Exploring Disentangled Feature Representation Beyond Face Identification

Yu Liu[1]*, Fangyin Wei[2]*, Jing Shao[2]*, Lu Sheng[1], Junjie Yan[2], Xiaogang Wang[1]
[1]CUHK-SenseTime Joint Lab, The Chinese University of Hong Kong
[2]SenseTime Group Limited
{yuliu,lsheng,xgwang}@ee.cuhk.edu.hk, weifangyin@pku.edu.cn,
{shaojing,yanjunjie}@sensetime.com

Abstract

This paper proposes learning disentangled but complementary face features with a minimal supervision by face identification. Specifically, we construct an identity Distilling and Dispelling Autoencoder (D^2AE) framework that adversarially learns the identity-distilled features for identity verification and the identity-dispelled features to fool the verification system. Thanks to the design of two-stream cues, the learned disentangled features represent not only the identity or attribute but the complete input image. Comprehensive evaluations further demonstrate that the proposed features not only preserve state-of-the-art identity verification performance on LFW, but also acquire comparable discriminative power for face attribute recognition on CelebA and LFWA. Moreover, the proposed system is ready to semantically control the face generation/editing based on various identities and attributes in an unsupervised manner.

1. Introduction

Learning distinctive yet universal feature representations has drawn long-lasting attention in the community of face analysis due to its pivotal role in various face-related problems such as face verification and attribute recognition [40, 37, 23, 5, 28, 25], as well as generative face modeling and controllable editing [29, 47, 21, 13, 21, 18]. Most contemporary methods learn the facial features specific to predefined supervision (*e.g.* identities, attributes) [43, 38, 37, 41, 39, 17, 5, 28], and thus hamper these features to be readily generalized to the feature space for a new task without careful fine-tuning. For example, without explicit supervision, the learned features are likely not to reflect the connection between two attributes *smile* and *mouth open*, nor to relate identity-relevant attributes like *gender* and *race* closely to identity. Therefore, learning an almighty feature representation generalizable to any face-related tasks is significant in the field of face analysis and possibly transferable to other fields such as pedestrian analysis.

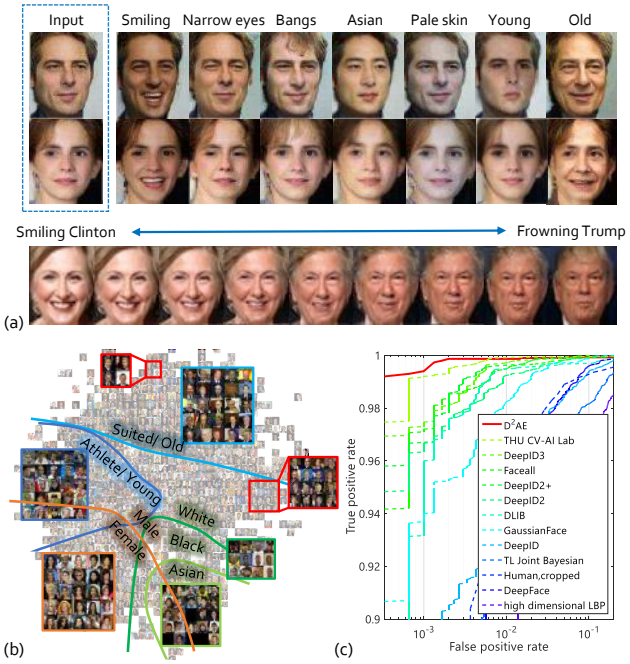


Figure 1. Representative face applications based on the learned face feature representations. (a) Semantic face editing such as identity-preserving attribute modification and identity transfer and interpolation. (b) The learned face features are trivially separable according to different attributes, visualized by Barnes-Hut *t*-SNE [30]. (c) The ROC curve on LFW face verification benchmark. The proposed face feature achieve the accuracy of 99.80% (single model), which outperforms most state-of-the-art methods without loss of ability in editing identity-related attributes.

Unlike prior arts that applied multi-task supervision [17] to extract quasi-universal features that are jointly effective across multiple predefined tasks, in this paper, we propose a novel feature learning framework with a minimal supervision by face identities. The learned representation not only produces *identity-distilled* features that discriminatively focus on inter-personal differences with identity supervision, but also effectively extracts the hidden *identity-*

*They contributed equally to this work

dispelled features to capture complementary knowledge including intra-personal variances and even background clutters. Analogous to the adversarial learning paradigm [8, 36, 7], the identity-dispelled features are fooled to make non-informative judgment over the identities. We claim that the learned face features own sufficient flexibility to improve face identification and are extensible to model diverse patterns like attributes for different tasks. Moreover, these features also enable controllable face generation and editing even without tedious training of the control units. Fig. 1 illustrates the superiority of the proposed feature representation over the state of the arts in representative applications.

In this study, we wish to highlight three advantages of this innovative feature learning framework:

(1) *Adversarial Supervision* – The identity-dispelled features are intactly encoded with the novel adversarial supervision. Distinct from those supervised by additional handcrafted tasks, the proposed scheme is simple yet effectively guarantees better generalization and completeness of the representation with complementary features.

(2) *Interpretability* – Our learning scheme provides a comprehensive and decomposable interpretation of the knowledge by adaptively assembling the identity-distilled and identity-dispelled features. We also find the learned features are compact and smoothly spread in a convex space. The extracted face features enhence face identity verification and are well prepared for various bypass tasks such as face attribute recognition and semantic face generation/editing.

(3) *Two-stream End-to-End Framework* – The proposed framework is end-to-end learned and solely supervised by face identities, distinguished from the conventional methods equipped with alternate adversarial supervision. By reusing the learned face features, other face-related tasks can be readily plugged in without fine-tuning the network.

(4) *Discriminative information preserving* – To be a minor contribution, the performance of face recognition gets improved if the attribute bias against identities occurs in the training set, which is often the case in small datasets.

The aforementioned advantages of the Distilling and Dispelling Autoencoder (D^2AE) framework are examined and analyzed through comprehensive ablation studies. The proposed approach is compared both quantitatively and qualitatively with state of the arts, achieving 1) accuracy of 99.80% on face verification benchmark LFW[12], 2) remarkable performance on attribute classification benchmarks LFWA[26] & CelebA[26], and 3) superior capability on various generative tasks such as semantic face editing.

2. Related Work

Learning Feature Representations. With the goal of disentangling distinct but informative factors in the data, representation learning has drawn much attention in the machine learning community [2, 3]. It is typically categorized into generative modeling and discriminative modeling. Given observations, *Discriminative Models* directly model the conditional probability distribution of the target variables and have accompanied and greatly nourished the rapid progress in classification and regression tasks, such as large-scale facial identity classification [43, 40, 38, 37, 41, 32] and attribute classification [28, 6]. *Generative Models*, as opposed to discriminative models, learn feature representations by modeling how the data was generated based on the joint distribution of the observed and target variables. For example, the autoencoder (AE) framework [19, 4, 11, 10] proposes that an encoder first extracts features from the data, followed by a decoder that maps from feature space back into input space. With the ability to automatically encode expressive information from the data space, various AE models [46, 35, 16] have been developed.

Combining Discriminative and Generative Models. While discriminative models generally perform better, they inherently require supervision, being less flexible than generative models. The pioneering work of GAN [8] combines them together, and a large body of literature has been built upon it. Impressive progress has been made on a variety of tasks, such as image translation [15], image editing [51], image inpainting [33, 1], and texture synthesis [20, 22].

Disentangled Representation. Despite impressive previous progress on improving either visual quality or recognition accuracy, disentangling the feature representation space is still under-explored. Some previous works tried to disentangle the representations in tasks such as pose-invariant recognition [45, 44] and identity-preserving image editing [13, 21, 18]. However, they usually require explicit attribute supervision and encode each attribute as a separate element in the feature vector. These methods are limited to representing a fixed number of attributes and need retraining once a new attribute is added. Makhzani *et al.* [31] encode class information into a discrete one-hot vector, with style information following a Gaussian distribution, but its training is likely to be unstable.

Our proposed D^2AE model overcomes these limitations. With no attribute supervision, the identity-dispelled feature encodes various attributes, to which the identity-distilled feature is invariant. In contrast, [49] extracts features that are only pose-invariant, which is a special case of our model. [34] learns a representation that is only invariant to pose and requires multi-source supervisions, while our method learns a representation invariant to any non-ID attributes and requires no supervision other than ID. Moreover, without popular regularization on distribution like VAE [16], our learned hidden space is naturally compact and smooth.

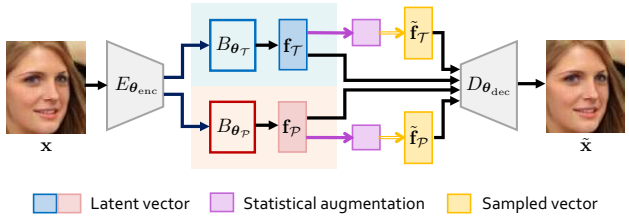


Figure 2. The Distilling and Dispelling Autoencoder model.

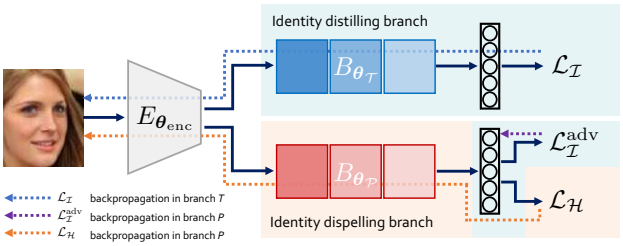


Figure 3. The encoding module for extracting disentangled face features.

3. Learning Disentangled Face Features

In this section, we introduce the identity Distilling and Dispelling Autoencoder (D^{2}AE) framework that end-to-end learns disentangled face features with an external supervision signal from face identity.

Given an input face image $\mathbf{x}$, the identity-distilled feature $\mathbf{f}_\mathcal{T} \in \mathbb{R}^{N_\mathcal{T}}$ and identity-dispelled feature $\mathbf{f}_\mathcal{P} \in \mathbb{R}^{N_\mathcal{P}}$ jointly serve as a complete representation of the face, as illustrated in Fig. 2. The encoding module is composed of a stack of shared convolutional layers $E_{\boldsymbol{\theta}_{\text{enc}}}(\mathbf{x})$, followed by the parallel identity distilling branch $B_{\boldsymbol{\theta}_\mathcal{T}}$ and identity dispelling branch $B_{\boldsymbol{\theta}_\mathcal{P}}$. Face identities supervise the training for $\mathbf{f}_\mathcal{T}$ and also adversarially guide the learning for $\mathbf{f}_\mathcal{P}$. Finally, a decoding module $D_{\boldsymbol{\theta}_{\text{dec}}}(\cdot)$ reconstructs $\tilde{\mathbf{x}}$ from the fused semantic features, so as to encourage the learned face features to encode a full representation of the input image.

3.1. Identity Distilling Branch

As visualized in Fig 3, the identity distilling branch D^{2}AE-$\mathcal{T}$ extracts $\mathbf{f}_\mathcal{T}$ by a convolutional subnet $B_{\boldsymbol{\theta}_\mathcal{T}}$ after $E_{\boldsymbol{\theta}_{\text{enc}}}(\mathbf{x})$, written as $\mathbf{f}_\mathcal{T} = B_{\boldsymbol{\theta}_\mathcal{T}}(E_{\boldsymbol{\theta}_{\text{enc}}}(\mathbf{x}))$. Specifically, $\mathbf{f}_\mathcal{T}$ is non-linearly mapped by $\texttt{softmax}$ function to an N_{ID}-dimentional identity prediction distribution, which corresponds to the N_{ID} identities provided by the applied large-scale training dataset for face identification [9, 26],

$$\mathbf{y}_\mathcal{T} = \texttt{softmax}(\mathbf{W}_\mathcal{T}\mathbf{f}_\mathcal{T} + \mathbf{b}_\mathcal{T}). \tag{1}$$

The predicted distribution $\mathbf{y}_\mathcal{T}$ is compared to the ground truth one-hot face labels $\mathbf{g}_\mathcal{I}$ via the cross-entropy loss

$$\mathcal{L}_\mathcal{I} = \sum_{j=1}^{N_{\text{ID}}} -\mathbf{g}_\mathcal{I}^j \log \mathbf{y}_\mathcal{T}^j = -\log \mathbf{y}_\mathcal{T}^t, \tag{2}$$

where t indicates the ground truth index. Please note that the optimization over $\mathcal{L}_\mathcal{I}$ only updates the identity distilled branch $B_{\boldsymbol{\theta}_\mathcal{T}}$ and the shared layers $E_{\boldsymbol{\theta}_{\text{enc}}}$.

3.2. Identity Dispelling Branch

The identity dispelling branch D^{2}AE-$\mathcal{P}$ suppresses the identity information and tries to encode the complementary facial information. Similar to the identity distilling branch D^{2}AE-$\mathcal{T}$, it also consists of a subnet $\mathbf{f}_\mathcal{P} = B_{\boldsymbol{\theta}_\mathcal{P}}(E_{\boldsymbol{\theta}_{\text{enc}}}(\mathbf{x}))$ appended with a fully connected layer towards the identity prediction distribution $\mathbf{y}_\mathcal{P} = \texttt{softmax}(\mathbf{W}_\mathcal{P}\mathbf{f}_\mathcal{P} + \mathbf{b}_\mathcal{P})$. To enable the complementary feature extraction, we propose an adversarial supervision.

On one hand, we also need to train an identity classifier based on the extracted features $\mathbf{f}_\mathcal{P}$ and supervised by the cross entropy loss $\mathcal{L}_\mathcal{I}^{\text{adv}} = -\log \mathbf{y}_\mathcal{P}^t$. The difference between the training of $\mathbf{y}_\mathcal{P}$ and $\mathbf{y}_\mathcal{T}$ is that the gradients of $\mathcal{L}_\mathcal{I}^{\text{adv}}$ are only back-propagated to the classifier but do not update the preceding layers in $B_{\boldsymbol{\theta}_\mathcal{P}}$ and $E_{\boldsymbol{\theta}_{\text{enc}}}$, analogous to the discriminator in GAN models [8].

On the other hand, we need to train the identity dispelling branch to fool the identity classifier, where the so-called "ground truth" identity distribution $\mathbf{u}_\mathcal{I}$ is required to be constant over all identities and equal to $\frac{1}{N_{\text{ID}}}$. Therefore, it is also equivalent to minimizing the negative entropy of the predicted identity distribution

$$\mathcal{L}_\mathcal{H} = \sum_{j=1}^{N_{\text{ID}}} \mathbf{u}_\mathcal{I}^j \log \mathbf{y}_\mathcal{P}^j = \frac{1}{N_{\text{ID}}} \sum_{j=1}^{N_{\text{ID}}} \log \mathbf{y}_\mathcal{P}^j, \tag{3}$$

where the gradients for $\mathcal{L}_\mathcal{H}$ are back-propagated to $B_{\boldsymbol{\theta}_\mathcal{P}}$ and $E_{\boldsymbol{\theta}_{\text{enc}}}$ with the identity classifier fixed.

It is worth mentioning that the proposed adversarial supervision does not introduce degenerated solutions for $\mathbf{f}_\mathcal{P}$ (*e.g.*, non-informative patterns). However, if we remove the identification loss $\mathcal{L}_\mathcal{I}^{\text{adv}}$ and allow the gradients in $\mathcal{L}_\mathcal{H}$ to update the identity classifier, few efforts are needed for this branch to deceive $\mathcal{L}_\mathcal{H}$, *e.g.*, by simply changing the identity classifier to produce non-informative outputs. In this case, there is certainly no guarantee that $\mathbf{f}_\mathcal{P}$ extracts the identity-dispelled features.

The total loss for this branch is the summation of $\mathcal{L}_\mathcal{I}^{\text{adv}}$ and $\mathcal{L}_\mathcal{H}$, and the two features can be learned simultaneously with the proposed feature-level adversarial training, no longer in need of a fragile alternate training process as is required in most GAN models [8].

3.3. Encoder-Decoder Architecture

While loss functions imposed on identity distilling and dispelling branches encourage a split of the input image represention, there is no guarantee that the combination of $\mathbf{f}_\mathcal{T}$ and $\mathbf{f}_\mathcal{P}$ form a complete encoding of the input image $\mathbf{x}$. In fact, we can only ensure that $\mathbf{f}_\mathcal{T}$ represents the identity while

$\mathbf{f}_{\mathcal{P}}$ wipes off the identity, but whether the remaining information has been encoded is not clear. An encoder-decoder architecture is used to further enhance the learned feature embedding by imposing a bijective mapping between an input image and its semantic features. For simplicity, we apply the ℓ_2 norm as the reconstruction loss

$$\mathcal{L}_{\mathcal{X}} = \frac{1}{2}\|\mathbf{x} - D_{\boldsymbol{\theta}_{\text{dec}}}(\mathbf{f}_{\mathcal{T}}, \mathbf{f}_{\mathcal{P}})\|_2^2. \quad (4)$$

Since $\mathcal{L}_{\mathcal{I}}$ encourages $\mathbf{f}_{\mathcal{T}}$ to distill identity-aware features, the reconstruction loss forces $\mathbf{f}_{\mathcal{P}}$ to encode all of the remaining identity-irrelevant information to recover the original image.

3.4. Statistical Augmentation

To encourage the channel-wise feature distribution in $\mathbf{f}_{\mathcal{T}}$ and $\mathbf{f}_{\mathcal{P}}$ to be sufficiently distinctive and concentrated, we may augment the features with Gaussian noises as

$$\tilde{\mathbf{f}}_{\iota}^{i} = \mathbf{f}_{\iota}^{i} + \varepsilon \boldsymbol{\sigma}_{\iota}^{i}, \forall i \in \{1, \ldots, N_{\iota}\} \text{ with } \varepsilon \sim \mathcal{N}(0, 1), \quad (5)$$

where $\iota \in \{\mathcal{T}_{id}, \mathcal{P}\}$ indicates the feature type. The scale is the standard deviation of each element in $\mathbf{f}_{\iota}$, which can be efficiently calculated via a strategy similar to batch normalization [14]. When plugging the augmentation operations right after $\mathbf{f}_{\mathcal{I}}$ and $\mathbf{f}_{\mathcal{P}}$, the loss functions aforementioned can be straightforwardly modified by the augmented features.

A slight perturbation on $B_{\boldsymbol{\theta}_{\mathcal{T}}}$ forces the ID-distilling space to learn larger margins between identities. Furthermore, since the perturbation in each channel is independent, it is useful for channel-decoupling, which is similar to the mechanism of dropout. Therefore, the resultant features are densely concentrated and nearly independent across channels. Moreover, it inherently condenses the semantic feature space expanded by $\mathbf{f}_{\mathcal{T}}$ and $\mathbf{f}_{\mathcal{P}}$, increasing the network interpretability for any face image.

3.5. Learning Algorithm

Learning the face features involves a single objective that consists of the feature extraction losses $\mathcal{L}_{\mathcal{I}}$, $\mathcal{L}_{\mathcal{I}}^{\text{adv}}$ and $\mathcal{L}_{\mathcal{H}}$, as well as the reconstructed loss $\mathcal{L}_{\mathcal{X}}$. Moreover, when statistically augmented by $\tilde{\mathbf{f}}_{\mathcal{T}}$ and $\tilde{\mathbf{f}}_{\mathcal{P}}$, we also incorporate the objective with the augmented reconstruction loss $\tilde{\mathcal{L}}_{\mathcal{X}}$. The final objective is a weighted combination:

$$\mathcal{L} = \lambda_{\mathcal{T}}\mathcal{L}_{\mathcal{I}} + \lambda_{\mathcal{P}}\left(\mathcal{L}_{\mathcal{I}}^{\text{adv}} + \mathcal{L}_{\mathcal{H}}\right) + \lambda_{\mathcal{X}}\left(\tilde{\mathcal{L}}_{\mathcal{X}} + \mathcal{L}_{\mathcal{X}}\right). \quad (6)$$

We apply the stochastic gradient descent solver to minimize the above objective and update the network parameters. As depicted in Fig. 3, the dotted blue line and the dotted orange line present the back-propagation routines for $\mathcal{L}_{\mathcal{I}}$ and $\mathcal{L}_{\mathcal{H}}$, respectively, and the purple line demonstrates the simultaneous back-propagation path for $\mathcal{L}_{\mathcal{I}}^{\text{adv}}$. Similarly, the gradient updates for the encoder-decoder network parameters are back-propagated through the whole autoencoder except the identity classifiers for both branches.

4. Experimental Setting

4.1. Datasets and Preprocessing

Datasets. The proposed D^{2}AE model is trained on the MS-Celeb-1M dataset [9], which is currently the largest face recognition dataset. For purpose of assessing its generalization ability, the trained model is evaluated on the LFW dataset [26], and the overlapped images both in the MS-Celeb-1M and LFW datasets are manually pruned from the MS-Celeb-1M dataset. Therefore, 4M checked images with 80K identities in the MS-Celeb-1M dataset are used for training and validation, with a split ratio of $9:1$.

Preprocessing. Faces in the images are detected and aligned by RSA [24]. Face patches are first cropped so that the interpupillary distance is equal to 35% of the patch width, and then they are resized to 235×235.

4.2. Detailed Implementation

The proposed D^{2}AE model consists of an encoding module $E_{\boldsymbol{\theta}_{\text{enc}}}$, two parallel subnets $B_{\boldsymbol{\theta}_{\mathcal{T}}}$ and $B_{\boldsymbol{\theta}_{\mathcal{P}}}$ to decompose the face features, and a decoding module $D_{\boldsymbol{\theta}_{\text{dec}}}$.

Encoding Module $E_{\boldsymbol{\theta}_{\text{enc}}}$. We use Inception-ResNet[42] as the backbone of $E_{\boldsymbol{\theta}_{\text{enc}}}$. The input size is modified to 235×235 and the final `AvePool` layer is replaced by $B_{\boldsymbol{\theta}_{\mathcal{T}}}/B_{\boldsymbol{\theta}_{\mathcal{P}}}$.

Subnets $B_{\boldsymbol{\theta}_{\mathcal{T}}}/B_{\boldsymbol{\theta}_{\mathcal{P}}}$. Each subnet has 3 `conv` layers, one global `AvePool` and one `FC` layer. These branches extract two 256 dimensional feature representations for $\mathbf{f}_{\mathcal{T}}$ and $\mathbf{f}_{\mathcal{P}}$.

Decoding Module $D_{\boldsymbol{\theta}_{\text{dec}}}$. $D_{\boldsymbol{\theta}_{\text{dec}}}$ decodes the concatenation $\{\mathbf{f}_{\mathcal{T}}, \mathbf{f}_{\mathcal{P}}\}$ into a face image with the same size as the input image. The concatenated feature vectors are firstly passed into an `FC` layer to increase the feature dimension and then reshaped to squared feature maps, which are fed into 20 `conv` layers interlaced with 6 `upsampling` layers to obtain the output image.

Model Training. The whole network is trained in an end-to-end manner with all of the supervisory signals simultaneously added to the system. The batch size of the input images is 192, distributed on 16 NVIDIA Titan X GPUs. The base learning rate is set to 0.01 and is declined by 0.1 every 10 epochs. It takes around 31 epochs in total for the training to converge. The weights in the training objective is set as $\lambda_{\mathcal{T}} = 1$ for $\mathcal{L}_{\mathcal{I}}$, $\lambda_{\mathcal{P}} = 0.1$ for $\mathcal{L}_{\mathcal{I}}^{\text{adv}}$ and $\mathcal{L}_{\mathcal{H}}$, and $\lambda_{\mathcal{X}} = 1.81 \times 10^{-5}$ for the $\mathcal{L}_{\mathcal{X}}$ and $\tilde{\mathcal{L}}_{\mathcal{X}}$ in the encoder-decoder architecture.

4.3. Model Evaluation

We select three representative face-related applications to demonstrate the effectiveness of the proposed face features. They share the same feature extraction pipeline that concatenates the face features from the proposed identity distilling and dispelling branches $\mathbf{f}_{\mathcal{C}}^{\top} = [\mathbf{f}_{\mathcal{T}}^{\top}, \mathbf{f}_{\mathcal{P}}^{\top}]$.

Face Identification. We select the LFW dataset as the test bed for face identification, following the standard evaluation protocols with two popular metrics: accuracy and

Branch	Identity		Attribute	
	Acc	TPR	Acc	#drop
D^{2}AE-$\mathcal{T}$	99.78	**99.63**	79.78	–
D^{2}AE-$\mathcal{P}$	64.13	5.3	81.99	–
D^{2}AE-$\mathcal{P}$ w/o $\mathcal{L}_{\mathcal{H}}$	71.2	8.67	80.47	36/40
D^{2}AE-$\mathcal{P}$ w/o $\mathcal{L}_{\mathcal{I}}^{\text{adv}}$	67.13	5.63	78.32	36/40
D^{2}AE	**99.80**	99.40	**83.16**	–

Table 1. Evaluation of the D^{2}AE model on identity verification and attribute recognition, comparing different combinations of branches and losses. The last column shows the number of attributes (out of the total number of 40) that suffer a performance drop compared to D^{2}AE-$\mathcal{P}$ with complete losses. Bold font marks the best result in each column.

TPR@0.001FPR[1]. The identity similarity is calculated by the cosine distance between two feature vectors.

Face Attribute Recognition. We further validate the discriminative power of the proposed face features on face attribute recognition over the CelebA [26] and LFWA [26] datasets. Each image in these datasets is annotated with $N_{\text{att}} = 40$ face attributes. The performance is evaluated by the metric of accuracy, as suggested by Liu *et al*. [26]. Since our model does not receive the attribute supervision, we extract the combined features $\mathbf{f}_{\mathcal{C}}$ and then train a linear SVM supervised by the labeled attributes in these datasets.

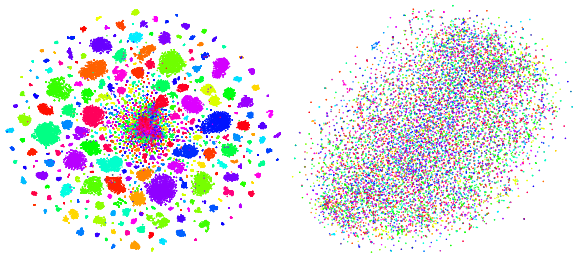
Face Editing. We also show the superiority of the proposed model in identity-preserving attribute editing and attribute-preserving identity exchanging. [2] By editing the semantic face features within the valid range of the feature space, we can observe rich semantic variations in the decoded image. (1) The identity-preserving attribute editing modifies $\mathbf{f}_{\mathcal{P}}$ by adding an incremental vector along the max-margin direction $\mathbf{w}_n$ of an attribute according to the trained linear SVM classifier for face attribute recognition. Thus the modified feature is $\mathbf{f}_{\mathcal{P}}^* = \mathbf{f}_{\mathcal{P}} + \alpha_n \mathbf{w}_n$, where $\alpha_n \mathbf{w}_n$ ranges within the confidence interval controlled by the learned standard deviation depicted in Sec. 3.4 for a reasonable modification of the input image. To support editing multiple attributes, $\mathbf{f}_{\mathcal{P}}^*$ can be extended to $\mathbf{f}_{\mathcal{P}}^* = \mathbf{f}_{\mathcal{P}} + \sum_{n=1}^{N_{\text{att}}} \alpha_n \mathbf{w}_n$, where $\boldsymbol{\alpha}$ is constrained in a similar fashion. (2) The attribute-preserving identity exchanging replaces $\mathbf{f}_{\mathcal{T}}^A$ from the image of one identity A with $\mathbf{f}_{\mathcal{T}}^B$ from another identity B, while keeping $\mathbf{f}_{\mathcal{P}}^A$ unchanged. Or more generally, the identity can be smoothly varied along the identity manifold, such as $\mathbf{f}_{\mathcal{T}}^* = \beta \mathbf{f}_{\mathcal{T}}^A + (1-\beta)\mathbf{f}_{\mathcal{T}}^B, \forall \beta \in [0,1]$. The generated face image has the target identity with the rest semantic information and background remaining the same.

5. Ablation Study

A unique advantage of the D^{2}AE model is its capability of learning complete and disentangled features from the input image, *i.e*., the identity-distilled feature and identity-dispelled feature. Successful extraction of the expected features is guaranteed by several pivotal components, *i.e*., two complementary branches for information selectivity, adversarial supervision for identity dispelling and statistical augmentation for a compact hidden space. In this section, we will validate their effectiveness by ablation studies, where the LFW(A) face dataset is employed for evaluation.

[1]We take TPR for short in the following experiments.

[2]To prove the robustness and consistency of our model, identities of visualized results are re-used for multiple times in the main paper.



(a) $\mathbf{f}_{\mathcal{T}}$ by D^{2}AE-$\mathcal{T}$ (b) $\mathbf{f}_{\mathcal{P}}$ by D^{2}AE-$\mathcal{P}$

Figure 4. Barnes-Hut t-SNE [30] visualization of the features extracted by two branches (a) D^{2}AE-$\mathcal{T}$ and (b) D^{2}AE-$\mathcal{P}$ on LFW. The colors indicate different identities. Best viewed in color.

5.1. Branch Selectivity

We find that the identity distilling branch D^{2}AE-$\mathcal{T}$ and the identity dispelling branch D^{2}AE-$\mathcal{P}$ indeed have distinctive capacities in representing different features.

Identity-distilled Feature $\mathbf{f}_{\mathcal{T}}$. Comparing TPR of identity verification in Table 1, $\mathbf{f}_{\mathcal{T}}$ from D^{2}AE-$\mathcal{T}$ is significantly superior to $\mathbf{f}_{\mathcal{P}}$ from D^{2}AE-$\mathcal{P}$. The extremely low value in TPR for D^{2}AE-$\mathcal{P}$ indicates that this branch has expelled most of the identity-related information from the input image. To further demonstrate their discrepancy on discriminative capability, we visualize the high-level features generated by these branches based on Barnes-Hut t-SNE [30]. As shown in Fig. 4 (a), D^{2}AE-$\mathcal{T}$ generates a set of densely clustered features for each identity with distinct boundaries between features from different identities. Moreover, D^{2}AE-$\mathcal{T}$ has almost the same identity verification result as that of the combined features (named as D^{2}AE in Table 1) and even outperforms the latter by the TPR metric. Not surprisingly, it also proves that the features by D^{2}AE-$\mathcal{T}$ have an extraordinary ability to represent identity-aware information.

Identity-dispelled Feature $\mathbf{f}_{\mathcal{P}}$. In contrast to its poor ability of extracting identity-aware features, D^{2}AE-$\mathcal{P}$ presents its superiority in face attribute recognition over D^{2}AE-$\mathcal{T}$, as shown in Table 1. Interestingly, the features learned by D^{2}AE-$\mathcal{T}$ also present certain discriminative ability to recognize some attributes. As shown in Fig. 5, the feature $\mathbf{f}_{\mathcal{P}}$ outperforms $\mathbf{f}_{\mathcal{T}}$ on 27 attributes in total. For most common attributes that are independent from identity such as *pale skin* and *smile*, the identity-dispelled feature exhibits more discriminative potential. However, other identity-aware attributes including genders (*e.g*., *male*) and races (*e.g*., *Indian* and *Asian*) tend to be better recognized through the

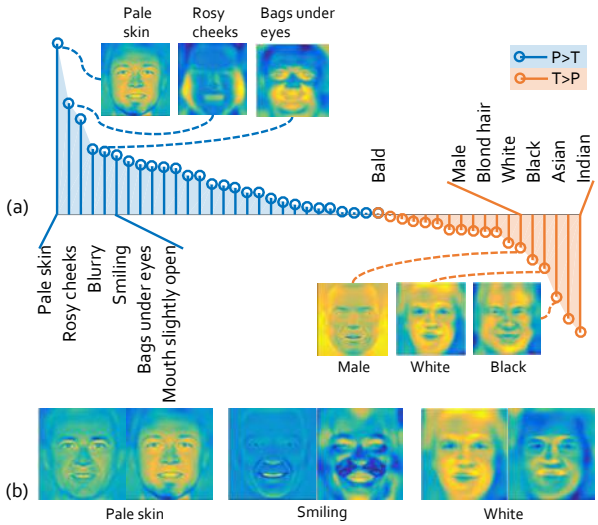


Figure 5. (a) Performance comparison on attribute recognition with features extracted from either D^{2}AE-$\mathcal{T}$ or D^{2}AE-$\mathcal{P}$. The magnitude of each bin illustrates the difference between the recognition accuracy of D^{2}AE-$\mathcal{T}$ and that of D^{2}AE-$\mathcal{P}$. Blue bins indicate attributes where features from D^{2}AE-$\mathcal{P}$ excel, and red bins show attributes where features from D^{2}AE-$\mathcal{T}$ win. (b) The residual maps correspond to three representative attributes. Images on the left of each pair are generated by D^{2}AE-$\mathcal{T}$ and the right ones are by D^{2}AE-$\mathcal{P}$.

identity-distilled feature. Besides, attributes on the borderline (*e.g.*, *bald*) with similar performance shared by $\mathbf{f}_\mathcal{T}$ and $\mathbf{f}_\mathcal{P}$, are mostly vaguely defined between identity-related and identity-irrelevant attributes.

To visualize the discriminative response of attribute onto the image space, we synthesize a set of residual images responsive to each attribute against the mean image from the LFW dataset. We synthesize the attribute-augmented face image by first adding a unit vector $\mathbf{w}_n, n \in \{1, \ldots, N_{\text{att}}\}$ to the mean feature $\bar{\mathbf{f}}_\mathcal{T}$ (or $\bar{\mathbf{f}}_\mathcal{P}$) and then decoding the combined feature to a face image. The residual images are attribute-augmented face images subtracted by the mean image. According to the results shown in Fig. 5 (a), residual maps with respect to D^{2}AE-$\mathcal{P}$ for identity-irrelevant attributes usually display high responses at local semantic regions, such as the facial skins for *pale skin* and cheeks for *rosy cheeks*. In contrast, the residual maps with respect to D^{2}AE-$\mathcal{T}$ for identity-related attributes usually have holistic responses, which are distributed throughout the whole image, *e.g.*, the maps for *gender* and *race*. Comparing the residual maps with respect to each branch, $\mathbf{f}_\mathcal{T}$ tends to be more responsive to identity-aware attributes while $\mathbf{f}_\mathcal{P}$ displays stronger responses to identity-irrelevant attributes, as presented in Fig. 5 (b).

In addition to quantitative comparison, in Fig. 6 we also provide qualitative results of face attribute editing by modifying features extracted from D^{2}AE-$\mathcal{T}$ and D^{2}AE-$\mathcal{P}$. Modifying $\mathbf{f}_\mathcal{T}$ has minimal influence on face variation when editing identity-irrelevant attributes like *smiling*, but it effectively controls the face warping to a different *race*. Conversely, modification on $\mathbf{f}_\mathcal{P}$ can hardly change the race of a face, but it continuously transforms a face from a neutral expression to *smiling*.

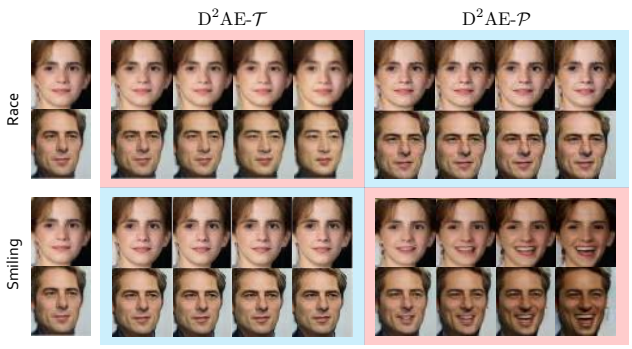


Figure 6. Modifying features extracted from two branches for face attribute editing. ID-related and ID-irrelevant attributes are extracted through D^{2}AE-$\mathcal{T}$ and D^{2}AE-$\mathcal{P}$, respectively.

5.2. Loss Functionality

The adversarial training in the identity dispelling branch is a distinctive feature of the D^{2}AE network. We examine the loss terms in D^{2}AE-$\mathcal{P}$ to demonstrate their necessities for the effective training of identity-dispelled features.

Identity Confusion Loss $\mathcal{L}_\mathcal{H}$. Removing the adversarial loss $\mathcal{L}_\mathcal{H}$ in D^{2}AE-$\mathcal{P}$ causes a failure of $\mathbf{f}_\mathcal{P}$ to completely dispel identity-related attributes. As a consequence, it performs better on identity verification than the model with combined losses, but its performance on attribute recognition is slightly degraded, as presented in Table 1. 36 out of 40 attributes experience drop of recognition accuracy. In contrast, both accuracy and TPR metrics for identity verification obtain remarkable gains. In fact, the identity classification loss cannot effectively constrain $\mathbf{f}_\mathcal{P}$ as it only has an impact on the identity classifier during gradient update, thus there is no guarantee that the resultant $\mathbf{f}_\mathcal{P}$ is independent from the identity as expected.

Identity Classification Loss $\mathcal{L}_\mathcal{I}^{\text{adv}}$. Removing the identity classification loss $\mathcal{L}_\mathcal{I}^{\text{adv}}$, we only regularize $\mathbf{f}_\mathcal{P}$ to fool the identity verification system based on $\mathcal{L}_\mathcal{H}$ which updates its identity classifier. Thus its identity dispelling ability, *i.e.*, the ability of pruning identity information from $\mathbf{f}_\mathcal{P}$, is weaker than the combined losses for lack of explicit identity supervision on the identity classifier. According to Table 1, without $\mathcal{L}_\mathcal{I}^{\text{adv}}$, the performance of $\mathbf{f}_\mathcal{P}$ on identity verification is slightly improved, but its performance on attribute recognition is degraded with drops happening in 36 over 40 attributes. As $\mathcal{L}_\mathcal{H}$ explicitly confuses $\mathbf{f}_\mathcal{P}$ about the identity, it renders poorer identity verification than that trained by $\mathcal{L}_\mathcal{I}^{\text{adv}}$. Moreover, with a weaker ability to extract the information complementary to identity, $\mathcal{L}_\mathcal{H}$ also produces infe-

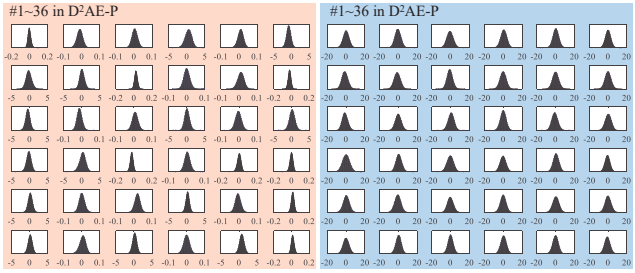


Figure 7. The distributions of the first 36 channels in features generated by D^{2}AE-$\mathcal{T}$ (left) and D^{2}AE-$\mathcal{P}$ (right). All the variables follow the Gaussian distribution.

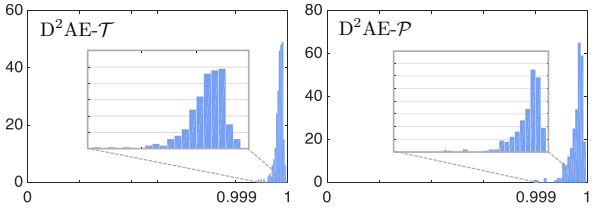


Figure 8. $adj - R^2$ scores of the channel-wise statistics of the features generated by D^{2}AE-$\mathcal{T}$ (left) and D^{2}AE-$\mathcal{P}$ (right).

rior attribute recognition results than that by $\mathcal{L}_{\mathcal{I}}^{\text{adv}}$.

5.3. Augmentation Necessity for Convex Space

The proposed statistical augmentation encourages the learned face features to be distinctive and densely Gaussian clustered in each channel as in Fig. 7. For quantitative evaluation, in Fig. 8, we plot two histograms to verify the required statistical property of $\mathbf{f}_{\mathcal{T}}$ and $\mathbf{f}_{\mathcal{P}}$ on the LFW dataset. These histograms are used to mimic the adjusted R-square ($adj - R^2$) score distribution for the channel-wise statistics of the features, where $adj - R^2$ is to measure how much the statistics look like a Gaussian distribution (a higher score is more alike). Obviously, both features are nearly Gaussian and almost all the channels have $adj - R^2$ scores higher than 0.99. They prove that the learned feature spaces for $\mathbf{f}_{\mathcal{T}}$ and $\mathbf{f}_{\mathcal{P}}$ are Gaussian and convex, whilst the features in these spaces are densely spread.

We also find that the learned feature space is compact and convex by densely interpolating two identities with different attributes. In Fig. 9, the interpolated face images change smoothly along the identity and attribute axes.

6. Performance Comparison

We also quantitatively and qualitatively compare the proposed D^{2}AE model with state-of-the-art approaches on the face-related tasks as mentioned above.

6.1. Face Identification

Comparison of results between the D^{2}AE model and the prior arts is plotted in Fig. 1 (c). According to the ROC curve and accuracy results listed in the legend, the D^{2}AE model achieves the best prformance.

Figure 9. Images generated by dense interpolation on space expanded by D^{2}AE-$\mathcal{T}$ and D^{2}AE-$\mathcal{P}$. D^{2}AE learns a convex hyper space. D^{2}AE-$\mathcal{T}$ controls identity and all identity-related attributes, such as *gender* and *race*. D^{2}AE-$\mathcal{P}$ controls all the other attributes like *smile* and *frontal*. Best viewed in color and zoomed in.

Method	MS-Celeb-1M		WebFace	
	Acc	TPR	Acc	TPR
Baseline	99.816	99.73	98.93	94.87
D^{2}AE	99.80	99.40	99.25	96.80

Table 2. Comparison of results on face identity verification.

In addition to the MS-Celeb-1M dataset, we also compare their face verification results on a smaller CASIA-WebFace dataset [48]. It only contains 0.49M images with around 10K identities, approximately one-tenth of the scale of the MS-Celeb-1M dataset. To further manifest the significance of the encoder-decoder structure compared to the encoder-only feature extraction scheme, we construct a baseline with the same encoder structure as included in the D^{2}AE architecture, denoted as the *Baseline* model.

As shown in Table 2, D^{2}AE achieves comparable performance with the Baseline model when trained on the MS-Celeb-1M dataset. Furthermore, if trained on the WebFace dataset, it even outperforms the Baseline model. This phenomenon occurs because the identity space may be biased towards some attributes due to the limited scale of the WebFace dataset. For example, it is possible that the face images of a certain identity in the dataset always appear in the same pose or expression. In this case, such particular pose or expression is likely to be used to define this identity, and hence it will be falsely encoded in the identity-related feature. Because the D^{2}AE model disentangles a face representation into an identity-distilled feature and a complementary identity-dispelled feature, it owns a superiority over the baseline model in the task of identity verification. In contrast, when trained on the MS-Celeb-1M dataset which has a sufficiently large scale, the baseline model is able to correctly extract identity-related information.

6.2. Face Attribute Recognition

We compare the proposed framework and two methods [27, 50] with supervision on face attribute. Perfor-

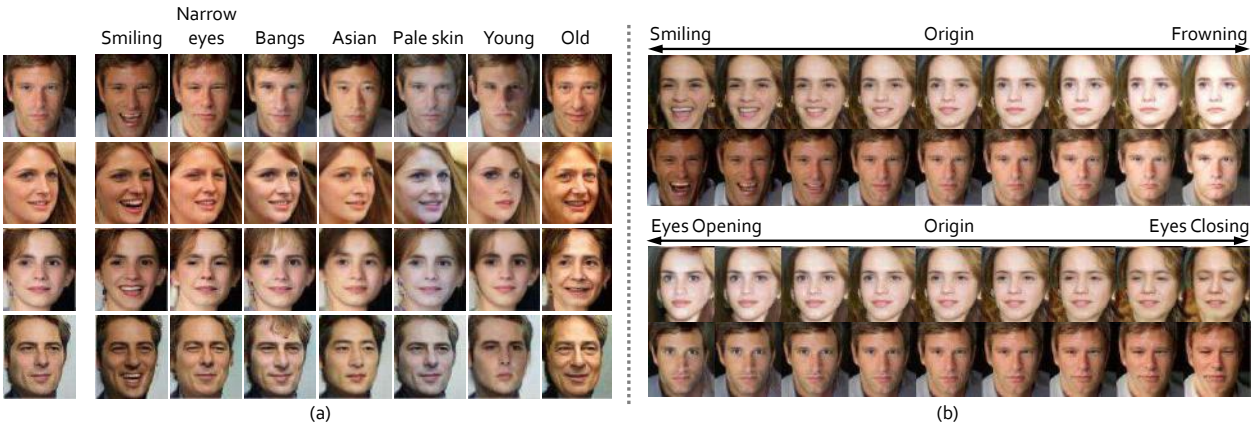


Figure 10. Results of (a) identity-aware attribute transfer and (b) identity-aware attribute interpolation. Zoom in for details.

Dataset	[27]	[50]	D^{2}AE
LFWA	83.85	81.03	83.16
CelebA	87.30	85.43	87.82

Table 3. Comparison of results on face attribute recognition.

mance is evaluated on two commonly employed datasets, *i.e*. LFWA [26] and CelebA [26]. As shown in Table 3, although the proposed method is entirely unsupervised in terms of face attribute, it achieves comparable results with the supervised methods.

In Fig. 1 (b) , we visualize the 2D embedding of all faces with attributes in LFWA dataset, based on the Barnes-Hut *t*-SNE method [30]. We can observe that with features extracted from D^{2}AE, the 2D embedding space can be automatically partitioned by either attributes or identities. Note that the attributes do not follow category boundaries. There are overlapping classification boundaries for different identity-aware attributes such as *sex* and *race*, while the face images for one identity are densely clustered.

6.3. Face Editing

We show that the proposed method presents superior performances on semantic face editing. We take several face images from the LFW dataset and reconstruct them with modification on different attributes as well as identities.

Identity-aware Attribute Editing. Fig. 10 (a) shows several portraits with one attribute changed at a time. Our model alters the attributes with well-preserved naturalness and identity. For either local (*e.g*., *smiling*, *narrow eyes*, and *bangs*) or global attributes (*e.g*., *Asian* and *age*), the model has successfully disentangled identity-distilled and identity-dispelled features. For example, even if the transformation of race certainly disturbs identity, almost all the identity-irrelevant attributes such as hair style, facial expression, and background color are well preserved. In the last column, the proposed model even synthesizes the unseen teeth and tongues when editing a portrait to *smile* and generates wrinkles when altering a person's age towards *older*.

Identity-aware Attribute Interpolation. Interpolation is performed by changing the weight of an attribute, which renders face images with different magnitudes of that attribute as shown in Fig. 10 (b). Our model enables smooth and natural change of either a female face from *smiling* to *frowning* or eyes of a male from being open to being closed.

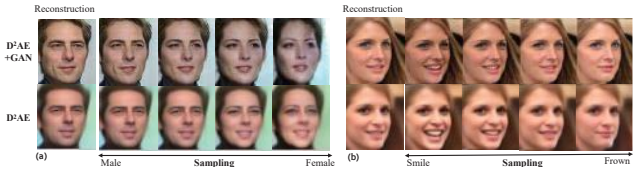


Figure 11. Identity-aware attribute transfer using the proposed model and its extension based on the GAN model.

Identity Transfer. To further illustrate the compactness of the convex feature space, a face image is gradually changed from one identity to another in the last row of Fig. 1 (a). It shows a smooth transition from a smiling female to a frowning male with the hair style gradually changed as well. More results are shown in Fig. 9.

Extension toward GANs. We find that the proposed model can be safely incorporated with the generative adversarial networks (GANs) [8], by switching the reconstruction loss $\mathcal{L}_{\mathcal{X}}$ to the adversarial loss from an additional discriminator. The reconstructed face images contain more realistic details and noises as shown in Fig. 11.

7. Conclusion

The proposed D^{2}AE disentangles the face representation into two orthogonal streams with novel adversarial supervision. Features in the two streams completely represent the information in the whole face, which are highly distinctive and densely distributed in a convex latent space. The learned features are ready for various applications such as face verification, attribute prediction and face editing, where the model all achieves state-of-the-art performances.

References

[1] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua. CVAE-GAN: fine-grained image generation through asymmetric training. *CoRR*, abs/1703.10155, 2017. 2

[2] Y. Bengio. Learning deep architectures for ai. *Foundations and Trends in Machine Learning*, 2:1–127, 2009. 2

[3] Y. Bengio, A. C. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35:1798–1828, 2013. 2

[4] H. Bourlard and Y. Kamp. Auto-association by multilayer perceptrons and singular value decomposition. *Biological Cybernetics*, 59:291–294, 1988. 2

[5] X. Cao, D. Wipf, F. Wen, and G. Duan. A practical transfer learning algorithm for face verification. In *Proc. ICCV*, 2013. 1

[6] D. Chen, X. Cao, F. Wen, and J. Sun. Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. In *Proc. CVPR*, 2013. 2

[7] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. S. Lempitsky. Domain-adversarial training of neural networks. In *Domain Adaptation in Computer Vision Applications.*, pages 189–209. 2017. 2

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014. 2, 3, 8

[9] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision*, pages 87–102. Springer, 2016. 3, 4

[10] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006. 2

[11] G. E. Hinton and R. S. Zemel. Autoencoders, minimum description length and helmholtz free energy. In *NIPS*, 1993. 2

[12] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled Faces in the Wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, 2007. 2

[13] R. Huang, S. Zhang, T. Li, and R. He. Beyond face rotation: Global and local perception GAN for photorealistic and identity preserving frontal view synthesis. *CoRR*, abs/1704.04086, 2017. 1, 2

[14] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167, 2015. 4

[15] P. Isola, J. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *CoRR*, abs/1611.07004, 2016. 2

[16] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[17] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *Proc. ICCV*, 2009. 1

[18] G. Lample, N. Zeghidour, N. Usunier, A. Bordes, L. Denoyer, and M. Ranzato. Fader networks: Manipulating images by sliding attributes. *CoRR*, abs/1706.00409, 2017. 1, 2

[19] Y. Lecun and F. Fogelman-Soulie. Modeles connexionnistes de l'apprentissage. 1987. 2

[20] C. Li and M. Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. *CoRR*, abs/1604.04382, 2016. 2

[21] M. Li, W. Zuo, and D. Zhang. Deep identity-aware transfer of facial attributes. *CoRR*, abs/1610.05586, 2016. 1, 2

[22] X. Liang, H. Zhang, and E. P. Xing. Generative Semantic Manipulation with Contrasting GAN. *ArXiv e-prints*, 2017. 2

[23] Y. Liu, H. Li, and X. Wang. Rethinking feature discrimination and polymerization for large-scale recognition. *arXiv preprint arXiv:1710.00870*, 2017. 1

[24] Y. Liu, H. Li, J. Yan, F. Wei, X. Wang, and X. Tang. Recurrent scale approximation for object detection in CNN. *CoRR*, abs/1707.09531, 2017. 4

[25] Y. Liu, J. Yan, and W. Ouyang. Quality aware network for set to set recognition. In *Proc. IEEE Int. Conf. Comput. Vision Pattern Recognit.*, pages 5790–5799, 2017. 1

[26] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proc. ICCV*, 2015. 2, 3, 4, 5, 8

[27] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015. 7, 8

[28] C. Lu and X. Tang. Surpassing human-level face verification performance on LFW with GaussianFace. Technical report, arXiv:1404.3840, 2014. 1, 2

[29] Y. Lu, Y. Tai, and C. Tang. Conditional cyclegan for attribute guided face image generation. *CoRR*, abs/1705.09966, 2017. 1

[30] L. V. D. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(2605):2579–2605, 2017. 1, 5, 8

[31] A. Makhzani, J. Shlens, N. Jaitly, and I. J. Goodfellow. Adversarial autoencoders. *CoRR*, abs/1511.05644, 2015. 2

[32] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *Proc. BMVC*, 2015. 2

[33] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2536–2544, 2016. 2

[34] X. Peng, X. Yu, K. Sohn, D. N. Metaxas, and M. Chandraker. Reconstruction-based disentanglement for pose-invariant face recognition. In *ICCV*, 2017. 2

[35] S. Rifai, P. Vincent, X. Muller, X. Glorot, and Y. Bengio. Contractive auto-encoders: Explicit invariance during feature extraction. In *ICML*, 2011. 2

[36] J. Schmidhuber. Learning factorial codes by predictability minimization. *Neural Computation*, 4(6):863–879, 1992. 2

[37] F. Schroff, D. Kalenichenko, and J. Philbin. FaceNet: A unified embedding for face recognition and clustering. In *Proc. CVPR*, 2015. 1, 2

[38] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. In *Proc. NIPS*, 2014. 1, 2

[39] Y. Sun, D. Liang, X. Wang, and X. Tang. Deepid3: Face recognition with very deep neural networks. *CoRR*, abs/1502.00873, 2015. 1

[40] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *Proc. CVPR*, 2014. 1, 2

[41] Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. In *Proc. CVPR*, 2015. 1, 2

[42] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. 2017. 4

[43] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. DeepFace: Closing the gap to human-level performance in face verification. In *Proc. CVPR*, 2014. 1, 2

[44] L. Tran, X. Yin, and X. Liu. Disentangled representation learning gan for pose-invariant face recognition, 07 2017. 2

[45] L. Tran, X. Yin, and X. Liu. Representation learning by rotating your faces. *CoRR*, abs/1705.11136, 2017. 2

[46] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *ICML*, 2008. 2

[47] X. Yan, J. Yang, K. Sohn, and H. Lee. Attribute2image: Conditional image generation from visual attributes. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*, pages 776–791, 2016. 1

[48] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 7

[49] X. Yin, X. Yu, K. Sohn, X. Liu, and M. K. Chandraker. Towards large-pose face frontalization in the wild. *CoRR*, abs/1704.06244, 2017. 2

[50] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev. Panda: Pose aligned networks for deep attribute modeling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1637–1644, 2014. 7, 8

[51] J. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros. Generative visual manipulation on the natural image manifold. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V*, pages 597–613, 2016. 2