Adversarial Learning of Disentangled and Generalizable Representations for Visual Attributes

James Oldfield
Goldsmiths, UoL
j.oldfield@gold.ac.uk

Yannis Panagakis
Imperial College London
i.panagakis@imperial.ac.uk

Mihalis A. Nicolaou
The Cyprus Institute
m.nicolaou@cyi.ac.cy

Abstract

Recently, a multitude of methods for image-to-image translation has demonstrated impressive results on problems such as multi-domain or multi-attribute transfer. The vast majority of such works leverages the strengths of adversarial learning in tandem with deep convolutional autoencoders to achieve realistic results by well-capturing the target data distribution. Nevertheless, the most prominent representatives of this class of methods do not facilitate semantic structure in the latent space, and usually rely on domain labels for test-time transfer. This leads to rigid models that are unable to capture the variance of each domain label. In this light, we propose a novel adversarial learning method that (i) facilitates latent structure by disentangling sources of variation based on a novel cost function and (ii) encourages learning generalizable, continuous and transferable latent codes that can be utilized for tasks such as unpaired multi-domain image transfer and synthesis, without requiring labelled test data. The resulting representations can be combined in arbitrary ways to generate novel hybrid imagery, as for example generating mixtures of identities. We demonstrate the merits of the proposed method by a set of qualitative and quantitative experiments on popular databases, where our method clearly outperforms other, state-of-the-art methods. Code for reproducing our results can be found at: https://github.com/james-oldfield/adv-attribute-disentanglement

1. Introduction

Image-to-image translation methods learn a non-linear mapping of an image in a source domain to its corresponding image in a target domain. The notion of domain varies, depending on the application. For instance, in the context of super-resolution the source domain consists of low-resolution images while the corresponding high-resolution images belong to the target domain. In a visual attribute transfer setting, 'domain' denotes face images with the same attribute that describes either intrinsic facial characteristics (e.g., identity, facial expressions, age, gender, etc.) or capture external sources of appearance variation related, for example, to different poses or illumination conditions. In the latter setting, the task is to change the attributes for a given face image.

Recently, deep generative models trained through adversarial learning have been shown capable of generating naturalistic images, that look authentic to the human observer. Deep generative models for image-to-image implement a mapping between between two [29] or multiple [3] image domains, in a paired [11] or unpaired [29, 3, 19, 14, 10] fashion. Despite their merits in pushing forward the state of the art in image generation, we posit that widely adopted image-to-image translation models (namely CycleGAN [29], Pix2Pix [11] and StarGAN [3]) also come with a set of shortcomings. For example, none of these methods provide semantically meaningful latent structure linked to specific attributes. In addition, the generated images do not cover the entire variance of the target domain and in most cases a single image is generated given a discrete attribute value, which is also required at test time, and thus can not generate images when the attribute label has not been observed during training. For example, changing the "smile" attribute of a facial image will always lead to a smile of specific intensity (Fig. 1)

In this paper, we propose a method that facilitates learning disentangled, generalizable, and continuous representations of visual data with respect to attributes acting as sources of variation. The proposed method can readily be used to generate varying intensity expressions in images (Fig. 1), while also being equipped with several other features that enable generating novel hybrid imagery on unseen data. Key contributions of this work are summarized below.

- Firstly, a novel loss function for learning disentangled representations is proposed. The loss function ensures that latent representations corresponding to an attribute (a) have discriminative power within the attribute class, while (b) being invariant to other sources linked

Figure 1: Expression transfers with the same expression label: the proposed method excels at preserving the variance within attribute values (zoom for quality).

to other attributes. For example, given a facial image with a specific identity and expression, representations of the *identity* attribute should classify all identities well, but should fail to classify the expressions well: that is, the conditional distribution of the class posteriors should be uniform over the expression labels.

- Secondly, we propose a novel method that encourages the disentangled representations to be *generalizable*. The particular loss function enables the network to generate realistic images that are classified appropriately, even when sampling representations from different samples, and without requiring paired data. This enables the representations to well-capture the attribute variation, as shown in Fig. 1, in contrast to other methods that simply consider a target label for transfer. We highlight that the expected value of the latent codes over a single attribute an be considered in our case as equivalent to an attribute label.

- Finally, we provide a set of rigorous experiments to demonstrate the capabilities of the proposed method on databases such as MultiPIE, BU-3DFE, and RaFD. Given generalizable and disentangled representations on a multitude of attributes (e.g., expression, identity, illumination, gaze, color), the proposed method can perform arbitrary combinations of the latent codes in order to generate novel imagery. For example, we can swap an arbitrary number of attribute representations amongst test samples to perform intensity-preserving multiple attribute transfer and synthesis, without knowing the test labels. Most interestingly, we can combine an arbitrary number of e.g., *identity* latent codes, in order to generate novel subjects that preserve a mixture of characteristics, and can be particularly useful for tasks such as data augmentation. Both qualitative and quantitative results corroborate the improved performance of the proposed approach over state-of-the-art image-to-image translation methods.

2. Related Work

Generative Adversarial Networks Generative Adversarial Networks [6] (GANs) approach deep generative models training from a game theory perspective by solving a minimax game. That is, GANs learn a distribution that matches the real data distribution and generate new image by sampling from the estimated distribution. Such an adversarial learning approach has been successfully employed in a wide range of computer vision tasks, e.g., [18, 25, 13]. Conditional variants of GANs [22], condition the generator on a particular class label. This allows fine-grained control over the generator in targeting particular modes of interest, facilitating predictable manipulation and generation of imagery [11, 26, 1, 23].

In a vanilla GAN paradigm however, there is no way to impose a particular semantic structure on the random noise vector from the prior distribution. Consequently, it is hard to drive desired change in generated images via latent space manipulation. Such limitations have been mitigated in [4] by imposing structure on the noise prior and also in the context of Variational Autoencoders (VAE) [21, 4, 15]. Similarly, InfoGAN [2] learn disentangled representations in a completely unsupervised manner. Nevertheless, the aforementioned deep generative models tend to yield blurry results, and having very low-dimensional bottlenecks often means the resolution of the generated imagery is compromised [5]. As opposed to VAE-based models, the proposed method is able to synthesize sharp, realistic images.

Adversarial Image-to-Image Translation Several GANs-based image-to-image translation models have achieved great success in both a paired and unpaired image-to-image translation setting [11, 29] by combining traditional reconstruction losses (e.g. L_1 reconstruction penalties) with adversarial terms to enhance the visual clarity of the model's outputs. In CycleGAN [29], DiscoGAN [14] and DualGAN [27] the so-called 'cycle-consistency' loss facilitates unpaired image-to-image domain translation. Coupled GAN [20] and its extension [19] assume a shared-latent space to learn a joint distribution of different, unpaired, domains. Recently, StarGAN [3] enables image-to-image translation across multiple domains with the use of a single conditional generative model. The latter can flexibly translate between multiple target domains with the generator being a function of both the input data and target domain label. FaderNet [16] translates input images to new ones

by manipulating attributes values via incorporating the discriminator onto the latent space.

3. Methodology

In this section, we provide a detailed description of the methodology proposed in this work, which focuses on learning disentangled and generalizable latent representations of visual attributes. Concretely, in Section 3.1 we describe the generative model employed in this work. In Section 3.2, we introduce the proposed loss functions and method towards disentangling these representations in latent space, such that they (i) well-capture variations that relate to a given attribute by enriching features with discriminative power (e.g., for identity), and (ii) *fail* to classify any other attributes well, by encouraging the classifier posterior distribution over values of other attributes (e.g., expression) to be uniform. In Section 3.3, we describe the optimization procedure that is tailored towards encouraging recovered representations to be *generalizable*, that is, can be utilized towards generating novel, realistic images from arbitrary samples and attributes. The full objective function employed is described in Section 3.4, while finally, implementation details regarding the full network that is trained in an end-to-end fashion are provided in Section 3.5. An overview of the proposed method is illustrated in Fig. 2

3.1. Generative Model

Given dataset with N samples, we assume that each sample $\mathbf{x}^{(i)}$ is associated with a set of M attributes that act as sources of visual variation (such as identity, expression, illumination). If not omitted, the superscript i denotes that $\mathbf{x}^{(i)}$ is the i-th sample in the dataset or mini-batch. We further assume a set of labels $y_m(\mathbf{x})$ corresponding to each attribute m. We aim to recover disentangled, latent representations $\mathbf{z}_m$ that capture variation relating to attribute m, while being invariant to variations sourced from the remaining $M-1$ attributes. We therefore assume the following generative model

$$\tilde{\mathbf{x}} = \mathcal{D}(\mathcal{E}(\mathbf{x})) = \mathcal{D}(\mathbf{z}) = \mathcal{D}(\mathbf{z}_0, \dots, \mathbf{z}_M) \tag{1}$$

where $\mathbf{z}_m = \mathcal{E}_m(\mathbf{x})$ is an encoder mapping to a space that preserves variance for attribute m, while $\mathcal{D}$ is a decoder mapping back to input space. Note that when $m = 0$, the corresponding representations $\mathbf{z}_0$ represent variation present in an image that does not relate to any attribute. For example, assuming a dataset of facial images, if $M = 2$ with $\mathbf{z}_1$ corresponding to *identity* and $\mathbf{z}_2$ to *expression*, $\mathbf{z}_0$ captures other variations such as e.g., background information. To ensure that the set of $M+1$ representations faithfully reconstruct the original image $\mathbf{x}$, we impose a standard reconstruction loss,

$$\mathcal{L}_{rec} = \mathbb{E}_{\mathbf{x}}\Big[||\mathbf{x} - \mathcal{D}(\mathcal{E}(\mathbf{x}))||_1\Big]. \tag{2}$$

3.2. Learning Disentangled Representations

Our aim is to define a transformation that is able to generate disentangled representations for specific attributes that act as sources of variation, while being invariant with respect to other variations present in our data. To this end, we introduce a method for training the encoders $\mathcal{E}_m$ arising in our generative model (Eq. 1), where resulting representations $\mathbf{z}_m$ have discriminative power over attribute m, while yielding maximum entropy class posteriors for each of the other $M-1$ attributes. In effect, this prevents any contents relating to other attributes besides m from arising in the resulting representations. To tackle this problem, we propose a composite loss function as discussed below.

Classification Loss. We firstly employ a loss function that is reminiscent of a typical classification loss, to ensure that the obtained representations $\mathbf{z}_m = \mathcal{E}(\mathbf{x})$ well-classify variation that is related to attribute m. This is done by feeding the representations directly into the the fully connected layers of a classifier $\mathcal{C}_m$, and minimizing the negative log-likelihood of the ground truth labels given an input sample,

$$\mathcal{L}_{cls}^{\mathbf{x}} = \mathbb{E}_{\mathbf{x}}\Bigg[\frac{1}{M}\sum_{m=1}^{M}\Big(-\log \mathcal{C}_m\Big(y_m(\mathbf{x})|\mathcal{E}_m(\mathbf{x})\Big)\Big)\Bigg]. \tag{3}$$

Disentanglement Loss. Classification losses, as defined above, ensure that the learned transformations leads to representations that are enriched with information related with the particular ground-truth label, and have been employed in different forms in other works such as [3]. However, as we demonstrate experimentally in Section 4.3, it is not reasonable to expect that the classification loss alone is sufficient to disentangle the latent representations $\mathbf{z}_m$ from other sources of variations arising from the other $M-1$ attributes. Hence, to further encourage disentanglement, we impose an additional "disentanglement" loss on the conditional label distributions of the classifiers, given the corresponding representations. In more detail, we posit that the class posterior for each attribute m given the latent represenations $\mathbf{z}_m$ induced by the encoder $\mathcal{E}_m(\mathbf{x})$ for every other distinct attribute m' should be a uniform distribution. We impose this soft-constraint by minimizing the cross-entropy loss between a uniform distribution and the classifier class posteriors, that is

$$\mathcal{L}_{dis}^{m} = \mathbb{E}_{\mathbf{x}}\Bigg[\frac{1}{(M-1)}\sum_{\substack{m'=1 \\ m'\neq m}}^{M}\frac{1}{|m'|}\log \mathcal{C}_{m'}\Big(y_{m'}|\mathcal{E}_m(\mathbf{x})\Big)\Bigg], \tag{4}$$

where m' iterates over all other attributes and $|m|$ is the number of classes for attribute m. In other words, we

impose that each encoder $\mathcal{E}_m$ must map to a representation that is correctly classified with respect to *only* the relevant attribute m, and that the representation is such that the conditional label distribution given its mapping, for every other attribute, has maximum entropy. In other words, this loss function filters-out information related to variation arising from other attributes. We note that the final loss function averages over all attributes, that is $\mathcal{L}_{dis} = \frac{1}{M+1}\sum_{m=0}^{M}\mathcal{L}_{dis}^{m}$. In particular, for $m = 0$ we ensure that the representations obtained via $\mathbf{z}_0 = \mathcal{E}_0(\mathbf{x})$ are invariant to *all* M attribute variations, while capturing only variations that are not related to any of the M attributes. This is an important distinction, as we can not always assume that all image variation is related to the given attributes.

3.3. Learning Generalisable Representations

The loss function described in Section 3.2 encourages the representations $\mathbf{z}_m$ generated by the corresponding encoders to capture the variation induced by attribute m, while being invariant to variation arising from other sources. In this section, we provide a simple, effective method for ensuring that the derived representations are *generalizable* over unseen data (e.g. new identities in facial images), while at the same time yielding the expected semantics in the generated images. We which utilizes the classifiers' distributions to learn generalizable representations *without* requiring the ground-truth pair for any combination of labels.

We assume a mini-batch of size b. During each forward pass, we randomly shuffle the representations for each attribute along the batch dimension, which when passed through the decoder provides a new synthesized sample, $\tilde{\mathbf{x}}'$,

$$\tilde{\mathbf{x}}' = \mathcal{D}(\mathbf{z}') = \mathcal{D}(\mathbf{z}_0, \mathbf{z}_1^{(r_1)}, \ldots, \mathbf{z}_M^{(r_M)}) \tag{5}$$

where $r_1, \ldots, r_M \in [1, b]$ are random integers indexing the mini-batch data. In essence, this leads to a synthesized sample $\tilde{\mathbf{x}}'$. Since we know what value the ground-truth labels for the attributes should be taking in the synthesized sample $\tilde{\mathbf{x}}'$, we can enforce a classification loss on $\tilde{\mathbf{x}}'$ by minimizing the negative log-likelihood of the expected classes for each attribute,

$$\mathcal{L}_{cls}^{\tilde{\mathbf{x}}} = \mathbb{E}_{\tilde{\mathbf{x}}'}[\frac{1}{M}\sum_{m=1}^{M} -\log \mathcal{C}_m\left(y_m(\tilde{\mathbf{x}}')|\tilde{\mathbf{x}}'\right)]. \tag{6}$$

We highlight that the above loss is at an advantage over *paired* methods (such as [11]) in that we don't require direct access to the corresponding target $\mathbf{x}'$.

Adversarial Loss In order to induce adversarial learning in the proposed model, and encourage generated images to match the data distribution, we further impose an adversarial loss in tandem with Eq. 6,

$$\mathcal{L}_{adv} = \mathbb{E}_{\mathbf{x}}[\log D(\mathbf{x})] + \mathbb{E}_{\tilde{\mathbf{x}}'}[\log(1 - D(\tilde{\mathbf{x}}'))] \tag{7}$$

This ensures that even when representations $\mathbf{z}_m$ are shuffled accross data points, the synthesized sample will both (i) be classified according to the sample/embedding combination (Eq. 7), as well as (ii) constitute a realistic image.

3.4. Full Objective

The proposed method is trained end-to-end, using the full objective as grouped by the set of variables we are optimizing for

$$\begin{aligned}\mathcal{L}_G &= -\mathcal{L}_{adv} + \mathcal{L}_{dis} + \mathcal{L}_{cls}^{\mathbf{x}} + \mathcal{L}_{cls}^{\tilde{\mathbf{x}}} + \mathcal{L}_{rec} \\ \mathcal{L}_D &= \mathcal{L}_{adv}, \quad \mathcal{L}_C = \mathcal{L}_{cls}^{c},\end{aligned} \tag{8}$$

where $\mathcal{L}_D$ is the combined loss for the discriminator, $\mathcal{L}_C$ is for the classifiers, and $\mathcal{L}_G$ is for the encoders and decoder. We control the relative importance of each loss term $\mathcal{L}_i$ with a corresponding λ_i hyperparameter.

3.5. Implementation

At train-time we sample mini-batches of size $b = 16$ at random. Each iteration we shuffle the M attributes' encodings along the batch dimension before concatenating depth-wise, and feeding into the decoder (i.e. Eq. (5)), to train the network to be able to flexibly pair any combination of values of the attributes. **Network Architecture.** We define $M + 1$ encoder instances (one for each explicitly modeled attribute, and an additional encoder to capture the remaining sources of variation). Each encoder $\mathcal{E}_m$ is a separate convolutional encoder based on the first half of [12] up to the bottleneck. Our decoder $\mathcal{D}$ depth-concatenates all $M+1$ latent encodings and then upsamples via [24] to reconstruct the input image. We adopt the deeper PatchGAN variant proposed in [3] for our discriminator. The classifiers $\mathcal{C}_m$ are simple shallow CNNs–trained on the images in the training set to correctly classify the labels of attribute m–with a final dense layer that outputs the logits for the classes of the appropriate attribute m.

4. Experiments

In this section, we present a set of rigorous qualitative and quantitative experiments on multiple datasets to validate the proposed method, and verify the derived representations are disentangled, generalizable, and continuous. In more detail, we experiment on databases such as Multi-PIE, BU-3DFE, and RaFD. We utilize the proposed method to learn disentangled and generalizable representations on various categorical attributes (), including *identity*, *expression*, *illumination*, *gaze*, and *color*. In more detail, we explore the properties of the proposed model in Section 4.3. In Section 4.4, we detail experiments related to expression synthesis in comparison to SOTA image-to-image translation models on the test set of each database. Subsequently, in Section 4.5 results on arbitrary multi-attribute transfer

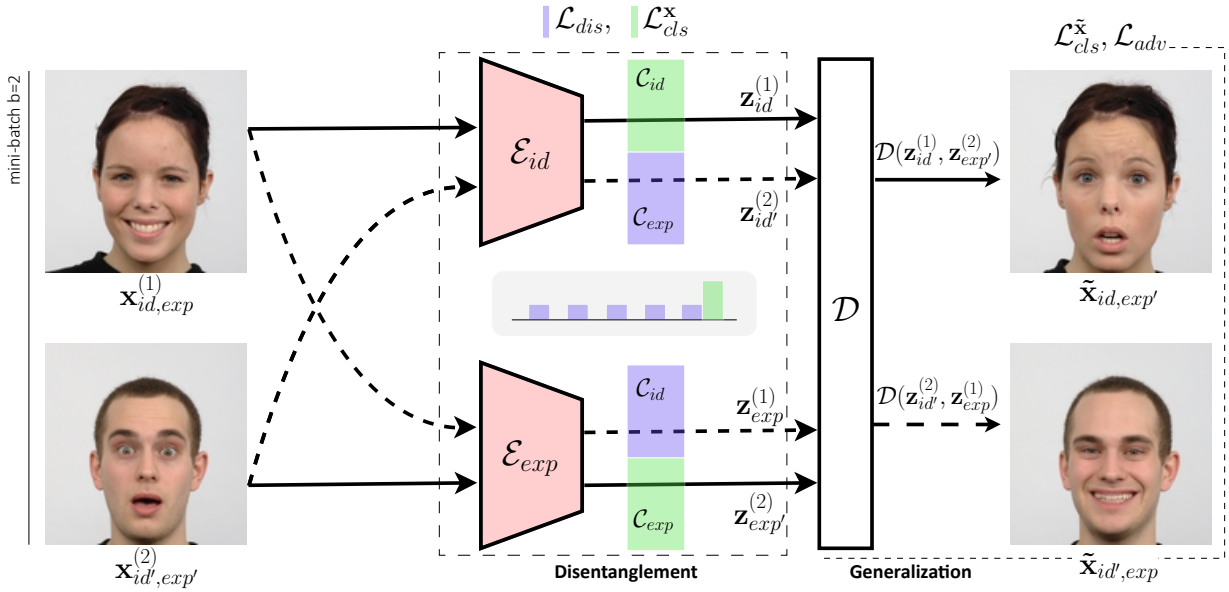

Figure 2: Overview of the proposed method, on a simple example with two samples, and attributes *identity* and *expression*. Samples are mapped to distinct disentangled representations $\mathbf{z}_m^{(i)}$ by utilizing the classification $\mathcal{L}_{cls}$ and disentanglement loss $\mathcal{L}_{dis}^{\mathbf{x}}$, encouraging the representations to well-classify identity (expression) and the classifier posteriors to be uniform over the expression (identity) labels. To enable the decoder $\mathcal{D}$ to generate novel images at test time, generalization is encouraged by permuting the latent representations for each attribute sample-wise. Utilizing a classification $\mathcal{L}_{cls}^{\tilde{\mathbf{x}}}$ and adversarial loss $\mathcal{L}_{adv}$, the synthesized images $\tilde{\mathbf{x}}$ are encouraged to be both realistic, as well as be classified correctly according to the given attribute. Our method does not require the synthesized images to exist in the dataset. Note: For clarity of presentation, we omit $\mathcal{E}_0$, the encoder handling variations unrelated to attributes M.

are presented, where as can be clearly seen the proposed representations well-capture attribute variation. Finally, in Section 4.6, we further evince the generalizable nature of the latent representations for each attribute. By performing weighted combinations of the represenations over multiple samples, we can generate novel unseen data, as for example novel identities. In the same section, we also perform latent space interpolations to demonstrate the continuous properties of the representations. Finally, we note that for all experiments, we adopt the Wasserstein-GP GAN objective [8] and set $\lambda_{rec} := 10$.

4.1. Datasets

MultiPIE. The MultiPIE [7] dataset consists of over 750,000 images, including a challenging range of variation. We use the forward-facing subset of MultiPIE, jointly modelling attributes 'identity', 'expression' and 'illumination'. We use 686 images for each of the emotions 'neutral', 'scream', 'squint', and 'surprise' for our training set, and holdout the first 10 identities for the test set. We utilise the same dataset splits across all models to ensure a fair comparison. **BU-3DFE (BU).** The BU [28] dataset is comprised of 100 identities, and a wide range of age, gender, race, and expression intensities. We utilize the 2D frontal projections for our training set, reserving 10 identities for the test set. **RaFD.** The RaFD [17] dataset contains 67 individuals, each in 8 expressions, each with 3 gazes, and in 3 poses. We holdout the first 8 images for the test set, and use the remaining for training. We use only the front-facing poses.

4.2. Baselines

StarGAN. We consider StarGAN [3] to be a state-of-the-art method for unpaired image-to-image translation between multiple domains, and consequently benchmark our model's performance against it. **CycleGAN.** To tackle unpaired image-to-image translation between two domains, CycleGAN [29] employ a cycle constraint to ensure the generator's inverse function approximates an identity function. We benchmark against CycleGAN by training it pairwise between each expression and 'neutral'. **Pix2Pix.** Pix2Pix [11] utilises a conditional GAN for learning the mapping between *paired* tuples from two domains.

4.3. Model Exploration

In this section, we present a set of exploratory experiments that verify the properties of the proposed model. In more detail, we train our model on the MultiPIE database, and plot the conditional PMFs of classifiers for expression

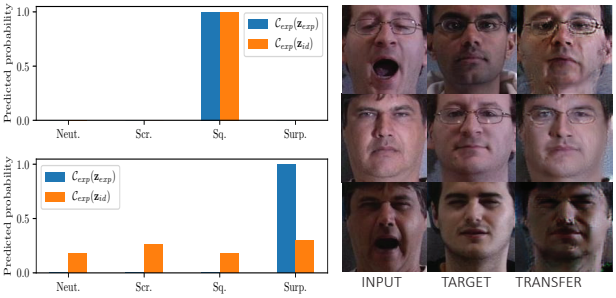

Figure 3: Left: Predicted classes (top - without disentanglement, bottom - disentanglement loss). Right: Ablation study showing that without disentanglement loss, identity content is transferred in the results.

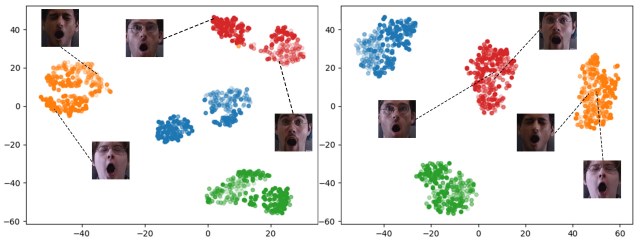

Figure 4: t-SNE clustering on expression representations with (right) and without (left) the disentanglement loss. Color indicates expression, while transparency to changes in illumination (zoom for better quality).

and identity in Fig. 3. As can be seen, the latent representations bear discriminative power for the desired attributes, whilst producing uniformly distributed class assignments for other attributes. This is not the case when we do not include the proposed loss, where variance from other attributes is contained. This is further evidenced by an ablation study, showing that without the disentanglement losses, identity components and ghosting artefacts from clothing are prone to mistakenly fall into the expression representations. Finally, in Fig. 4 we train our model on the MultiPIE database and visualize the recovered embeddings $\mathbf{z}_{exp}$ with and without the disentanglement loss. As can be seen, without the disentanglement loss, clusters with the same expression are split according to illumination, while our representations appear invariant to such variations and find the correct expression clusters.

4.4. Expression Synthesis

In this section we present a set of expression synthesis results across several databases. In more detail, most GAN-based methods are able to synthesize facial images given a specific target expression (e.g., "neutral" to "smile"). Our method is able to capture the variability associated with each attribute (expression in this case), and we can therefore generate varying intensity images of the same expression. We can also obtain a representation equivalent to an expression label as used in other models (such as [3]) by simply taking the expected value of the embeddings, $\mathbb{E}_{i=1}^{N}[\mathbf{z}_m^{(i)}]$, while we can also generate combinations of embeddings to obtain novel data. We compare our method against SOTA image-to-image translation models, applied on the test sets of BU, MultiPIE and RaFD in Fig. 5 (a), (b), and (c) respectively. As can be seen, the proposed method can generate sharp, realistic images of target expressions, and can capture expression intensity as can be particularly seen in Fig. 5 (a). Quantitative results on classification accuracy and FID are show in in Table 1. We note that since CycleGAN and pix2pix can only be trained between pairs of domains, a separate instance is trained for each pair of expressions. Our method outperforms compared techniques on nearly all databases and metrics.

	MultiPIE			BU			RaFD		
Model	**mean**	**std**	**FID**	**mean**	**std**	**FID**	**mean**	**std**	**FID**
pix2pix	0.99	±0.01	121.86	**0.89**	**±0.02**	58.99	**0.99**	**±0.01**	91.57
CycleGAN	0.96	±0.01	97.12	0.63	±0.02	65.55	0.93	±0.01	71.44
StarGAN	0.99	±0.01	**75.39**	0.74	±0.03	59.58	0.95	±0.02	55.99
Ours	**1.0**	**±0.00**	90.95	0.77	±0.02	**46.62**	0.94	±0.01	**52.44**

Table 1: Classification Accuracy and FID[9] for expression syntheses on the test set. We note that whilst pix2pix performs well under classification, it performs poorly under the FID metric, and the samples are noticeably less sharp.

4.5. Attribute Transfer

In this section, we present experiments that involve arbitrary, intensity-preserving transfer of attributes. While most other methods require a target domain or a label, in our case we can simply swap the obtained representations arbitrarily from sample to sample, while also being able to perform operations on them. In more detail, in Fig. 6 we demonstrate that the proposed method is able to map to disentangled and generalizable attributes, by successfully transferring intrinsic facial attributes such as identity, expression, and gaze, as well as appearance based attributes such as illumination and image color. Note that this demonstrates the generality of our method, handling arbitrary sources of variation. Finally, we also show that it is entirely possible to transfer several attributes jointly, by using the corresponding representations. We show examples where we simultaneously transfer expression and illumination, as well as expression and gaze.

4.6. Generating Novel Identities and Interpolation

Our method produces generalizable representations that can be combined in many ways to synthesize novel samples. This is an important feature that can be used towards tasks such as data augmentation, as well as enhancing the robustness of classifiers for face recognition. In this section, we

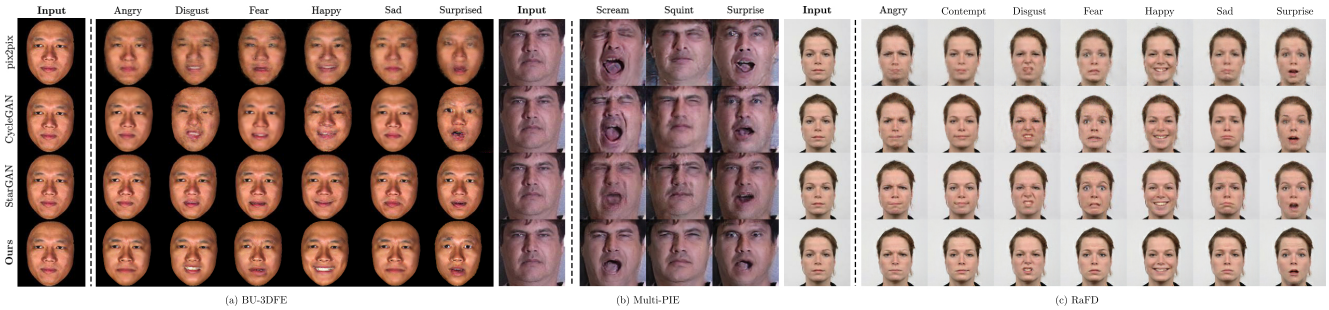

Figure 5: Expression synthesis comparison against baseline methods, on the test set of all datasets.

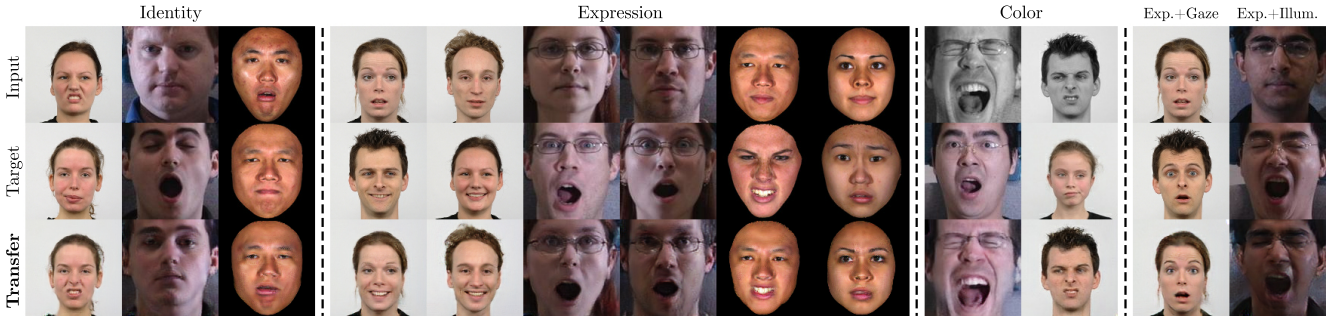

Figure 6: Multiple-attribute transfer across databases and attributes. **Row 1**: Input image; **Row 2**: Target image; **Row 3**: Single or Joint attribute transfer (zoom in for better quality).

Figure 7: Combining multiple identity representations.

present experiments to demonstrate that the learned representations are both generalizable and continuous in the latent domain for all attributes. Consequently, latent contents can be manipulated accordingly in order to synthesize entirely novel imagery. Compared to dedicated methods for interpolation employing GANs such as [16], our method immediately allows for categorical attributes and also preserves intensity when interpolating between specific values. In more detail, in Fig. 7 we decode the convex combination of 3 identity embeddings from distinct samples, and synthesize a realistic mixture of the given identities rendered as a new person. Finally, in Fig. 8, we show that we can smoothly interpolate the representations for attributes such as *identity*, *illumination*, and *expression*.

5. Conclusion

In this paper, we presented a novel method for learning disentangled and generalizable representations in an adversarial manner. The proposed method offers many benefits over other methods, while learning a meaningful latent structure that corresponds semantically to image characteristics. We provided experimental evidence to showcase some of the possibilities that arise with such a rich latent structure, including being able to interchange and manipulate latent contents to generate a large, rich gamut of new hybrid imagery.

References

[1] A. Brock, T. Lim, J. M. Ritchie, and N. Weston. Neural photo editing with introspective adversarial networks, 2016. 2

[2] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets, 2016. 2

[3] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation, 2017. 1, 2, 3, 4, 5, 6

[4] C. Donahue, Z. C. Lipton, A. Balsubramani, and J. McAuley. Semantically decomposing the latent spaces of generative adversarial networks, 2017. 2

[5] A. Dosovitskiy and T. Brox. Generating images with perceptual similarity metrics based on deep networks. In *Ad-*

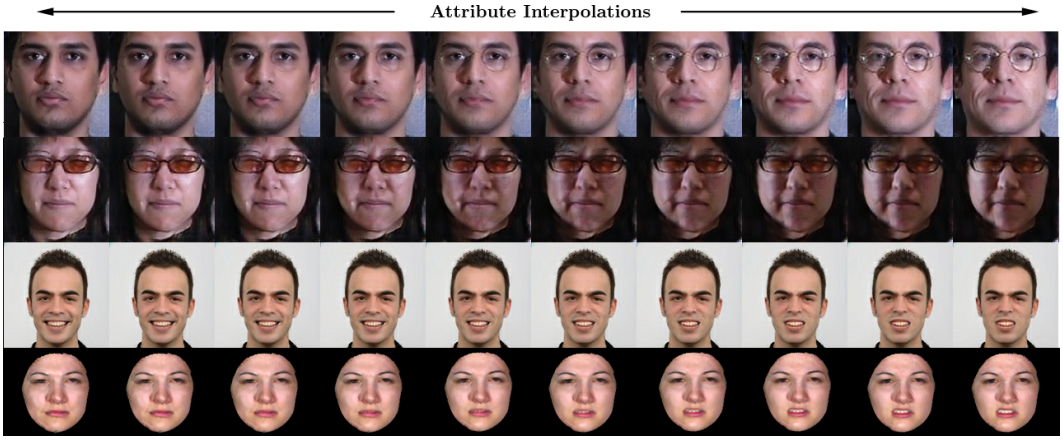

Figure 8: Linearly interpolating between various categorical target attributes.

vances in Neural Information Processing Systems (NIPS), pages 658–666, 2016. 2

[6] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks, 2014. 2

[7] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE Computer Society, September 2008. 5

[8] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville. Improved training of wasserstein gans, 2017. 5

[9] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. 2017. 6

[10] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz. Multimodal unsupervised image-to-image translation, 2018. 1

[11] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. 2016. 1, 2, 4, 5

[12] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution, 2016. 4

[13] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation, 2017. 2

[14] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim. Learning to discover cross-domain relations with generative adversarial networks, 2017. 1, 2

[15] T. D. Kulkarni, W. Whitney, P. Kohli, and J. B. Tenenbaum. Deep convolutional inverse graphics network, 2015. 2

[16] G. Lample, N. Zeghidour, N. Usunier, A. Bordes, L. DENOYER, et al. Fader networks: Manipulating images by sliding attributes. In *Advances in Neural Information Processing Systems*, pages 5963–5972, 2017. 2, 6

[17] O. Langner, R. Dotsch, G. Bijlstra, D. H. J. Wigboldus, S. T. Hawk, and A. van Knippenberg. Presentation and validation of the radboud faces database. *Cognition & Emotion*, 24(8):1377–1388, dec 2010. 5

[18] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-realistic single image super-resolution using a generative adversarial network, 2016. 2

[19] M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks, 2017. 1, 2

[20] M.-Y. Liu and O. Tuzel. Coupled generative adversarial networks, 2016. 2

[21] M. Mathieu, J. Zhao, P. Sprechmann, A. Ramesh, and Y. LeCun. Disentangling factors of variation in deep representations using adversarial training, 2016. 2

[22] M. Mirza and S. Osindero. Conditional generative adversarial nets, 2014. 2

[23] V. Nguyen, T. F. Y. Vicente, M. Zhao, M. Hoai, and D. Samaras. Shadow detection with conditional generative adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, oct 2017. 2

[24] A. Odena, V. Dumoulin, and C. Olah. Deconvolution and checkerboard artifacts. *Distill*, 2016. 4

[25] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. 2016. 2

[26] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans, 2017. 2

[27] Z. Yi, H. Zhang, P. Tan, and M. Gong. Dualgan: Unsupervised dual learning for image-to-image translation, 2017. 2

[28] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato. A 3d facial expression database for facial behavior research. In *Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition*, FGR '06, pages 211–216, Washington, DC, USA, 2006. IEEE Computer Society. 5

[29] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks, 2017. 1, 2, 5