Exploring Unlabeled Faces for Novel Attribute Discovery

Hyojin Bahng[1] Sunghyo Chung[2] Seungjoo Yoo[1] Jaegul Choo[3]
[1]Korea University [2]Kakao Corp. [3]KAIST

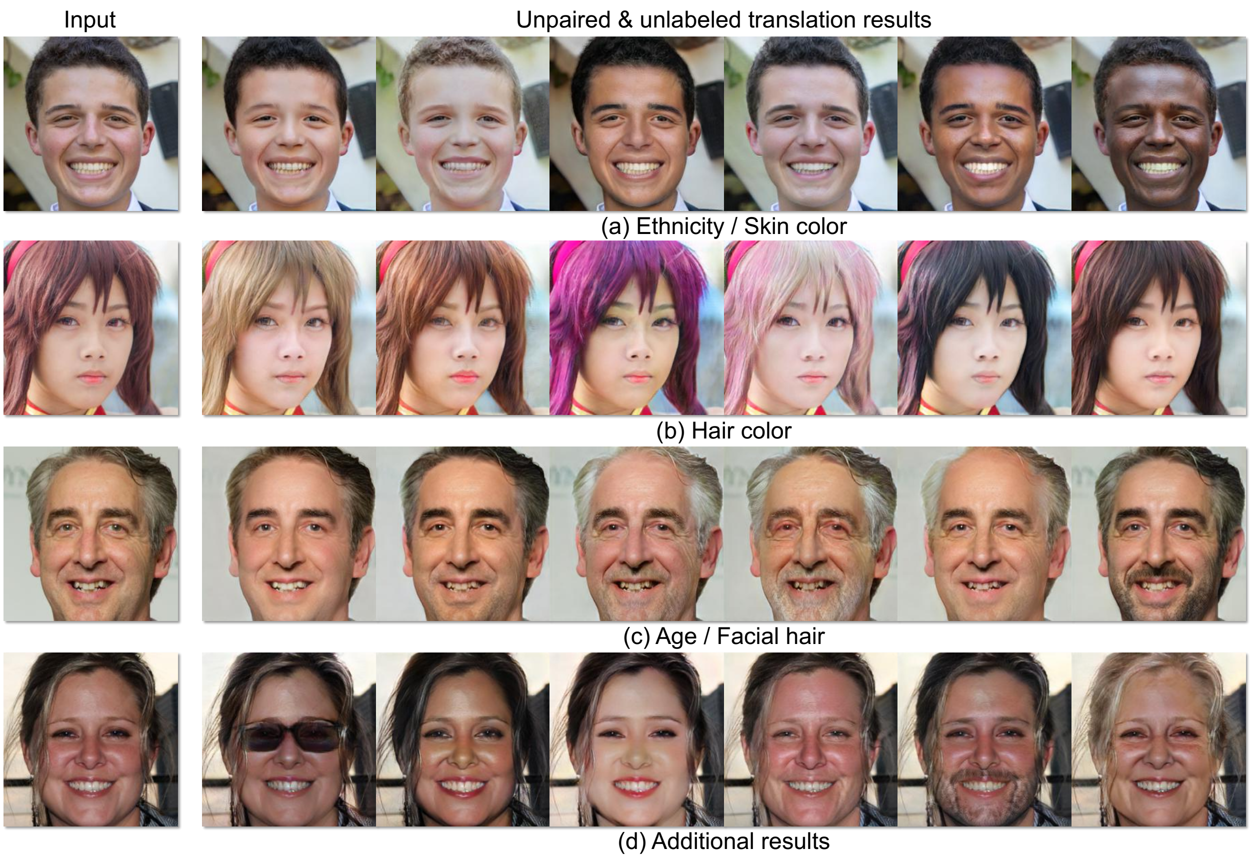

Figure 1: Given raw, unlabeled data, our algorithm discovers novel facial attributes and performs high-quality multi-domain image translation. All results are based on *newly-found* attributes from our algorithm (e.g., a wide range of ethnicity, skin and hair color, age, facial hair, accessories, and makeup). We did not use any pre-defined attribute labels to generate the results.

Abstract

Despite remarkable success in unpaired image-to-image translation, existing approaches still require a large amount of labeled images. This is a bottleneck against their real-world applications; in practice, a model trained on a labeled dataset, such as CelebA dataset, does not work well for test images from a different distribution – limiting their applications to unlabeled images of a much larger quantity. In this paper, we attempt to alleviate this necessity for labeled data in the facial image translation domain. We aim to explore the degree to which you can discover novel attributes from unlabeled faces and perform high-quality translation. To this end, we use prior knowledge about the visual world as guidance to discover novel attributes and transfer them via a novel normalization method. Experiments show that our method trained on unlabeled data produces high-quality translations, preserves identity, and is perceptually realistic, as good as, or better than, state-of-the-art methods trained on labeled data.

1. Introduction

In recent years, unsupervised image-to-image translation has improved dramatically [45, 4, 23, 17]. Existing translation methods use the term *unsupervised* for translating with *unpaired* training data (i.e., provided with images in domain X and Y, with no information on which x matches which y). However, existing systems, in essence, are still trained with supervision, as they require a large amount of *labeled* images to perform translation. This acts as a bottleneck against their applications in the real world. In practice, a model trained on the labeled CelebA dataset [27] does not work well for images of a different test distribution due to dataset bias [36, 39]. For instance, a model trained on CelebA images are biased towards Western, celebrity faces, which necessitates collecting, labeling, and training with new data to match a different test distribution. Hence, the need for labels greatly limits their applications to unlabeled images of a much larger quantity.

In this paper, we attempt to alleviate the necessity for labeled data by automatically discovering novel attributes from unlabeled images – moving towards *unpaired* and *unlabeled* multi-domain image-to-image translation. In particular, we focus on image translation of facial images, as they require annotation of multiple attributes (e.g., 40 attributes for 202,599 images in CelebA), which makes labeling labor- and time-intensive. While existing benchmark datasets attempt to label as many attributes as they can, we notice that much is still unnamed, e.g., CelebA only contains 'pale skin' attribute among all possible skin colors. This makes us wonder: *can't we make the attributes "emerge" from data?*

This paper aims to explore the degree to which you can discover novel attributes from unlabeled faces X, thus proposing our model called XploreGAN. To this end, we utilize pre-trained convolutional neural network (CNN) features – making the most out of *what we have already learned* about the visual world. Note that classes used for CNN pre-training (ImageNet classes) differ from the unlabeled data (facial attributes). The goal is to transfer not its specific classes, but the general knowledge on what properties make a good class in general [13]. We use it as guidance to group a new set of unlabeled faces, where each group contains a common attribute, and transfer that attribute to an input image by our newly proposed attribute summary instance normalization (ASIN). Unlike previous style normalization methods that generate affine parameters from a single image [7, 16], resulting in translation of *entangled* attributes (e.g., hair color, skin color, and gender) that exist in the style image, ASIN summarizes the common feature (e.g., blond hair) among a group of images (cluster) and only transfers its common attribute (style) to the input (content). Experiments show that XploreGAN trained on *unlabeled* data produces high-quality translation results as good as, or better than, state-of-the-art methods trained with *labeled* data. To the best of our knowledge, this is the first method that moves towards both *unpaired* and *unlabeled* image-to-image translation.

2. Proposed Method

While existing methods use facial images that annotate a single image with multiple labels (i.e., one-to-many mapping) to achieve multi-domain translation, we slightly modify this assumption to achieve high-quality performance with no attribute labels at all. We first utilize a pre-trained feature space as guidance to cluster unlabeled images by their common attribute. Using the cluster assignment as *pseudo-label*, we utilize our newly proposed attribute summary instance normalization (ASIN) to summarize the common attribute (e.g., blond hair) among images in each cluster and perform high-quality translation.

2.1. Clustering for attribute discovery

The features extracted from a pre-trained CNN on ImageNet [5] have been used to assess *perceptual similarity* among images [19, 43]. In other words, images with similar pre-trained features are perceived as similar to humans. Exploiting this property, we propose to discover novel attributes existing in unlabeled data by clustering their feature vectors obtained from pre-trained networks and using these cluster assignments as our *pseudo-label* for attributes. In other words, we utilize the pre-trained feature space as guidance to group images by their dominant attributes.

We adopt a standard clustering algorithm, k-means, and partition the features from pre-trained networks $\{f(x_1), ..., f(x_n)\}$ into k groups by solving

$$\min_{\mu, C} \sum_{i=1}^{k} \sum_{x \in C_i} \|f(x) - \mu_i\|_2^2, \tag{1}$$

which results in a set of cluster assignments C, centroids μ, and their standard deviations σ. We use C as pseudo-labels for training the auxiliary classifier of the discriminator and use μ and σ for conditioning the normalization layer of the generator in our generative adversarial networks (GANs).

2.2. Attribute summary instance normalization

Normalization layers play a significant role in modeling style. As Huang et al. [16] put it, a single network can "generate images in completely different styles by using the *same* convolutional parameters but *different* affine parameters in instance normalization layers". That is, to inject a particular style to a content image, it is sufficient to simply tune the scaling and shifting parameters corresponding to the style after properly normalizing the content image.

Previous style normalization methods generate affine parameters from a single image instance [7, 16], resulting in translation of *entangled* attributes (e.g., hair color/shape, skin color, and gender) that exist in the given style image. In contrast, our approach summarizes and transfers the *common attribute* (e.g., blond hair) within a group of images by generating affine parameters from the feature statistics of each cluster. We call this *attribute summary instance normalization* (ASIN). We use a multilayer perceptron (MLP) f to map cluster statistics to the affine parameters of the normalization layer, defined as

$$\mathrm{ASIN}(x;\mu_k,\sigma_k) = f_\sigma(\sigma_k)\left(\frac{x-\mu(x)}{\sigma(x)}\right) + f_\mu(\mu_k). \quad (2)$$

As the generator is trained to generalize the common feature among each subset of images (cluster), ASIN allows us to discover multiple attributes in unlabeled data. ASIN can also be used in *supervised* settings to summarize the common attribute among images with the same label (e.g., black hair). You may generate affine parameters from both the centroid and the variance of each cluster, only the centroid information, or the domain pseudo-label (i.e., cluster assignments). We will use the first option in subsequent equations in the paper, so as not to confuse the readers.

2.3. Objective function

Cluster classification loss. To translate an input image x to a target domain k, we adopt a domain classification loss [4] to generate those images properly classified as its target domain. However, we use cluster assignments as *pseudo-labels* for each attribute unlike previous multi-domain translation approaches that utilize pre-given labels for classification [4, 33]. We optimize the discriminator D to classify real images x to its original domain k' via the loss function defined as

$$\mathcal{L}^r_{cls} = \mathbb{E}_{x,k'}[-\log D_{cls}(k' \mid x)]. \quad (3)$$

Similarly, we optimize the generator G to classify fake images $G(x,\mu_k,\sigma_k)$ to its target domain k via the loss function defined as

$$\mathcal{L}^f_{cls} = \mathbb{E}_{x,k}[-\log D_{cls}(k \mid G(x,\mu_k,\sigma_k)]. \quad (4)$$

The cluster statistics act as conditional information for translating images to its corresponding pseudo-domain.

Reconstruction and latent loss. Our generator should be sensitive to changes in content but robust to other variations. To make translated images preserve the content of its input images while changing only the domain-relevant details, we adopt a cycle consistency loss [22, 45] to the generator, defined as

$$\mathcal{L}_{rec} = \mathbb{E}_{x,k,k'}\left[\|x - G(G(x,\mu_k,\sigma_k),\mu_{k'},\sigma_{k'})\|_1\right], \quad (5)$$

where the generator is given the fake image $G(x,\mu_k,\sigma_k)$ and the original cluster statistics $\mu_{k'},\sigma_{k'}$ and aims to reconstruct the original real image x. We use the L_1-norm for the reconstruction loss.

However, solely using the pixel-level reconstruction loss does not guarantee that translated images preserve the high-level content of its original images in settings where a single generator has to learn a large number of domains simultaneously (e.g., more than 40). Inspired by Yang et al. [41], we adopt the latent loss, where we minimize the distance between real and fake images in the feature space, i.e.,

$$\mathcal{L}_{lnt} = \mathbb{E}_{x,k,k'}\left[\|h(x) - h(G(x,\mu_k,\sigma_k))\|_2\right]. \quad (6)$$

We denote h as the encoder of G and use the L_2-norm for the latent loss. The latent loss ensures that the real and the fake images have similar high-level feature representations (i.e., perceptually similar) even though they may be quite different at a pixel level.

Adversarial loss. We adopt the adversarial loss used in GANs to make the generated images indistinguishable from real images. The generator G attempts to generate a realistic image $G(x,\mu_k,\sigma_k)$ given the input image x and the target cluster statistics μ_k,σ_k, while the discriminator D tries to distinguish between generated images and real images. To stabilize GAN training, we adopt the Wasserstein GAN objective with gradient penalty [1, 11], i.e.,

$$\begin{aligned}\mathcal{L}_{adv} = \mathbb{E}_x\left[D_{adv}(x)\right] - \mathbb{E}_{x,\mathrm{k}}\left[D_{adv}(G(x,\mu_k,\sigma_k))\right] \\ -\lambda_{gp}\mathbb{E}_{\hat{x}}\left[(\|\nabla_{\hat{x}}D_{adv}(\hat{x})\|_2 - 1)^2\right],\end{aligned} \quad (7)$$

where $\hat{x}$ is sampled uniformly from straight lines between pairs of real and fake images.

Full objective function. Finally, our full objective function for D and G can be written as

$$\mathcal{L}_D = -\mathcal{L}_{adv} + \lambda_{cls}\mathcal{L}_{cls}, \quad (8)$$

$$\mathcal{L}_G = \mathcal{L}_{adv} + \lambda_{cls}\mathcal{L}_{cls} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{lnt}\mathcal{L}_{lnt}. \quad (9)$$

The hyperparameters control the relative importance of each loss function. In all experiments, we used $\lambda_{gp} = 10$, $\lambda_{rec} = 10$, and $\lambda_{lnt} = 10$. At test time, we used the pseudo-labels to generate translated results. It is surprising that the pseudo-labels correspond to meaningful facial attributes; results are demonstrated in Section 3.

2.4. Implementation details

Clustering stage. We use the final convolutional activations (i.e., `conv5` for BagNet-17 and ResNet-50) to cluster images according to high-level attributes. We use BagNet-17 [3] pre-trained on ImageNet (IN) [5] as the feature extractor for FFHQ [21] and CelebA [27] dataset, and ResNet-50 [14] pre-trained on Stylized ImageNet (SIN) [9] as the feature extractor of EmotioNet [8] dataset. The former is effective in detecting local texture cues, while the latter ignores texture cues but detects global shapes effectively. For clustering, the extracted features are L_2-normalized and PCA-reduced to 256 dimensions. We utilize the k-means implementation by Johnson et al. [20], with $k = 50$ for images with 256×256 resolution and $k = 100$ for images with 128×128 resolution.

Translation stage. Adapted from StarGAN [4], our encoder has two convolutional layers for downsampling, followed by six residual blocks [14] with spectral normalization [29]. Our decoder has six residual blocks with attribute summary instance normalization (ASIN), with per-pixel noise [21] added after each convolutional layer. It is then followed by two transposed convolutional layers for upsampling. We also adopt stochastic variation [21] to increase generation performance on fine, stochastic details of the image. For the discriminator, we use PatchGANs [24, 18, 44] to classify whether image patches are real or fake. As a module to predict the affine parameters for ASIN, our multi-layer perceptron consists of seven layers for FFHQ and EmotioNet datasets and three layers for CelebA dataset. For training, we use the Adam optimizer, a mini-batch size of 32, a learning rate of 0.0001, and decay rates of $\beta_1 = 0.5$, $\beta_2 = 0.999$.

3. Experiments

3.1. Datasets

Flickr-Faces-HQ (FFHQ) [21] is a high-quality human face image dataset with 70,000 images, offering a wide variety in age, ethnicity, and background. The dataset is not provided with any attribute labels.

CelebFaces Attributes (CelebA) [27] is a large-scale face dataset with 202,599 celebrity images, each annotated with 40 binary attribute labels. In our experiments, we do not utilize the attribute labels when training our model.

EmotioNet [8] contains 950,000 face images with diverse facial expressions. The facial expressions are annotated with action units, yet we do not utilize them for training our model.

3.2. Baseline models

We compare our approach with the baselines that utilize *unpaired* yet *labeled* datasets. For their implementations, we used the original source codes and hyperparameters. As XploreGAN does not use any labels during training, at test time, we select pseudo-labels that best estimate the labels used by other baseline models (e.g., the best pseudo-label corresponding to 'blond'). Each result of our model is generated from the statistics of a single cluster.

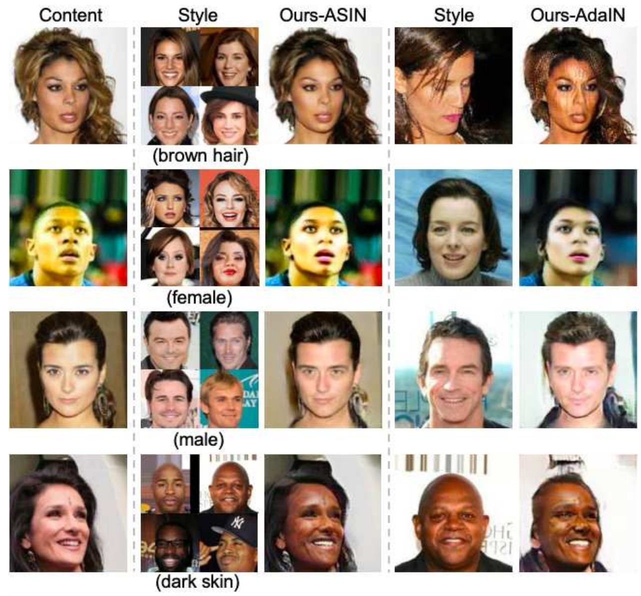

Figure 2: **Comparison on style normalization methods.** As AdaIN is conditioned on a single image instance to transfer style, it tends to translate entangled attributes of the style image (last three rows). In contrast, ASIN summarizes a common attribute within a group (cluster) of images and transfers its specific feature, while keeping all other attributes (identity) of the content image intact.

StarGAN is a state-of-the-art *multi-domain* image translation model that uses the attribute label during training.

DRIT and MUNIT are state-of-the-art models that perform *multi-modal* image translation between two domains.

3.3. Comparison on style normalization

We show qualitative comparisons of group-based ASIN and instance-based AdaIN. For fair comparison, we substitute the ASIN layers with AdaIN as implemented in [17] while maintaining all the other network architecture and training settings. As shown in Fig. 2, AdaIN depends on a single image instance to transfer style. AdaIN results in translation of entangled attributes (e.g., hair color/shape, gender, background color; last three rows of Fig. 2) that exist within the reference image. In contrast, ASIN is able to summarize the common attribute within a group of images (e.g., hair color) and transfer its specific attribute. This makes it easy for users to transfer a particular attribute they desire while preserving all other attributes (identity) of the content image intact.

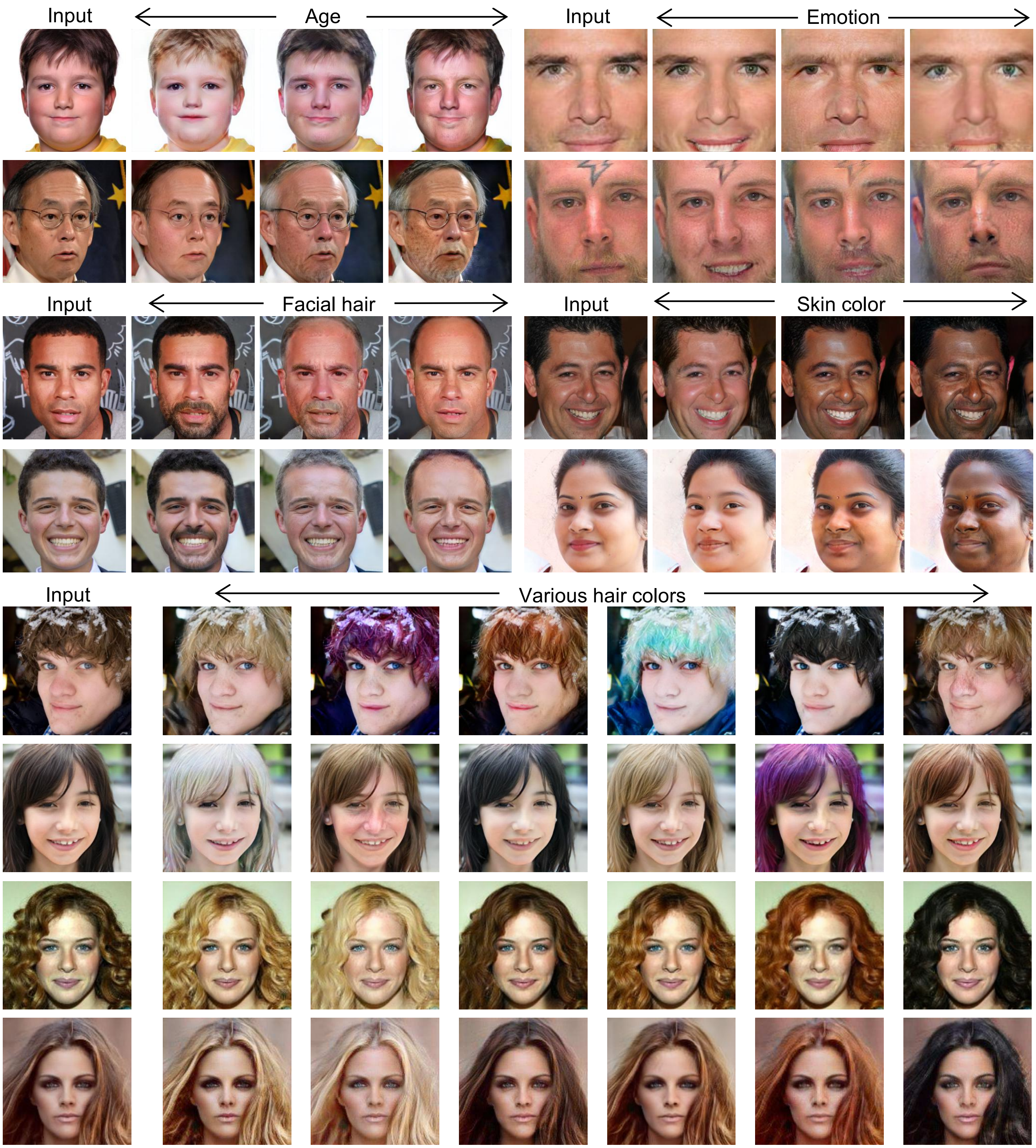

Figure 3: **Translation results from multiple datasets.** XploreGAN can discover various attributes in data such as diverse hair colors, ethnicity, degree of age, and facial expressions from unlabeled images. Note that labels in the figure are assigned post-hoc to enhance the interpretability of the results.

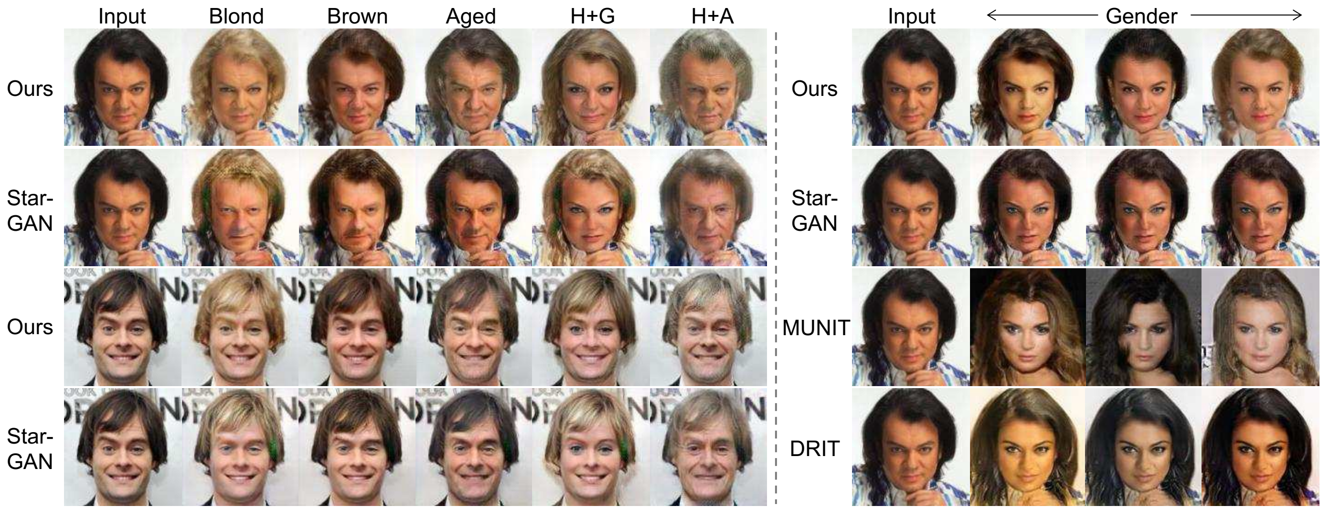

Figure 4: **Qualitative comparisons.** Facial attribute translation results on CelebA dataset. (a) compares multi-domain translation quality (H: hair color, G: gender, A: aged), and (b) compares multi-modal translation quality. Each result of our model is generated from the statistics of a single cluster.

3.4. Qualitative evaluation

As shown in Fig. 4, we qualitatively compare face attribute translation results on CelebA dataset. All baseline models are trained using the attribute labels, while XploreGAN is trained with unlabeled data. As we increase the number of clusters in k-means clustering, we can discover multiple subsets of a single attribute (e.g., diverse styles of 'women'; further discussed in Section 3.6). This can be thought of as discovering multiple different modes in data. Thus, we can compare our model to not only multi-domain translation but also multi-modal translation between two domains. Fig. 4 demonstrates that our method can generate translation results of as high quality as other models trained with labels. Also, Fig. 3 shows that XploreGAN can perform high-quality translation for various datasets (FFHQ [21], CelebA [27], and Emotionet [8]). We present additional qualitative results in the Appendix.

3.5. Quantitative evaluation

A high-quality image translation should i) properly transfer the target attribute while ii) preserving the identity of the input image and iii) maintain realism to human eyes. We quantitatively measure the three quality metrics by attribute classification, face verification, and a user study.

Attribute classification. To measure how well a model transfers attributes, we compare the classification accuracy of synthesized images on face attributes. We train a binary classifier for each of the selected attributes (blond, brown, old, male, and female) in CelebA dataset (70%/30% split for training and test sets), which results in an average accuracy of 95.8% on real test images. We train all models with the same training set and perform image translation on the same test set. Finally, we measured the classification accuracy of translated images using the trained classifier above. Surprisingly, XploreGAN outperforms all baseline models in almost all attribute translation as shown in Table 1. This shows that our method trained on *unlabeled* data can perform high-quality translation as well as, or sometimes even better than, those models trained on *labeled* data.

Method	Blond	Brown	Aged	Male	Female
Ours	**90.2**	77.4	**90.0**	**99.7**	**99.6**
StarGAN	90.0	**86.1**	88.4	97.5	98.0
MUNIT	-	-	-	95.7	99.1
DRIT	-	-	-	98.8	98.5
Real Image	97.2	92.4	93.3	98.5	97.6

Table 1: Classification accuracy for translated images, evaluated on five CelebA attributes.

Identity preservation. We measure the identity preservation performance of translated images using a state-of-the-art face verification model. We use ArcFace [6] pre-trained on Celeb-1M dataset [12], which shows an average accuracy of 89.76% on the CelebA test set. Next, we perform image translation on the same unseen test set regarding five face attributes (blond hair, brown hair, aged, male, and female). To measure how well a translated image preserves identity of the input image, we measure the face verifica-

Method	Blond	Brown	Aged	Male	Female
Ours	**99.3**	**99.4**	**99.1**	90.1	**94.8**
StarGAN	96.8	99.0	98.8	**97.5**	93.7
MUNIT	-	-	-	9.7	16.3
DRIT	-	-	-	72.2	62.0

Table 2: Facial verification accuracy for identity preservation of translated images from different methods.

Method	Hair	Aged	Gender	H+G	H+A
Ours	**54.7**	43.8	**64.5**	**89.6**	**53.1**
StarGAN	45.3	**56.2**	14.6	10.4	46.9
MUNIT	-	-	4.2	-	-
DRIT	-	-	16.7	-	-

Table 3: User study results. Last two columns correspond to simultaneous translations of multiple domains. (H+G: Hair+Gender, H+A: Hair+Aged)

(a) BagNet-17 (IN) (b) ResNet-50 (SIN)

Figure 5: **Comparing different pre-trained feature spaces.** Different pre-trained feature spaces provide highly different attribute clusters. (a) Texture-based representation: As ImageNet pre-trained BagNets are constrained to capture only small local features, it is effective in detecting texture cues (e.g., skin color, age, hair color, and lighting). (b) Shape-biased representation: ResNets trained on Stylized ImageNet (SIN) are effective in ignoring texture cues and focusing on global shape information (e.g., facial expressions, gestures, and viewpoints). We used BagNet-17 as the feature extractor for CelebA dataset (first four rows) and ResNet-50 pre-trained on SIN for EmotioNet dataset (last four rows).

tion accuracy on pairs of real and fake images using the pre-trained verification model above. As shown in Table 2, our method produces translation results that preserve the identity of an input image comparable to or sometimes even better than most baseline models trained using attribute labels. Although multi-modal image translation models (MUNIT and DRIT) show high classification accuracy by properly transferring target attributes, they often modify the input to the extent that it greatly hinders identity preservation.

User study. To evaluate how realistic the translated outputs are from human eyes, we conduct a user study with 32 participants. Users are asked to choose which output is most successful in producing high-quality images, while preserving content and transferring the target attribute well. 20 questions were given for each of the six attributes, with a total of 120 questions. Note that since MUNIT and DRIT produce multi-modal outputs, a single image was chosen randomly for the user study. Table 3 shows that our model performs as well as supervised models across diverse attributes. Though StarGAN achieves promising results, its results on H+G frequently exhibit green artifacts, which decreases user preference.

3.6. Analysis on the clustering stage

Comparison on pre-trained feature spaces. The pre-trained feature space provides guidance to group the unlabeled images. We found that differences in model architectures and datasets it was pre-trained on lead to significantly different feature spaces, i.e., representation bias. We attempt to exploit this "skewness" towards recognition of particular types of features (e.g., texture or shape) – to group novel images in different directions. We will mainly compare two feature spaces: texture-biased BagNets and shape-biased ResNets. It has been found that ImageNet pre-trained CNNs are strongly biased towards recognizing textures rather than shapes [9]. Related to this characteristic, BagNets [3] are designed to be more sensitive to recognizing local textures compared to vanilla ResNets [14] by limiting the receptive field size. They are designed to focus on small local image features rather than their larger spatial relationships. On the other hand, ResNets trained on Stylized ImageNet [9] (denoted ResNet (SIN)) ignore texture cues altogether and focus on global shapes of images.

Fig. 5 shows the characteristics of these two feature spaces. BagNets trained on ImageNet are effective in detecting detailed texture cues (e.g., skin color and texture, degree of age, hair color/shape, lighting). However, BagNets or vanilla ResNets fail to detect facial emotions, as they often produce clustering results biased towards local texture cues. In fact, we found that ResNet (SIN) is highly effective in ignoring texture cues and focusing on global shape information (e.g., facial expressions, gestures, view-points). Based on this observation, we have adopted BagNet-17 (IN) as the feature extractor for CelebA [27] and FFHQ [21] datasets, and ResNet-50 (SIN) for EmotioNet [8] dataset.

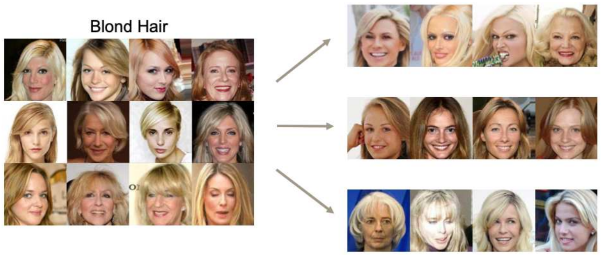

Figure 6: **Effects of increasing the number of clusters.** One can discover hidden attribute subsets previously entangled in a single cluster.

Choosing the number of clusters. As shown in Fig. 6, a single blond cluster is further divided to different types of blond hair as the number of clusters increases. As such, a small number of clusters produces compact clusters with highly distinctive features, while its large number produces clusters with similar yet detailed features. However, it is not clear to determine the optimal number of clusters, but instead it can be subjective depending on how a human labeler defines a single attribute in a given dataset (e.g., 'pale makeup' itself can be a single attribute, or it may be further divided into 'pale skin', 'wearing eyeshadow', and 'wearing lipstick'). In our model, users can flexibly control such granularity by adjusting the number of clusters.

4. Related Work

Generative adversarial networks (GANs). GANs [10] have achieved remarkable success in image generation. Its key to success is the adversarial loss, where the discriminator distinguishes between real and fake images while the generator attempts to fool the discriminator by producing realistic fake images. Existing studies leverage *conditional* GANs in order to generate samples conditioned on the class [28, 30, 31], text description [34, 42, 35], domain information [4, 33], input images [18], and color features [2]. Our approach adopts the adversarial loss conditioned on the cluster statistics to generate corresponding translated images indistinguishable from real images.

Unpaired image-to-image translation. Image-to-image translation [18, 46] has recently shown remarkable progress. CycleGAN [45] extends it to *unpaired* settings. Multi-domain translation models [4, 33] generate diverse outputs when given domain labels. DRIT [23] and MUNIT [17] further advance image translation models to produce diverse multi-modal outputs using unpaired data. FUNIT [26] is trained on labeled images and performs translation based on few images of a novel object class at test time. As such, existing methods mostly rely on labeled data. Unlike previous approaches that define the term 'unpaired' as synonymous to unsupervised, we define unsupervised to encompass both unpaired and unlabeled. According to our definition, no previous work on image-to-image translation has tackled this setting.

Clustering for discovering the unknown. Clustering is a powerful unsupervised learning method that groups data by their similarity. It has been used to discover novel object classes in images [25] and videos [32, 37, 15, 40]. Instead of discovering new object classes, our work aims to discover attributes within unlabeled data through clustering. Finding attributes is a complicated task, as a single image can have multiple different attributes. To the best of our knowledge, our work is the first to perform image-to-image translation using newly discovered attributes from unlabeled data.

Instance normalization for style transfer. To facilitate training of neural networks, batch normalization (BN) was originally introduced. BN normalizes each feature channel by its respective mean and standard deviation from mini-batches of images. Instance normalization [38] utilizes the mean and standard deviation from a given image. As its extension, conditional instance normalization [7] learns different sets of parameters for each style. Adaptive instance normalization (AdaIN) [16] performs normalization without additional trainable parameters, to which MUNIT adds trainable parameters for flexible translation cability. In contrast to existing normalization methods that perform style transfer on image instances, our attribute summary instance normalization (ASIN) uses cluster statistics to summarize the common attribute within each cluster, which allows translation of fine, detailed attributes.

5. Conclusions

In this paper, we attempt to alleviate the necessity for labeled data in the facial image translation domain. Provided with raw, unlabeled data, we propose an *unpaired* and *unlabeled* multi-domain image-to-image translation method. We utilize prior knowledge from pre-trained feature spaces to group unseen, unlabeled images. Attribute summary instance normalization (ASIN) can effectively summarize common attributes within clusters, enabling high-quality translation of particular attributes. We demonstrate that the results of our model is comparable to or sometimes better than most of the state-of-the-art methods.

Acknowledgments This work was partially supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-00075, Artificial Intelligence Graduate School Program (KAIST)), Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (2019R1A2C4070420), and Korea Electric Power Corporation (Grant number:R18XA05).

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017. 3

[2] Hyojin Bahng, Seungjoo Yoo, Wonwoong Cho, David Keetae Park, Ziming Wu, Xiaojuan Ma, and Jaegul Choo. Coloring with words: Guiding image colorization through text-based palette generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 431–447, 2018. 8

[3] Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. In *International Conference on Learning Representations (ICLR)*, 2018. 4, 7

[4] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2, 3, 4, 8

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. Ieee, 2009. 2, 4

[6] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 6

[7] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. 2017. 2, 8

[8] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5562–5570, 2016. 4, 6, 7

[9] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations (ICLR)*, 2019. 4, 7

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2672–2680, 2014. 8

[11] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5767–5777, 2017. 3

[12] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. MS-Celeb-1M: A dataset and benchmark for large scale face recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 6

[13] Kai Han, Andrea Vedaldi, and Andrew Zisserman. Learning to discover novel visual categories via deep transfer clustering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 8401–8409, 2019. 2

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 4, 7

[15] Evan Herbst, Xiaofeng Ren, and Dieter Fox. Rgb-d object discovery via multi-scene analysis. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2011. 8

[16] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1501–1510, 2017. 2, 8

[17] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2, 4, 8

[18] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1125–1134, 2017. 4, 8

[19] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 694–711. Springer, 2016. 2

[20] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*, 2017. 4

[21] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 4, 6, 7

[22] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017. 3

[23] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Kumar Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2, 8

[24] Chuan Li and Michael Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 702–716. Springer, 2016. 4

[25] Liangchen Liu, Feiping Nie, Teng Zhang, Arnold Wiliem, and Brian C Lovell. Unsupervised automatic attribute discovery method via multi-graph clustering. In *International Conference on Pattern Recognition (ICPR)*, pages 1713–1718. IEEE, 2016. 8

[26] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsueprvised image-to-image translation. In *Proceedings of*

International Conference on Computer Vision (ICCV), 2019. 8

[27] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015. 2, 4, 6, 7

[28] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 8

[29] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2018. 4

[30] Augustus Odena. Semi-supervised learning with generative adversarial networks. *arXiv preprint arXiv:1606.01583*, 2016. 8

[31] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017. 8

[32] Aljosa Osep, Paul Voigtlaender, Jonathon Luiten, Stefan Breuers, and Bastian Leibe. Large-scale object mining for object discovery from unlabeled video. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2019. 8

[33] A. Pumarola, A. Agudo, A.M. Martinez, A. Sanfeliu, and F. Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 3, 8

[34] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text-to-image synthesis. In *Proceedings of The 33rd International Conference on Machine Learning (ICML)*, 2016. 8

[35] Qiuyuan Huang Han Zhang Zhe Gan Xiaolei Huang Xiaodong He Tao Xu, Pengchuan Zhang. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 8

[36] Antonio Torralba, Alexei A Efros, et al. Unbiased look at dataset bias. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Citeseer, 2011. 2

[37] Rudolph Triebel, Jiwon Shin, and Roland Siegwart. Segmentation and unsupervised part-based discovery of repetitive objects. *Robotics: Science and Systems VI*, pages 1–8, 2010. 8

[38] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 8

[39] Sheng-Yu Wang, Oliver Wang, Andrew Owens, Richard Zhang, and Alexei A Efros. Detecting photoshopped faces by scripting photoshop. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2019. 2

[40] Christopher Xie, Yu Xiang, Dieter Fox, and Zaid Harchaoui. Object discovery in videos as foreground motion clustering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 8

[41] Hongyu Yang, Di Huang, Yunhong Wang, and Anil K Jain. Learning face age progression: A pyramid architecture of gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 31–39, 2018. 3

[42] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 8

[43] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018. 2

[44] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of International Conference on Computer Vision (ICCV)*, pages 2223–2232, 2017. 4

[45] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networkss. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 2, 3, 8

[46] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 465–476, 2017. 8