geometry. As shown in Fig. 3, with the ge-

ometry maps as the conditional input, our method can generate different shapes of face, which are consistent with geometry maps while maintaining the appearance of one specific input. It indicates that by estimating the pure geometry cues of objects, our method can disentangle geometry and appearance within a domain in a completely unsupervised manner.

3.3. Appearance Transformer

With the disentangled appearance geometry space, we can discompose the image translation into two separate problems. In this section, we first consider the transformation Φ^a on the appearance latent space A_X and A_Y. One may address this latent to latent transformation problem as a CycleGAN [52], with the cycle consistency loss and the adversarial loss. However, this does not guarantee g_x and mapped appearance transformer $\Phi_{X\to Y}(g_x)$ associated with two images to have a visual relationship. Since these two constraints can only lead to a translation between two distributions, which is arbitrary and multimodal. To this end, we introduce a cross-domain appearance consistency loss to constrain the appearance transformer:

$$\mathcal{L}^a_{\text{con}} = \|\zeta(x) - \zeta(D_y\left(\Phi^g_{x\to y}\cdot E^g_x(x), \Phi^a_{x\to y}\cdot E^a_x(x)\right))\|, \tag{4}$$

where ζ is the *Gram matrix* [9, 17] calculated with a pretrained VGG-16 [35] network, $\Phi^g_{x\to y}\cdot E^g_x(x)$ is the geometry code transformed from X to Y, $\Phi^a_{x\to y}\cdot E^a_x(x)$ is the appearance code transformed from X to Y, and $D_y(,)$ refers to decoder of Y domain. This loss ensures the image associated with g_x and translated appearance $\Phi_{X\to Y}(g_x)$ to have a similar appearance. In our experiment, it is observed that CycleGAN without appearance constraint can also converge, but it yields different results in each time of training with the same settings. The appearance consistency constraint stables the training and provide a more explainable results.

Single and Multimodal Transition: In our framework, the transform function is learned both in appearance and geometry latent spaces. For single-modal translation, the appearance transform Φ^a is constrained to guarantee transformed samples to have an associated appearance on the image domain. However, as aforementioned, a complex transformation problem is always multimodal. In our method, by replacing the transformed appearance representation by any feasible vector in the target appearance space A, we can achieve the results for multimodal generation. For example, with only the geometry transform Φ^g, by taking different human faces as a reference, we can obtain different results by just one cat face input. The multimodal ability is brought by the fine-disentangled representation within the domain. Qualitative results are shown in Sec. 4.2.

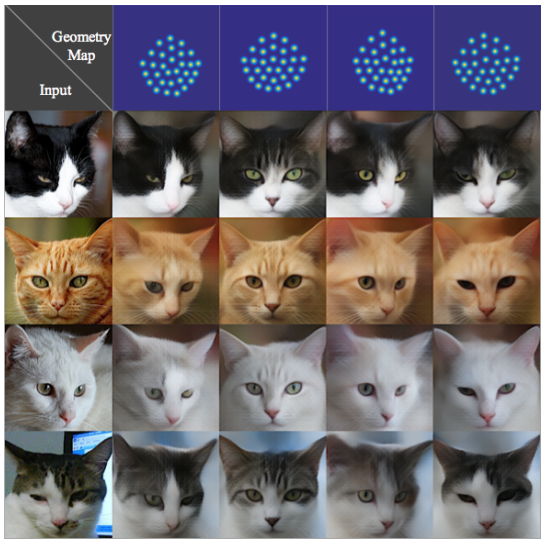

Figure 3: **Disentangled representation.** The top-row shows the corresponding geometry heatmaps of the faces in the left-most column. We illustrate the explicitly disentangled latent space with a grid of structure&appearance swapping results. In each column, the shape of the generated images is shown to be consistent with the geometry heatmaps. In each row, the appearance of the generated images are shown to be consistent with the left-most ones.

3.4. Geometry Transformer

We found it hard to learn a transfer between unsupervised learned geometry heatmaps directly since CNNs are usually not well-suited at capturing geometry information. Instead, we extract the coordinates information of each landmark from the heatmaps directly with the differentiable re-normalisation operator [16] R. Thus, the *de facto* geometry transformation is performed in the landmark coordinate space.

Specifically, for each landmark's heatmap, we compute a weighted average of coordinates over all activations across each heatmap. Although the dimensionality of landmarks with 2D coordinates is lower than the image representation, we still use the PCA to reduce dimensions of the landmark representation. The reason behind it is that we observe the result is more sensitive to small errors in geometries than in image pixel values, since slight errors of coordinates may cause severe artifacts (*e.g.* foldover and zigzag contour). It indicates that geometry translation is sometimes harder than image translation.

It is noteworthy that we have tried three kinds of representations for Geometry Transformer (*i.e.*, geometry heatmaps, landmark coordinates and PCA embedding of coordinates). All of the three representations can be used for training in our experiments. PCA embedding of coordinates works best in terms of *stability* and *convergence* of model training while other representations sometimes fail in some specific tasks. PCA constrains the geometry structure in

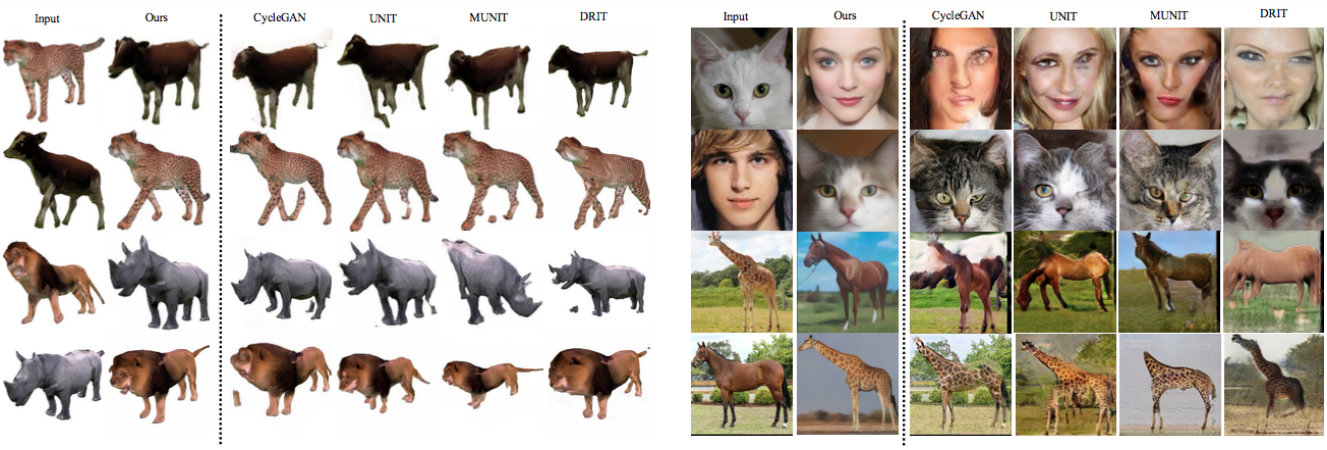

Figure 4: **Comparison in geometry-preserving.** Results on (a) synthesis datasets (cow↔cheetah and lion↔rhino) (b) real-world datasets (cat↔human face and giraffe↔horse). From left to right: input, ours, CycleGAN [52], UNIT [27], MUNIT [14] and DRIT [23].

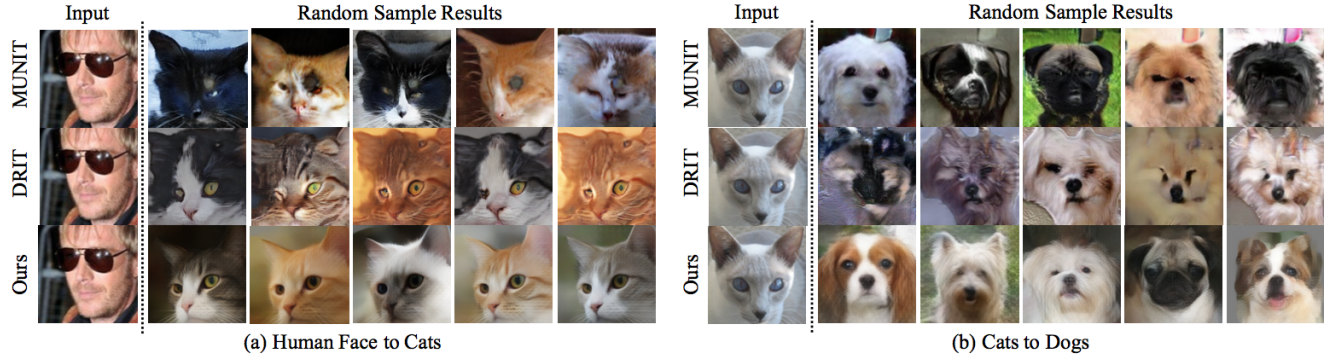

Figure 5: **Comparison in multi-modal generation.** Results on (a) human→cat face (b) cat→dog face. From top to bottom: MUNIT[14], DRIT [23] and ours (zoom in for more details).

the output. It constructs an embedding space for geometry shapes, where each principal component represents a reasonable dimension. Therefore, any sample in the embedding space will maintain the basic object structure, which reduces the risk of mode collapse.

To incorporate the PCA landmark representation with GAN, we replace all Conv-ReLU blocks with FC-ReLU blocks in both generators and discriminators. Though we incorporate a similar transformer structure as in CariGANs [4], our work differs in that unlike CariGANs that uses landmarks' PCA embeddings directly as the source and target domain defined in CycleGAN, we train the corresponding cycle on image pixel level as discussed in Sec. 3.4, which is more direct and powerful for pose-preserving generation task.

3.5. Other Constraints

Other than the proposed geometry prior loss and style consistency loss, we also leverage cycle-consistency and adversarial loss functions to facilitate the model training.

Cycle-consistency Loss. We apply three types of cycle-consistency loss, *i.e.*, $\mathcal{L}^{a}_{\text{cyc}}$, $\mathcal{L}^{g}_{\text{cyc}}$ and $\mathcal{L}^{pix}_{\text{cyc}}$. These three types of cycle-consistency constraints are performed in the geometry space, appearance space and pixel space respectively. Our ablation study in Sec. 4.3 demonstrates that cycle-consistency constraints are important for pose-preserving in translation.

Adversarial Loss. We impose adversarial losses $\mathcal{L}^{a}_{\text{adv}}$, $\mathcal{L}^{g}_{\text{adv}}$ and $\mathcal{L}^{pix}_{\text{adv}}$, which correspond to the geometry, appearance and pixel space, respectively. LSGAN is used for more stable training and convergence.

Total Loss. In summary, the full loss function of our method is:

$$\begin{aligned}\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CVAE}} + \mathcal{L}_{\text{prior}} + \mathcal{L}^{a}_{\text{con}} + \mathcal{L}^{a}_{\text{cyc}} \\ + \mathcal{L}^{g}_{\text{cyc}} + \mathcal{L}^{pix}_{\text{cyc}} + \mathcal{L}^{a}_{\text{adv}} + \mathcal{L}^{g}_{\text{adv}} + \mathcal{L}^{pix}_{\text{adv}}\end{aligned} \tag{5}$$

More details of the implementation of these losses are described in the supplementary material.

4. Experiments

Datasets. We conduct extensive comparisons and ablation studies on four datasets that cover both synthesis and real-world data. (1). Synthesis Animals: We use the publicly

Table 1: **Human perceptual study.** Pairwise A/B tests on horse→giraffe and human→cat face task.

Method	**horse → giraffe** % Testers labeled *better*	**human → cat face** % Testers labeled *better*
CycleGAN [52]	15.0%	15.4%
UNIT [27]	19.3%	18.9%
MUNIT [14]	20.4%	17.8%
DRIT [23]	16.1%	23.4%
Ours	**50.0**%	**50.0**%

(a) Score of "realism".

Method	**horse → giraffe** % Testers labeled *better*	**human → cat face** % Testers labeled *better*
CycleGAN [52]	11.9%	25.7%
UNIT [27]	16.5%	23.3%
MUNIT [14]	19.2%	31.7%
DRIT [23]	23.6%	34.4%
Ours	**50.0**%	**50.0**%

(b) Score of "geometry-consistency".

available CAD model provided by [54] to render six different non-rigid animals, *i.e*., Cheetah, Cow, Lion, Rhino, Bear and Wolf. For each population of animal, we rendered $10,000$ images (9000 for training and 1000 for testing) with different shapes through the randomly sampled parameters. (2). Real-world Animals: We collected 5000 images (4500 for testing and 500 for testing) of horse and giraffe from Flickr. (3). Unconstrained Face: We collected images of three typical domains, *i.e*., human, dog and cat faces. We randomly sampled $5,000$ images (4500 for testing and 500 for testing) from YFCC100M [18], Stanford Dog [19] and CelebA [28] datasets respectively. Note that the faces in each dataset are completely unconstrained rather than within four given modes in [14].

Baselines. We compare our approach with the four most related state-of-the-art methods: CycleGAN [52], UNIT [27], MUNIT [14] and DRIT [23], All of these methods can perform image-to-image translation with unpaired training data. In particular, MUNIT [14] and DRIT [23] can generate multimodal results. Thus, we compare to them also in multi-modal generation task. We trained these four baselines on the newly collected datasets with their public implementation with default settings.

Evaluation Metric. For quantitative comparison, we evaluate both the realism and diversity of the generated images. Following [41, 45], we perform human subjective study for geometry-consistency/realism evaluation. To measure visual quality, rather than general image quality assessment methods [44, 25, 26] or perceptual loss [50], Fréchet Inception Distance (FID) [2] is adopted. To measure diversity, similar to [53, 14], we use the LPIPS metric [48] to calculate the distance among images.

Implementation Details. Images of all datasets are cropped and resized to 256×256. Taking X domain for example. We adopt the architecture for our structure encoder E_x^g from Stack-Hourglass network [31] which have shown impressive results for landmark localisation task [7, 3]. For the mapping from g_x to $\hat{x}$ (E_x^c and D_x with skip-connection), we use the UNet architecture [33] provided by [52]. The same architecture of E_x^c is adopted for the appearance encoder E_x^a. We use a simple 4-layer fully-connection network followed with ReLU for the transformer $\Psi_{X \leftrightarrow Y}$ and the discriminators. For pixel level adversarial loss, we use the discriminator provided by [27].

We train our model in two main steps. First, to obtain the geometry heatmap $g_x(g_y)$, $E_x^a(E_y^a)$, $E_x^g(E_y^g)$ and $D_x(D_y)$ are trained together for 40 epochs. Then, structure encoders are frozen and all of the networks except E_x^g and E_y^g are trained end-to-end for 20 epochs. We train all of the models use the Adam [21] optimizer with initial $lr = 0.0001$ and $(\beta_1, \beta_2) = (0.5, 0.999)$ on eight NVIDIA V100 GPUs. More details on the training and network architecture are provided in the supplementary material.

4.1. Comparisons with State-of-the-Arts

Qualitative Comparison. Recall the motivation of our work: by introducing the unsupervised latent geometry representation, we hope our framework has a higher capacity for translation between more complicated objects. Here we perform visual quality comparison to state-of-the-art methods in Fig. 4. We evaluate the quality of generated results on both near-rigid (*i.e*., faces) and non-rigid (animals) objects. Our approach is able to achieve superior results to all of the baselines. Although results of the baselines are recognizable to be settled in the target domain, the geometry tends to be broken due to the neglect of geometry cues. For near-rigid objects, the baselines are likely to yield distorted results. For non-rigid objects, which are more challenging due to the large inter and intra-domain shape variations, the baselines always obtain results with missing parts. By contrast, the translations by our approach are more robust to large shape variations and unconstrained appearance in both rigid and non-rigid scenarios.

For multimodal generation, we compare our approach with MUNIT [14] and DRIT [23] in Fig. 5. Both of the baselines can obtain diverse outputs. However, in some unconstrained scenarios, *e.g*., profile face with sun-glass and large face shape difference between domains, the results of baselines degrade and suffer from severe artifacts. It can be observed that our approach achieve better visual quality than others. More results on other datasets are demonstrated in the supplementary material.

Quantitative Comparison. We use both subjective and objective metrics for the quantitative performance evaluation. For the realism of generation images, we ask volunteers to perform subjective pairwise A/B tests. Following the metric in MUNIT [14], the preference score of our work indicates the percentage that one method (CycleGAN [52], UNIT [27], MUNIT [14], DRIT [23]) is preferred over *our method*. For each time of a test, partici-

Table 2: **Quantitative Results.** We use FID (lower is better) and diversity (higher is better) with LPIPS distance to evaluate the quality and diversity of the generated images.

	Real Data		**CycleGAN [52]**		**UNIT [27]**		**MUNIT [14]**		**DRIT [23]**		**Ours**	
	FID	Diversity	FID	Diversity	FID	Diversity	FID	Diversity	FID	Diversity	FID	Diversity
cats → human face	0.00	0.54	57.92	-	98.39	-	40.91	**0.41**	69.53	0.20	**32.25**	0.39
human face → cats	0.00	0.65	44.23	-	35.26	-	23.24	0.53	33.14	0.52	**21.88**	**0.56**
cats → dogs	0.00	0.66	143.14	-	104.32	-	100.26	0.59	67.01	0.54	**65.77**	**0.60**
dogs → cats	0.00	0.65	75.75	-	66.84	-	27.60	0.56	31.04	**0.59**	**23.23**	0.58
dogs → human face	0.00	0.54	105.09	-	103.35	-	37.84	0.40	46.70	0.32	**31.06**	**0.41**
human face → dogs	0.00	0.66	149.61	-	91.38	-	73.98	0.60	68.84	0.57	**52.20**	**0.67**
Average	0.00	0.62	95.96	-	83.26	-	50.64	0.52	52.71	0.46	**37.73**	**0.54**

pants can vote for A/B/Not Sure. Two metrics are evaluated as shown in Table. 1, *realism* for evaluation the similarity with real data while *geometry-consistency* for evaluation the geometry consistency with input image. Participants were given 10 seconds to choose which image has better *realism* or *geometry-consistency* in a pair of generated images from two different methods. All 500 test images of each dataset are compared 100 times by different participants. Our method obtains the highest preference rate.

For the evaluation of visual quality and diversity, following [51], we use 100 input images in the test set and sample 19 output pairs per input. We compute the average LPIPS distance in ImageNet pre-trained AlexNet feature space between the $1,900$ pairs of images. FID is calculated between the real data and the generated results. As shown in Table 2, our method significantly outperforms all of the baselines both in visual quality and diversity. In particular, even though MUNIT and DRIT obtain reasonable performance in diversity, they get a poor score in the subjective metric, suggesting the shortcoming of these methods in handling translation across a large geometry gap.

4.2. Representation Disentanglement

Exemplar-guided Image Translation. In Fig. 6, we illustrate the exemplar-guided translation results of several typical shapes of faces, *e.g.*, frontal, profile, eye-closed and mouth-opening. From input to output, we observe that the geometry feature maintains faithfully. Thanks to the pure geometry representation translation schema, which endow the model with the ability of appearance-agnostic image-to-image translation. In addition, once the geometry is translated successfully, the model can take images in the target domain as exemplars to guide multimodal generations. Results in Fig. 6 show successful *disentanglement* of the geometry and appearance in two aspects. First, the geometry maintains to be the same no matter what shape of the exemplar is. As a concrete example, as shown in Fig. 6 (b), the generated faces maintain to be profile even with large variations of the exemplars. Second, the appearance of exemplars can be successfully transferred to the generated images, even for the detail textures, *e.g.*, the beard of the man in Fig. 6 (a) and the blue eyes of the cat in Fig. 6 (d).

Interpolation. To evaluate whether disentangled latent space is densely populated, we perform a linear interpolation to geometry code and appearance code respectively in Fig. 7. The interpolation results show that both the geometry and appearance of images can change smoothly along with the latent space from source to target. It is noteworthy that the datasets have only one geometry and appearance for each sample and only discrete features are supplied from standalone individuals in raw datasets. The smooth interpolation results show that our model has captured a reasonable coverage of the manifold successfully.

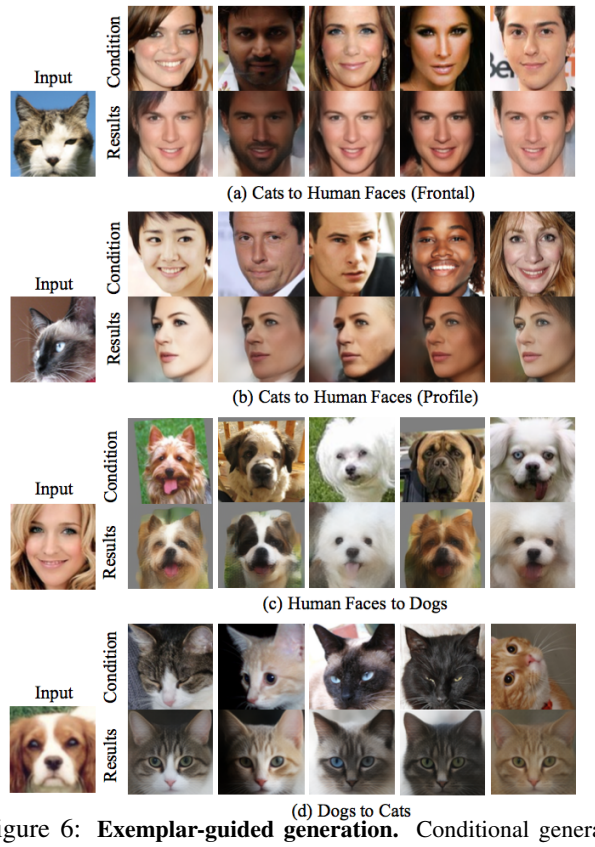

Figure 6: **Exemplar-guided generation.** Conditional generation with different images as appearance reference on cat→human face, human→dog face, dog→cat face tasks.

Table 3: **Ablation study.** Fooling rate of "real vs fake".

Method	**horse → giraffe** % Testers labeled *real*	**human → cat face** % Testers labeled *real*
Ours w/o Trans.	0.0%	0.0%
Ours w/o $\mathcal{L}_{\text{cyc}}$	14.2%	16.8%
Ours w/o KL	10.2%	15.4%
Ours w/o VGG	12.6%	18.4%
Ours	**16.2**%	**23.9**%

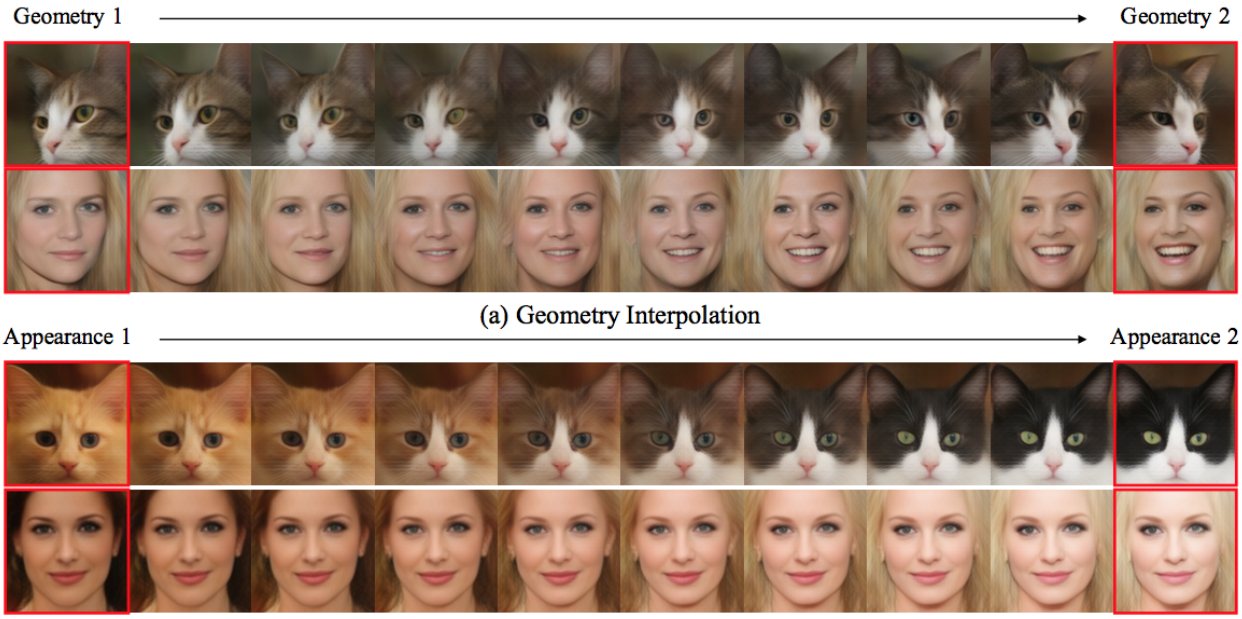

Figure 7: **Interpolation.** Linear interpolation results of geometry and appearance latent code on cat and human face datasets.

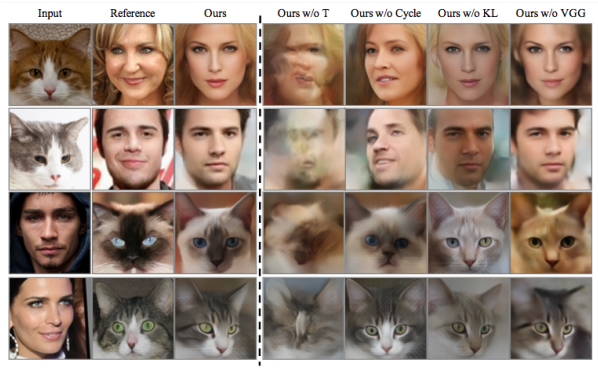

Figure 8: **Quantitative ablation study.** Visualisation results on human↔cat task.

4.3. Ablation Study

To isolate the effectiveness of pivotal components of our method, we perform an ablation study on the quality of generated images. We evaluate several variants of our method: 1) Ours w/o T: our methods without appearance and geometry transformers. 2) Ours w/o cycle: our methods without the cycle-consistency loss term. 3) Ours w/o KL: our methods without the KL loss term. 4) Ours w/o VGG: replacing VGG loss with L1 loss in our method.

Figure 8 shows the qualitative results of the variants. Without the transformer, our method is unable to generate plausible results to cross the large gap of the geometry representation between two domains. Without using the cycle-consistency loss, our method can still obtain plausible results. However, the pose-consistency with the input image cannot be guaranteed, suggesting that the cycle-consistency loss is a key component for pose-preserving. Without using the KL loss, the consistency with the reference image cannot be maintained. Without the VGG loss, we obtain blurred results, which is consistent with the observation of [34, 8].

We quantify these observations with perceptual studies on the human→cat face and horse→giraffe task in Table 3. The scores obtained by our method on these two tasks demonstrate its capability in generating realistic results. Note that without cycle-consistency loss, a comparable perceptual score can also be achieved with our method, which indicates that this loss is more important for pose-preserving than generated quality.

5. Conclusion

We have presented a novel geometry-aware disentangle-and-translate framework for image-to-image translation, in which we introduced an unsupervised geometry latent branch based on CycleGAN system. Specifically, we first disentangled each domain on the geometry space and appearance space, then established the translation on each latent space. Our extensive qualitative and quantitative experiments showed that our method is effective for translation between objects with complex structures. Moreover, our model can also support multimodal translation and outperform previous state-of-the-art methods. Future work includes extending this framework to more unconstrained scenarios, such as images in ImageNet and YouTube videos.

Acknowledgement. We would like to thank Kwan-Yee Lin and Jingtan Piao for insightful discussion and their exceptional support.

References

[1] Guha Balakrishnan, Amy Zhao, Adrian V. Dalca, Frédo Durand, and John Guttag. Synthesizing images of humans in unseen poses. In *CVPR*, 2018.

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

[3] Adrian Bulat and Georgios Tzimiropoulos. Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans. In *CVPR*, 2018.

[4] Kaidi Cao, Jing Liao, and Lu Yuan. Carigans: Unpaired photo-to-caricature translation. In *Siggraph Asia*, 2018.

[5] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *NIPS*, 2016.

[6] Xipeng Chen, Kwan-Yee Lin, Wentao Liu, Chen Qian, and Liang Lin. Weakly-supervised discovery of geometry-aware representation for 3d human pose estimation. In *CVPR*, 2019.

[7] Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan L. Yuille, and Xiaogang Wang. Multi-context attention for human pose estimation. In *CVPR*, 2017.

[8] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. In *CVPR*, 2018.

[9] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. In *NIPS*, 2015.

[10] Aaron Gokaslan, Vivek Ramanujan, Daniel Ritchie, Kwang In Kim, and James Tompkin. Improving shape deformation in unsupervised image-to-image translation. *ECCV*, 2018.

[11] 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*, 2014.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[13] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017.

[14] Xun Huang, Ming-Yu Liu, Serge J. Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018.

[15] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

[16] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Conditional image generation for learning the structure of visual objects. In *NeurIPS*, 2018.

[17] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016.

[18] Sebastian Kalkowski, Christian Schulze, Andreas Dengel, and Damian Borth. Real-time analysis and visualization of the yfcc100m dataset. In *MM Workshop*, 2015.

[19] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *CVPR Workshop*, 2011.

[20] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *ICML*, 2017.

[21] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

[22] 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 *CVPR*, 2017.

[23] 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*, 2018.

[24] Jianxin Lin, Yingce Xia, Tao Qin, Zhibo Chen, and Tie-Yan Liu. Conditional image-to-image translation. In *CVPR*, 2018.

[25] Kwan-Yee Lin and Guangxiang Wang. Hallucinated-iqa: No-reference image quality assessment via adversarial learning. In *CVPR*, 2018.

[26] Kwan-Yee Lin and Guangxiang Wang. Self-supervised deep multiple choice learning network for blind image quality assessment. In *BMVC*, 2018.

[27] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *NIPS*, 2017.

[28] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015.

[29] Liqian Ma, Xu Jia, Stamatios Georgoulis, Tinne Tuytelaars, and Luc Van Gool. Exemplar guided unsupervised image-to-image translation. In *NeurIPS*, 2018.

[30] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. In *CVPR*, 2018.

[31] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016.

[32] Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. 2016.

[33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

[34] Mehdi S. M. Sajjadi, Bernhard Schölkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *ICCV*, 2017.

[35] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[36] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *NIPS*, 2015.

[37] James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object frames by dense equivariant image labelling. In *NIPS*, 2017.

[38] James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks by factorized spatial embeddings. In *ICCV*, 2017.

[39] Aaron van den Oord, Nal Kalchbrenner, Lasse Espeholt, koray kavukcuoglu, Oriol Vinyals, and Alex Graves. Conditional image generation with pixelcnn decoders. In *NIPS*, 2016.

[40] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In *NeurIPS*, 2018.

[41] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, 2018.

[42] Wei Wang, Xavier Alameda-Pineda, Dan Xu, Pascal Fua, Elisa Ricci, and Nicu Sebe. Every smile is unique: Landmark-guided diverse smile generation. In *CVPR*, 2018.

[43] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. ESRGAN: enhanced super-resolution generative adversarial networks. In *ECCV Workshop*, 2018.

[44] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing (TIP)*, 13(4):600–612, 2004.

[45] Wayne Wu, Yunxuan Zhang, Cheng Li, Chen Qian, and Chen Change Loy. Reenactgan: Learning to reenact faces via boundary transfer. In *ECCV*, 2018.

[46] Zili Yi, Hao (Richard) Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *ICCV*, 2017.

[47] Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. In *ECCV*, 2016.

[48] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[49] Yuting Zhang, Yijie Guo, Yixin Jin, Yijun Luo, Zhiyuan He, and Honglak Lee. Unsupervised discovery of object landmarks as structural representations. In *CVPR*, 2018.

[50] 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*, 2017.

[51] 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*, 2017.

[52] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networkss. In *ICCV*, 2017.

[53] 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*, 2017.

[54] Silvia Zuffi, Angjoo Kanazawa, and Michael J. Black. Lions and tigers and bears: Capturing non-rigid, 3d, articulated shape from images. In *CVPR*, 2018.