The Universal Face Encoder: Learning Disentangled Representations Across Different Attributes

Sandipan Banerjee[1], Ajjen Joshi[2], and Jay Turcot[2]
[1] Samsung Research America, [2] Smart Eye
sandipan.b@samsung.com, {ajjen.joshi, jay.turcot}@smarteye.ai

Abstract

*Models that can learn orthogonal representations for different facial attributes (*e.g. *pose, lighting, identity, expressions) have proven to be beneficial for both discriminative and generative tasks. In this work, we propose the universal facial encoder (UFE) that can simultaneously encode different facial attributes as disentangled features from a single face image. We propose a variety of qualitative and quantitative metrics to evaluate feature orthogonality of the UFE and demonstrate superior disentanglement compared to traditional single-attribute encoding. We also show that these features can then be used to train lightweight prediction heads for multiple downstream classification tasks. Moreover, coupling the UFE with a style-based decoder enables hallucination of new face images composed of attributes taken from different samples. As experimentally demonstrated, the UFE allows us to pick and choose these attributes from label-disjoint datasets. A catalog of such synthetic composites can be used as supplemental training data or simply as stock photos.*

1. Introduction

Training a model to learn meaningful features that are composed of disentangled representations for different dimensions of variations can have several applications. A disentangled feature space can be beneficial for downstream discriminative (*e.g.* training feature classifiers) and generative tasks (*e.g.* generating synthetic images) as well as improve model interpretability and robustness. Consider the task of learning disjoint encodings for different facial attributes, such as pose, lighting, identity, expressions [12]. Training a single encoder to learn disentangled encodings for such attributes can be used to train robust but lightweight downstream classifiers from a single, shared encoder [7]. The disentangled representations can also be used as input to generative models that can produce images where the attributes can be implicitly or explicitly controlled [43].

In this work, we build a universal face encoder (UFE) that can simultaneously encode different facial attributes like identity, expression and lighting from a single image, irrespective of its domain association (*e.g.* facial pose variation or cross dataset training). Using these encoded features, coupled with dedicated lightweight prediction heads, the class association for each attribute can be predicted for an image. When deployed for real time inference, such global encoding can serve as a common backbone for multiple functionalities, and reduce the computation overhead.

To further reduce any leakage of information between one subspace to another, we formulate the sampling and objectives for model training to make the feature spaces orthogonal to each other. While the prediction heads (simple MLPs) are trained on the corresponding generated features with a classification loss, to enforce orthogonality we make use of a metric based contrastive loss. The relationship (positive or negative) between a face image pair, and consequently their corresponding embeddings, in different attribute spaces is specifically used for this task. These attribute embeddings are then fed to a style based decoder [29] for input reconstruction for pixel space interpretability. The decoder also enables us to utilize unlabeled samples, by creating new examples directly in the feature space, and use them for training by virtue of a self-supervised disentanglement loss, described in 5.

To evaluate and quantify the degree of orthogonality between the attribute sub-spaces we propose multiple metrics: (a) weak classifiers [16] to learn representations from one attribute with labels from another, (b) silhouette score between attribute clusters [1], (c) tSNE visualization [48] of generated features, and (d) a human rater study to visually evaluate feature disentanglement in decoded synthetic images [43]. The experimental results demonstrate significant subspace separability, both in the representation and pixel spaces, even when the input images are acquired across different camera positions [4] or belong to different datasets (*e.g.* MultiPIE [22] and FFHQ [28]) that have no common labels. Additionally, we show the sub-spaces in the proposed multi-attribute UFE to possess a higher degree of separability compared to traditional single attribute encoding models [5] while maintaining the same level of task accuracy when tested on in-the-wild data [27].

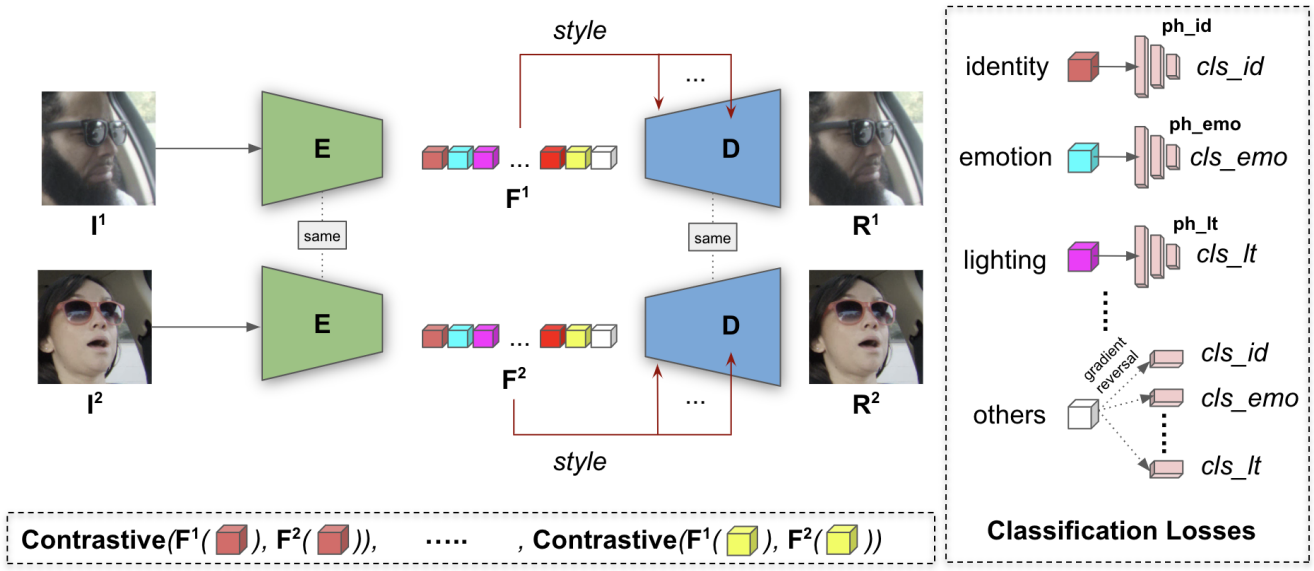

Figure 1. Our UFE framework has a single encoder **E** and decoder **D** that can take face images from different domains I^1 and I^2 and generate multi-attribute representations F^1 and F^2 respectively. Each attribute element (*e.g.* id) in this representation is classified using prediction heads **ph**, while smoother, pairwise relationships across attributes can be gleaned from contrastive learning. The representations also act as the guiding style for reconstructions R^1 and R^2 using **D**.

The main contributions of this paper are as follows:

(1) We build a universal face encoder (UFE) that can simultaneously and orthogonally encode different facial attributes from a single face image. The UFE can then act as a common backbone for training lightweight prediction heads for different downstream tasks.

(2) By attaching a style based decoder to the UFE, we hallucinate new images by compositing different encoded attributes from different images together and learn from them in an unsupervised manner by virtue of a disentanglement loss. We experimentally verify this loss' efficacy with training samples belonging to different datasets, even when their attribute labels are non-uniform or missing.

(3) We propose a set of numeric and visual metrics to evaluate the degree of feature orthogonality in the UFE representation space and demonstrate its embeddings to be better disentangled than those of a traditional encoder model.

2. Related Work

Image Translation: While style transfer models [18, 26] formulate exemplar based layerwise style manipulation in images, image-to-image translation for content editing took off with pix2pix [25] and CycleGAN [50]. These architectures work with encoder-decoder pairs that focus on target based transformation with a single latent vector [10, 11]. The StyleGAN models [28, 29] marry these two concepts by adaptively fusing style latents [24] from coarse to fine layers of the decoding generator. Recent papers [30, 33, 36] still utilize the decoding structure of StyleGAN but compartmentalize the encoding space for easy tweaking of editable features (*e.g.* facial attributes [8], RGB to anime [30], garment and body editing [36]).

Disentangled Learning: A common tactic for encoding space separation is to hallucinate or 3D render synthetic images varying in different attributes, and then to jointly learn the attribute differences from these images through an encoder [12, 32]. Explicit decoder and encoder pairs for attribute separation can also be learned for shape and geometry [45], content and motion [46], lighting and expression [3] disentanglement. By treating one attribute as static input and others as layerwise style, [39] build an age translation model while preserving the subject identity. The other tactic for subspace separation is to add a metric learning based objective to model training, where the encoder learns positive and negative relationship between input samples for different attribute spaces (*e.g.* color, shape, pose, identity). In [38], a triplet loss is used for this task while a pairwise contrastive loss is utilized to understand such relationships from synthetic image pairs in [43]. Finally, explicit latent spaces can also be built for labeled and unlabeled samples using separate encoding spaces [14–16].

Domain Invariance: To learn meaningful representations across different domains, researchers either build networks with shared embedding space [19] or introduce domain confusion for learning agnostic features [47]. This can be achieved by reversing domain-specific gradients during back propagation [17] or minimizing inter-domain feature distance [40]. Alternatively, teacher models can be used to distill useful information [21] or fine-tune [7, 49] for specific downstream tasks. These guiding models can also be used

to identify discriminative and challenging training samples from different domains [41] to further improve model generalization in a supervised [44] or unsupervised manner [4].

While most of the style based disentanglement models [39, 43] focus on the visual fidelity in pixel space, we pay more attention to the encoding aspect instead. Specifically, we design the UFE to not only produce style tensors for the generator but also meaningful attribute features on their own. This design allows the UFE to generate representations that can hallucinate visually pleasing results and be simultaneously used for downstream applications.

3. The Proposed UFE Framework

Architecture: As illustrated in Figure 1, our framework consists of a single downsampling residual encoder E that takes in a face image I and generates 1-D feature vectors $F = [f_{a1}, f_{a2}, f_{a3}, ..., f_{an}]$ pertaining to attributes $a1$ to an. A combination of these features, generated from face images collected from different domains (*e.g.* camera type/position or dataset) can be used for multi-domain, multi-attribute classification. The individual features in the encoder are separately generated from the residual maps, normalized to maintain conformity in value range. This is different from [38], where a single encoded feature F is adaptively allocated to the participating attributes. Once the feature vectors are extracted, the classification is done using individual prediction heads ph for each attribute space (*e.g.* ph_{ak} for the k-th attribute). We design ph as lightweight dense layers to map the feature f^i_{ak} to the corresponding label l^i_{ak} for i-th sample in the k-th attribute subspace.

Finally, a combination of these features $[f_{a1}, f_{a2}, ..., f_{ak}, ..., f_{an}]$ is fed to a style-based [29] upsampling decoder D to reconstruct I. While an ensemble of decoders could be trained for this task [37], where each decoder is assigned to a particular domain (*e.g.* camera position, facial pose, subject identity or dataset association), our goal is to make E domain agnostic. Therefore, we use a single D and utilize the gradient for domain invariance.

Loss Functions: The objectives used to train our framework are described below:

1. Classification: For each input I in a training iteration, we apply a classification objective to teach E and ph class association of the image within each of the different attribute spaces $[a1, a2, ..., an]$. This loss is applied separately for each attribute ak with separate one-hot vectors for each class association. The loss L_{cls} is computed as:

$$L_{cls} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{n}\sum_{j}^{C}(\mathrm{l}^i_{ak})_j log(ph_{ak}(f^i_{ak})_j), \quad (1)$$

where f^i_{ak} denotes the k-th attribute based feature from i-th sample I^i (*i.e.* $F = E(\mathrm{I}^i)$), and N, n, C are the batch size, number of attributes and classes for an attribute respectively. For each I, the gradient is backpropagated through the corresponding ph_{ak} and finally to E. Thus, E's weights get updated from the propagated gradients for all the attributes and domains during each training iteration.

2. Contrastive: Since the L_{cls} does not handle interactions between sample pairs that come from similar or different domains, we add a contrastive loss L_{con} to teach the encoder such relationships without requiring additional labels. To implement the objective, a pair of images I^i and I^j are taken as input and fed to E to extract the corresponding features $F^i = [f^i_{a1}, f^i_{a2}, ..., f^i_{an}]$ and $F^j = [f^j_{a1}, f^j_{a2}, ..., f^j_{an}]$ respectively. Based on the class association l^i_{ak} and l^j_{ak} of I^i and I^j for the attribute ak respectively, it is computed as:

$$L_{con} = \frac{1}{N_p}\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N}\sum_{k=1}^{n} y \cdot dist(f^i_{ak}, f^j_{ak}) + (1-y)\cdot max(m - dist(f^i_{ak}, f^j_{ak}), 0), \quad (2)$$

where N_p denotes the number of valid (i, j) pairs from the batch with size N. y is an indicator function set to 1 if $\mathrm{l}^i_{ak} = \mathrm{l}^j_{ak}$, otherwise 0. The $dist(\mathrm{x}, \mathrm{y})$ function computes the L2 distance between x and y, and m is a margin we set empirically. L_{con} pushes f^i_{ak} and f^j_{ak} away beyond m when $y = 0$ (*i.e.* different attribute labels) while they are pushed towards each other otherwise. Thus, L_{con} helps establish explicit pairwise relationships between batch samples irrespective of their domain.

As negative hard mining is not required for this particular version of the contrastive metric, we found it to be significantly faster than the triplet loss [38, 42], and more memory-efficient (*i.e.* controlled N_p range) compared to the harder-to-train temperature-normalized contrastive loss [9].

3. Reconstruction: To enable E to learn representations at a spatial level, we additionally include a reconstruction loss L_{rec}. The loss is fairly simple and computed as:

$$L_{rec} = \frac{1}{N}\sum_{i=1}^{N}\left|\mathrm{I}^i - D(E(\mathrm{I}^i))\right|, \quad (3)$$

The loss is simply a pixelwise L1 distance measure between the input and its reconstruction, and regularizes the encoder output by introducing a secondary task [4].

4. Other unlabeled attributes: Although the UFE can tackle multiple feature spaces at the same time, it is still limited by the number of attributes actually labeled in a dataset. Any unlabeled attribute can get distributed to one or more proximal feature spaces, depending on the data and model architecture, and hamper subspace separation.

To explicitly target such muddying of the feature spaces, we add an additional output feature f_{oth} [43] from E and add separate prediction heads $ph_{oth} = [ph_{oth1}, ph_{oth2}, ..., ph_{othn}]$ for all labeled attributes $[a1, a2, ..., an]$. However, when backpropagating from ph_{oth}, we reverse the gradient to push f_{oth} away from the other subspaces, similar to [17], as shown below:

$$L_{oth} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{n} GR(CE(\mathrm{l}^i_{ak}, ph_{othk}(f_{oth}))), \quad (4)$$

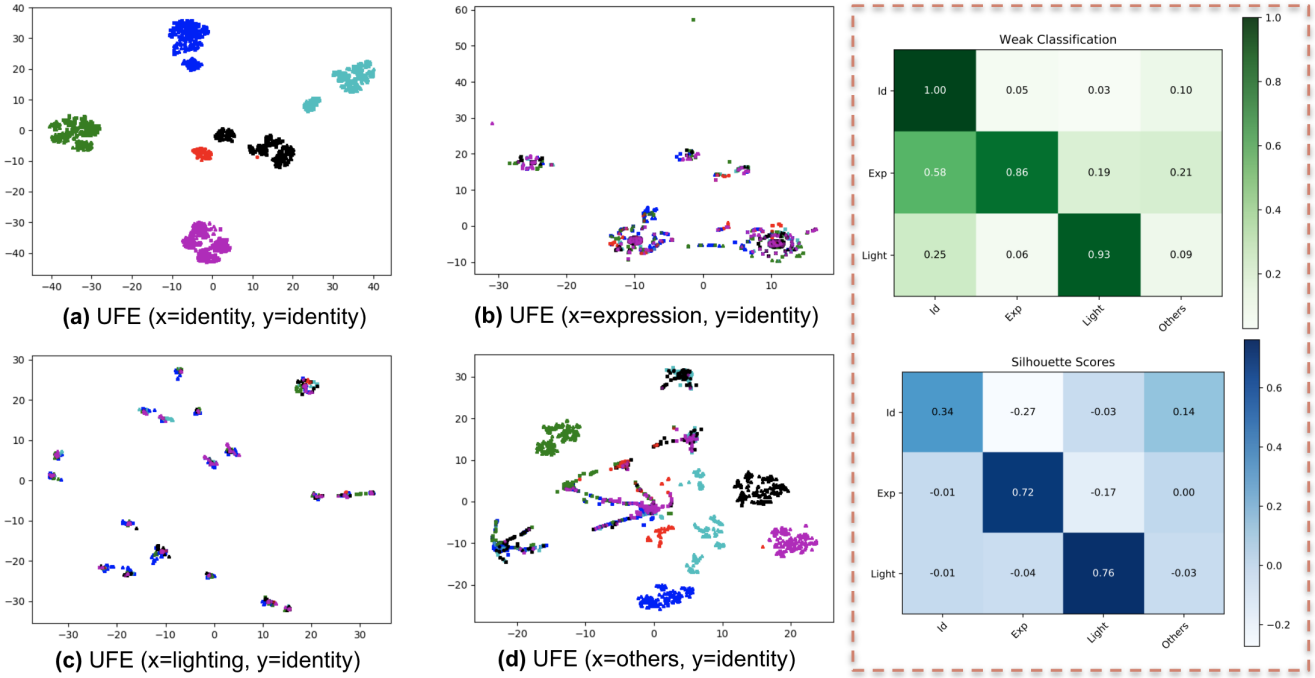

Figure 2. Results on the MultiPIE test set [22]. Using both frontal and posed (yaw=30°) images from 6 subjects, we show the (a) UFE based identity features (x=identity) cluster compactly for identity labels (y=identity), (b) expression features are not discriminative of identity, (c) lighting features are disentangled from identity, and (d) other unlabeled attributed are uncorrelated with identity. Interestingly, the features implicitly cluster (*e.g.* (c) (x=lighting, y=identity) into the corresponding # of classes (20 lighting classes in [22]) signifying well separated representations. The weak classification and silhouette score based confusion matrices on the right panel quantify this separability.

where CE denotes cross entropy, similar to 1, and GR denotes gradient reversal, i.e. shifting the gradient sign and pushes f_{oth} away from other subspaces by essentially penalizing any correct prediction.

5. Disentanglement: While some datasets might have a lot of samples with multi-labeled attributes, most do not. To learn from any unlabeled or semi-labeled samples (*e.g.* when few attribute labels are missing), we include a self-supervision based disentanglement loss L_{dis} in our training. L_{dis} enables learning from additional feature combinations and can further disentangle attribute spaces.

In order to compute such a loss, we take attribute features $[f^i_{a1}, ..., f^i_{an}]$ and $[f^j_{a1}, ..., f^j_{an}]$ from two samples I^i and I^j respectively, that may or may not belong to the same domain. We generate two random combinations of such features f^i_{dis} and f^j_{dis}, by mixing and matching attributes across i and j, and synthesize their decoded images I^i_{dis} and I^j_{dis}. They are again passed through E to generate $E(\mathrm{I}^i_{dis})$ and $E(\mathrm{I}^j_{dis})$ respectively. L_{dis} is then computed as:

$$L_{dis} = \frac{1}{N_p}\sum_{i=1}^{N}\sum_{j=1}^{N}(\left|f^i_{dis} - E(\mathrm{I}^i_{dis})\right| + \left|f^j_{dis} - E(\mathrm{I}^j_{dis})\right|), \quad (5)$$

where N_p are the total number of valid (i, j) pairs in the batch.The loss being an L1 distance measure not only pushes E features to be consistent but also stabilizes D's weights to produce feasible combinations.

The final loss a weighted sum of these objectives:

$$L_{full} = \lambda_{cls}L_{cls} + \lambda_{con}L_{con} + \lambda_{rec}L_{rec} + \lambda_{oth}L_{oth} + \lambda_{dis}L_{dis}, \quad (6)$$

4. Experiments & Results

4.1. Training Data

MultiPIE [22]: We utilize the labels for six facial expression (*Neutral*, *Smile*, *Sadness*, *Surprise*, *Anger*, *Disgust*), 337 subject identities, and 20 directional lighting conditions presented in the dataset. For facial pose variation, we choose images from the 0°and 30°yaw buckets. Around 37k face images are extracted from the dataset in this way.
FFHQ [28]: We gauge the self-supervision ability of the UFE by using the 70k high resolution images from the FFHQ dataset. The dataset contains plenty of identity, expression and pose variations but the images were originally released unlabeled. Annotations for certain attributes (*e.g.* age, pose, gender, eyeglasses) were later shared in [39]. We use this dataset, along with MultiPIE [22], to test the capability of the disentanglement loss 5 in Section 4.7.
In-house: To test the UFE on in-the-wild data, we prepare an in-house dataset, consisting of driver videos captured using both RGB and near infra-red (NIR) cameras placed

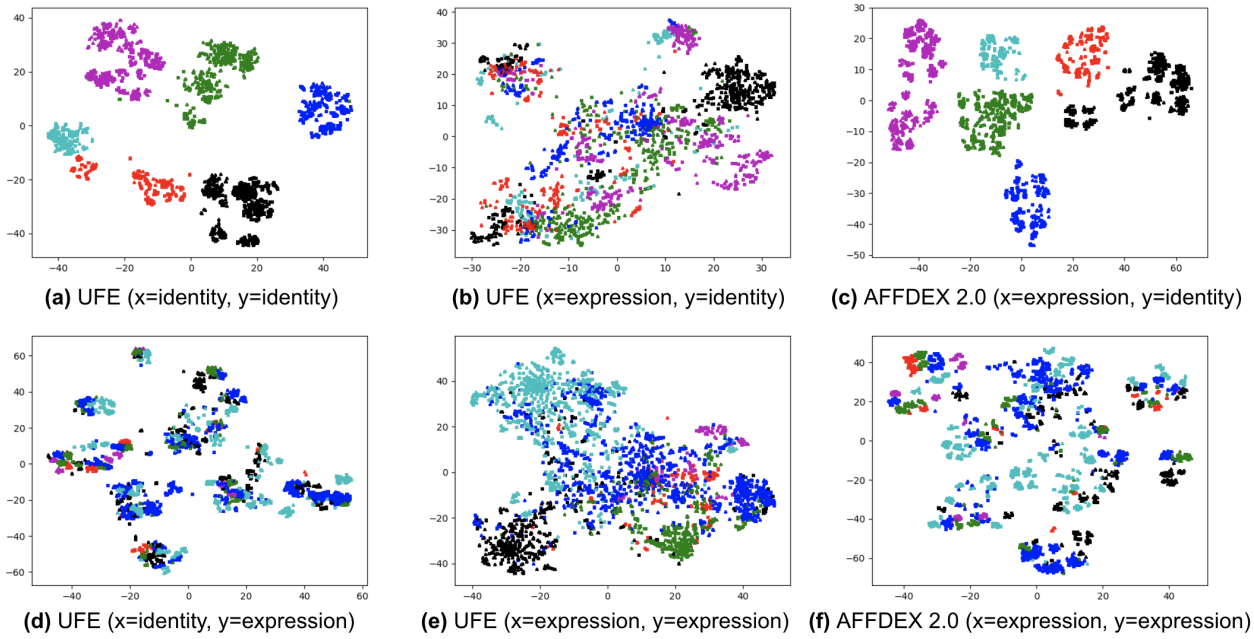

Figure 3. tSNE visualization of the feature space distribution using representations from the proposed UFE and the well-established AFFDEX 2.0 toolkit [5]. In the top row when identity labels are taken as basis for clustering (y=identity), the expression features (x=expression) from the UFE (b) are shown to be disentangled from the identity (a), while the expression features from AFFDEX 2.0 can separate identities well (c). In the bottom row, when y is set to expression labels, UFE based expression features are shown to have higher separability (e) compared to the AFFDEX 2.0 features (f) while maintaining orthogonality from the identity space (d). The test samples are taken from the In-house dataset.

Table 1. Silhouette score based quantitative comparison of UFE with the AFFDEX 2.0 [5] for different (x=feature, y=label) combination.

Model (Camera)	x=id, y=id ↑	x=emo, y=id ↓	x=id, y=emo ↓	x=emo, y=emo ↑
AFFDEX 2.0 [5] (CC)	-	0.122	-	0.011
AFFDEX 2.0 [5] (RVM)	-	0.127	-	0.005
UFE (CC)	0.176	**-0.22**	-0.029	**0.052**
UFE (RVM)	0.212	**-0.104**	-0.028	**0.138**

at rear view mirror (RVM) and center console (CC) locations inside an automotive vehicle. While lighting conditions and other attributes (*e.g.* gender, ethnicity and age) are unlabeled in the dataset, subject identity (121 in total) and expression (6 positive + *Neutral* class) labels are available. The dataset is especially challenging due to non-uniform image quality, noisy annotations and missing labels. Example frames can be seen in Figure 4.

Pre-processing: We extract facial landmarks from each image using the pre-trained FAN from [6], and crop the images to a 128×128 square after aligning the eye centers to the horizontal. For all datasets we randomly pick 90% for training, the rest for testing.

4.2. Implementation Details

We design the encoder E as a sequence of 5 downsampling residual blocks [23] to extract meaningful feature maps from an image. This set of downsampled (flattened) feature maps is split into one dimensional attribute spaces by using the corresponding number of dense layers (+1 for *Others*). Each dense layer has 128 neurons and preserves a uniform range of scale across different subspaces with a sigmoid activation. The decoder D is based on StyleGAN2 [29], consisting of 6 upsampling blocks. The concatenation of the extracted features is fed to each decoder block, serving as the style. The prediction heads ph are 3 dense layers each, with softmax activation in the final layer.

For training the framework on Tensorflow 2 [2], we use a single NVIDIA T4 card with 16GB of memory, and set a batch size of 8. The model is trained for a total of 100 epochs, using the Adam optimizer [31] and initial learning rate of 0.0001. The α (in inverted residuals) and m parameters are both empirically set to 0.9 and 1.0 respectively; the $[\lambda_{cls}, \lambda_{con}, \lambda_{rec}, \lambda_{oth}, \lambda_{dis}]$ coefficients to [1, 1, 100, 1, 0.1] respectively. While training, we slowly introduce the disentanglement loss 5 after a few epochs of UFE training. This allows the E's features to mature in a supervised manner before smoothly transitioning to an additional

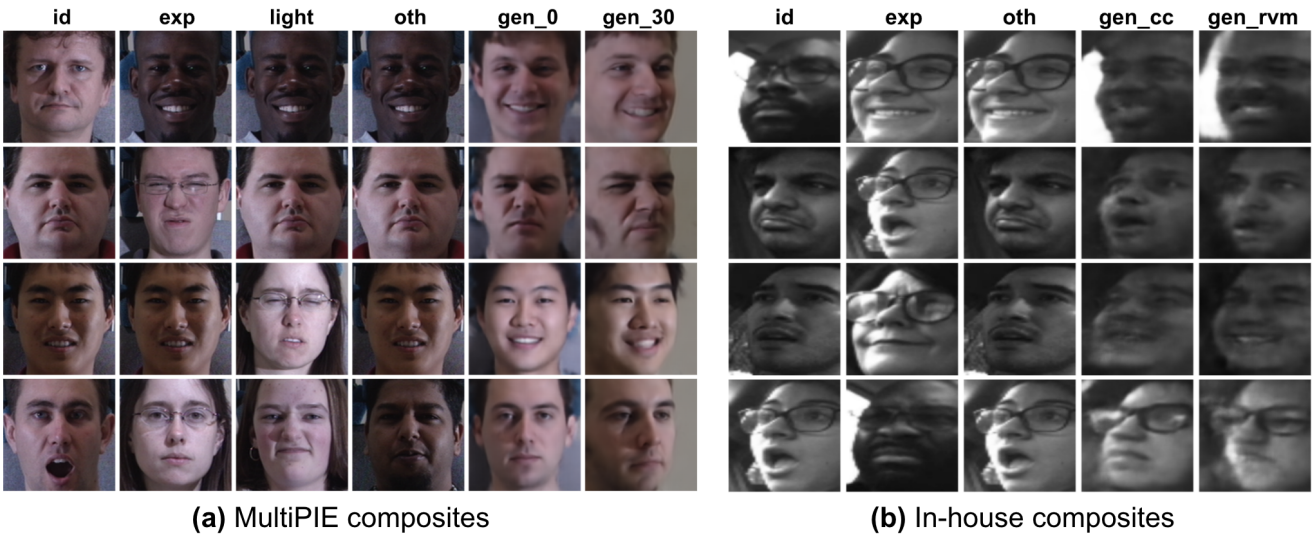

Figure 4. UFE based composite generation: (a) $0^{\circ}$and 30°MultiPIE and (b) CC and RVM camera angle In-house image synthesis. For each row, specific attribute features are extracted from the sample in that column (*e.g.* identity under 'id' column in (a)) and the decoded output can be seen in the last two columns. As the features are invariant to (a) pose and (b) camera angle, the decoder also serves as an implicit reposing function. Compared to MultiPIE, the In-house samples are noisy and low-res, hence the relative decrease in visual fidelity.

self-supervision mode. Post training, only E can be used to extract disentangled features, (E, ph) for any classification task, and (E, D) for new image synthesis.

4.3. Measuring Feature Disentanglement

Weak Classifiers: Since the encoded features should ideally be orthogonal to each other, by virtue of the contrastive objective 2, using features from one attribute as data x (*e.g.* expression) to predict another attribute label y (*e.g.* identity) should not generate strong results for E's representations to be deemed disentangled. Models like CNNs can memorize directly from the training data and project the non-linear relationships directly to generate above-chance scores. Weak classifiers on the other hand (*e.g.* regression) project semi-linear boundaries that do not memorize. Thus, they can be beneficial in quantifying feature disentanglement.

Silhouette Score: To quantitatively estimate the tightness of clusters in each attribute space, we use the silhouette score metric from [1]. It is calculated as $(b-a)/max(a,b)$, where a and b are the mean intra-cluster distance and the mean nearest-cluster distance for each sample respectively. For the same attribute features (*e.g.* expression), a score of 1 suggests perfect tightness while a score of -1 between different attribute features (*e.g.* expression & identity) signifies perfect separation; 0 indicates overlapping clusters.

tSNE: A more visual mode of evaluating E is to take the attribute features and project them into a low dimensional space using tSNE [48]. The spatial positioning of the features in the low dimensional embedding can reveal the extent of disentanglement – similar features (*e.g.* expression) should be clustered together in the same attribute space while mismatching features (*e.g.* expression & identity) and attribute labels should result in imperfect clustering.

4.4. Results on MultiPIE

Using the metrics described in 4.3, we evaluate the efficacy of E trained on the MultiPIE [22] on the testing split of 34 subjects not used in training. For the weak classification task, we utilize logistic regression to the perform the weak classification on the 128-D encoded representations from E. The results are presented in Figure 2 (panel top). As can be seen, the features are quite predictive of their own attribute space and produce the best results (*e.g.* identity data-identity labels = 1.0). In most cases, the features are well-disentangled and mixing data and labels typically result in below chance accuracy, *e.g.* lighting data-identity label $<$ chance labeling for identity (1/34). However, in certain cases the features are not as separated as they can ideally be. Although not strong, the identity features seem to have some relationship of emotion labels with a 0.58 classification accuracy (chance = 1/6).

Using the same snapshot of the trained E and the same test data, we generate the table of silhouette scores as shown in Figure 2 (panel bottom). The intra-attribute numbers (diagonal) are generally in the [0.3, 0.76] range which suggests tight clusters. Additionally, the inter-attribute silhouette scores are all < 0 and therefore suggest relatively looser clustering in such situations. Although the ideal case would be the intra-scores (inter-scores) to be 1 (-1), this can be used as proof of disentanglement in encoded features even in sub-spaces that are correlated from acquisition.

Using samples from 6 test subjects, and their identity as clustering labels, we get the tSNE visualizations, as shown in Figure 2, using the same trained E as before. As can be

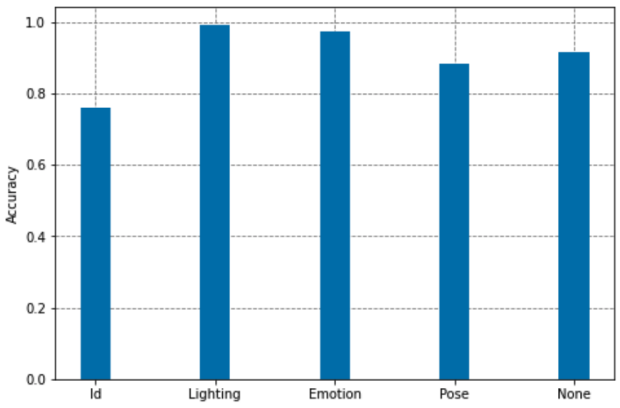

Figure 5. Disentanglement user study: for each manipulated attribute in MultiPIE [22], the mean rater accuracy is shown.

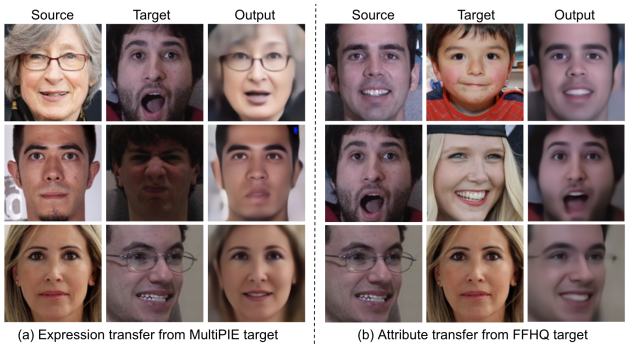

Figure 6. Mixing multi-dataset features: we transfer (a) expression features from MultiPIE [22] target to FFHQ [28] source, (b) age, gender and eyeglasses features from FFHQ [28] to MultiPIE [22] source (1st, 2nd & 3rd rows respectively).

seen, the identity features are tightly clustered with uniform colors across all blobs. However, when embeddings from other attributes are used the clusters are all heterogeneous suggesting that they are not tied to identity. Moreover, the embeddings cluster implicitly based on their spatial location in the embedding space *e.g.* you still get 6 clusters for the 6 emotions and roughly 20 for the 20 lighting conditions.

4.5. Comparison with Traditional Encoding

To gauge the range of the UFE's representation capacity, we compare E with the newly released AFFDEX 2.0 toolkit [5] for expression estimation. While Affdex [35] is an established SDK, used in many downstream applications [20, 34], AFFDEX 2.0 is its enhanced version capable of recognizing facial expressions and emotional states in challenging conditions. The model is specifically trained to estimate facial expressions, and therefore a competitive baseline to measure against for the UFE. To enable an apples-to-apples comparison, we set UFE's E architecture same as that of AFFDEX 2.0 while keeping the sampling and learning strategy same as described in Section 3. We use the In-house dataset for this experiment, and unlike AFFDEX 2.0, the UFE is also trained to learn both the subject identity and facial expressions from the samples. Post training, we analyze emotion estimation and identity-emotion subspace separation from both model's penultimate layer representations in terms of tSNE figures and silhouette scores in Figure 3 and Table 1 respectively.

As expected, the AFFDEX 2.0 features are indeed predictive of the expression classes (Figure 3.f) but not as tightly as the UFE model (Figure 3.e). The AFFDEX 2.0 features are also found to be tightly coupled with identity labels (Figure 3.c) while the same expression features from the UFE are shown to be uncorrelated with any identity (Figure 3.b). The UFE features are shown to be equally effective across the CC & RVM camera angles as well.

4.6. Compositing Features: User Study

Similar to mixing features across domains for tuning the model for self-supervision, we can also generate new samples in the image space by mixing representations in the feature space. By simply upsampling the mixed features gathered from images across different domains, novel views of existing samples can be synthesized by our decoder D. As shown in Figure 4.a, each row represents MultiPIE images from the frontal domain that we take attribute features from and then output the corresponding hallucinations in $0°$(frontal) and $30°$facial yaw. In the first row, every attribute comes from the same image while the identity features are extracted for a (presumed) Caucasian man. As can be seen in the results, the identity traits (gender, age, ethnicity) are preserved while the lighting and expression is mirrored from the dark skinned subject. Additionally, we composite a reposed face ('gen_30') from the same features by simply switching the pose information in D. In this way, we can generate more data. Corresponding results for the In-house samples and attributes are shared in Figure 4.b.

By leveraging these composites, we conduct a user study to evaluate the quality of the perceived feature disentanglement in UFE. For a MultiPIE image in a randomly sampled collection, we pass it through E to get its attribute features. We randomly select an attribute (*e.g.* lighting) and pass a second image with a different label for the selected attribute through E. D is then fed with the selected attribute feature from the second image, while keeping all other attribute features from the first image intact. For a small percent of images, we changed no attributes. For each input-output image pair, we asked users to select the attribute that they perceived to have changed the most. In total, 30 users evaluated 100 image pairs. Figure 5 shows that for all attributes, users are able to identify the correct feature that changed, with a mean accuracy much higher than chance, highlighting the fact that the UFE features are perceivably disentangled.

4.7. Multi-dataset Training

Since the UFE's design and proposed batch sampling enable learning from unseen instances by leveraging L_{dis} (5), we put the model to the ultimate test of learning mul-

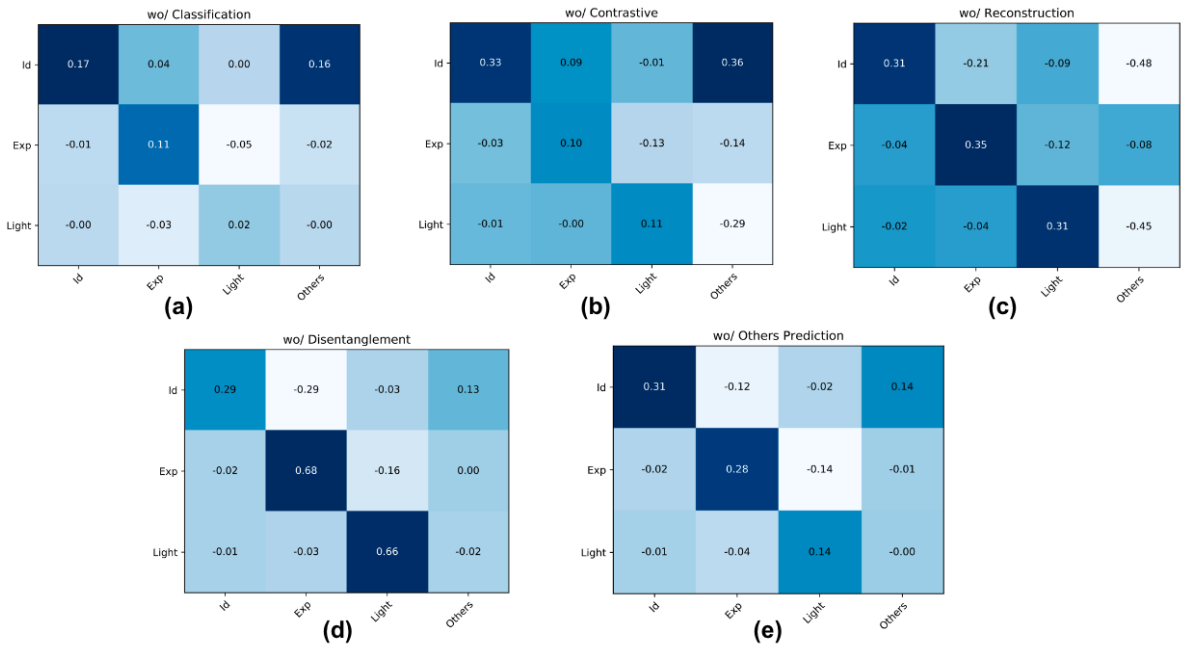

Figure 7. Ablation Study: silhouette scores with UFE features after ablating (a) L_{cls}, (b) L_{con}, (c) L_{rec}, (d) L_{dis}, (e) L_{oth}.

tiple attributes across datasets that have no labels in common. Specifically, we design the model to learn [identity, expression, lighting, pose] labels from MultiPIE [22] and [age, gender, eyeglasses] labels from FFHQ [28]. Alternate (supervised) batches from both datasets are iterated during training, with L_{dis} bridging the two attribute sets by self-supervision. For this particular experiment, we found L_{dis} to benefit from a more mature encoder and hence assigned cyclic annealing [13] based weighing of the loss between [0.0025, 0.1] before stabilizing at 0.1 after three quarters of training. After training, we examine the model's composites by transferring expression features from a MultiPIE target to an FFHQ source and eyeglasses/gender/age from an FFHQ target to MultiPIE, as shown in Figure 6.

We find expression transfer to be seamless across datasets, even when the facial pose of the source and target images are different (Figure 6.a, last row). No other attributes (*e.g.* identity, age, eyeglasses) undergo any noticeable alteration as well. The FFHQ based age and gender transfers work more subtly, especially latter (Figure 6.b, second row), where facial hair and eyebrows are more affected by the transformation. This might be due to the gender features being somewhat correlated to the identity subspace learned from MultiPIE. Eyeglasses can also be removed from bespectacled MultiPIE subjects by pairing with a bare faced FFHQ target (6.b, third row). This validates the UFE's efficacy in learning reliable features across disjoint datasets and its utility as a cross dataset generative tool.

4.8. Ablation Study

Finally, to estimate the contribution of each objective in UFE's multi-attribute representation, we remove L_{cls} (wo/ Classification), L_{con} (wo/ Contrastive), L_{rec} (wo/ Reconstruction), L_{dis} (wo/ Disentanglement) and L_{oth} (wo/ Others Prediction) one at a time. This essentially provides us with 5 different ablated models that we evaluate on the MultiPIE [22] test set. The corresponding confusion matrices for identity, expression and lighting features can be seen in Figure 7. As expected, L_{cls} and L_{con} are key to effective representation learning, with L_{dis} refining the features. Both L_{rec} and L_{oth} further regularize in model training and help in feature visualization and separation.

4.9. Limitations

While the UFE can disentangle attibute sub-spaces, it is able to do so only when such attributes are labeled in at least one dataset used. Additionally, it requires categorical datasets in its current form for training. We plan to scale the model to work with multi-hot attribute vectors in the future.

5. Conclusion

We propose a universal face encoder that simultaneously encodes different facial attributes like identity, expression, lighting from a single image, irrespective of its domain association (*e.g.* facial pose variation or cross dataset training). We evaluate the degree of feature orthogonality in the UFE representation space and show that the encoded features, coupled with dedicated lightweight prediction heads, can be used to predict the class association for each attribute and image. By attaching a style based decoder to the UFE, we hallucinate new images by compositing different encoded attributes from different images, including when sampling from different datasets with disjoint labels.

References

[1] Silhouette score. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html. 1, 6

[2] M. Abadi and et al. Tensorflow: A system for large-scale machine learning. In *USENIX Symposium on Operating Systems Design and Implementation*, 2016. 5

[3] S. Banerjee, A. Joshi, P. Mahajan, S. Bhattacharya, S. Kyal, and T. Mishra. Legan: Disentangled manipulation of directional lighting and facial expressions whilst leveraging human perceptual judgements. In *CVPR Workshops*, 2021. 2

[4] S. Banerjee, A. Joshi, J. Turcot, B. Reimer, and T. Mishra. Driver glance classification in-the-wild: Towards generalization across domains and subjects. In *FG*, 2021. 1, 3

[5] M. Bishay, K. Preston, M. Strafuss, G. Page, J. Turcot, and M. Mavadati. Affdex 2.0: A real-time facial expression analysis toolkit. In *FG*, 2023. 1, 5, 7

[6] A. Bulat and G. Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *IEEE International Conference on Computer Vision*, 2017. 5

[7] M. Chen, A. Radford, R. Child, J. Wu, H. Jun, P. Dhariwal, D. Luan, and I. Sutskever. Generative pretraining from pixels. 2020. 1, 2

[8] S-Y. Chen, F-L. Liu, Y-K. Lai, P.L. Rosin, C. Li, H. Fu, and L. Gao. DeepFaceEditing: Deep generation of face images from sketches. In *SIGGRAPH*, 2021. 2

[9] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. *arXiv:2002.05709*, 2020. 3

[10] Y. Choi, M. Choi, M. Kim, J-W. Ha, S. Kim, and J. Chool. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, 2018. 2

[11] Y. Choi, Y. Uh, J. Yoo, and JW. Ha. Stargan v2: Diverse image synthesis for multiple domains. In *CVPR*, 2020. 2

[12] Y. Deng, J. Yang, D. Chen, F. Wen, and X. Tong. Disentangled and controllable face image generation via 3d imitative-contrastive learning. In *CVPR*, 2020. 1, 2

[13] H. Fu, C. Li, X. Liu, J. Gao, A. Celikyilmaz, and L. Carin. Cyclical annealing schedule: A simple approach to mitigating kl vanishing. In *NAACL*, 2019. 8

[14] A. Gabbay, N. Cohen, and Y. Hoshen. An image is worth more than a thousand words: Towards disentanglement in the wild. In *NeurIPS*, 2021. 2

[15] A. Gabbay and Y. Hoshen. Demystifying inter-class disentanglement. In *ICLR*, 2020. 2

[16] A. Gabbay and Y. Hoshen. Scaling-up disentanglement for image translation. In *ICCV*, 2021. 1, 2

[17] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, and V. Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 17(59):1–35, 2016. 2, 3

[18] L. Gatys, A.S. Ecker, and M. Bethge. A neural algorithm of artistic style. *arXiv:1508.06576*, 2015. 2

[19] T. Gebru, J. Hoffman, and F-F. Li. Fine-grained recognition in the wild: A multi-task domain adaptation approach. In *ICCV*, 2017. 2

[20] G. Gordon, S. Spaulding, JK. Westlund, JJ. Lee, L. Plummer, M. Martinez, M. Das, and C. Breazeal. Affective personalization of a social robot tutor for children's second language skills. In *AAAI*, 2016. 7

[21] J. Gou, B. Yu, S.J. Maybank, and et al. Knowledge distillation: A survey. *IJCV*, 129:1789–1819, 2021. 2

[22] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *IVC*, 28(5):807–813, 2010. 1, 4, 6, 7, 8

[23] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CVPR*, 2016. 5

[24] X. Huang and S. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 2

[25] P. Isola, J-Y. Zhu, T. Zhou, and A.A. Efros. Image-to-image translation with conditional adversarial nets. In *CVPR*, 2017. 2

[26] J. Johnson, A. Alahi, and F-F. Li. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 2

[27] A. Joshi, S. Kyal, S. Banerjee, and T. Mishra. In-the-wild drowsiness detection from facial expressions. In *IV*, 2020. 1

[28] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *arXiv:1812.04948*, 2018. 1, 2, 4, 7, 8

[29] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of StyleGAN. In *CVPR*, 2020. 1, 2, 3, 5

[30] J. Kim, M. Kim, H. Kang, and K.H. Lee. U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. In *ICLR*, 2020. 2

[31] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[32] M. Kowalski, S. J. Garbin, V. Estellers, T. Baltrušaitis, M. Johnson, and J. Shotton. Config: Controllable neural face image generation. In *ECCV*, 2020. 2

[33] J. Li, R. Zhang, F. Ganz, S. Han, and J-Y. Zhu. Anycost gans for interactive image synthesis and editing. In *CVPR*, 2021. 2

[34] A. Lopez-Rincon. Emotion recognition using facial expressions in children using the nao robot. In *CONIELECOMP*, 2019. 7

[35] D. McDuff, A. Mahmoud, M. Mavadati, M. Amr, J. Turcot, and R. el Kaliouby. Affdex sdk: A cross-platform real-time multi-face expression recognition toolkit. In *CHI EA*, 2016. 7

[36] Y. Men, Y. Mao, Y. Jiang, W-Y. Ma, and Z. Lian. Controllable person image synthesis with attribute-decomposed gan. In *CVPR*, 2020. 2

[37] J. Naruniec, L. Helminger, C. Schroers, and RM. Weber. High-resolution neural face swapping for visual effects. In *ESR*, 2020. 3

[38] I. Nigam, P. Tokmakov, and D. Ramanan. Towards latent attribute discovery from triplet similarities. In *ICCV*, 2019. 2, 3

[39] R. Or-El, S. Sengupta, O. Fried, E. Shechtman, and I. Kemelmacher-Shlizerman. Lifespan age transformation synthesis. In *ECCV*, 2020. 2, 3, 4

[40] X. Peng, Z. Huang, X. Sun, and K. Saenko. Domain agnostic learning with disentangled representations. In *ICML*, 2019. 2

[41] V. Prabhu, A. Chandrasekaran, K. Saenko, and J. Hoffman. Active domain adaptation via clustering uncertainty-weighted embeddings. In *ICCV*, 2021. 3

[42] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 3

[43] A. Shoshan, N. Bhonker, I. Kviatkovsky, and G. Medioni. Gan-control: Explicitly controllable gans. In *ICCV*, 2021. 1, 2, 3

[44] G. Song and W. Chai. Collaborative learning for deep neural networks. In *NeurIPS*, 2018. 3

[45] A. Tewari, M. B R, X. Pan, O. Fried, M. Agrawala, and C. Theobalt. Disentangled3d: Learning a 3d generative model with disentangled geometry and appearance from monocular images. In *CVPR*, 2022. 2

[46] S. Tulyakov, M-Y. Liu, X. Yang, and J. Kautz. Mocogan: Decomposing motion and content for video generation. In *CVPR*, 2018. 2

[47] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv:1412.3474*. 2

[48] L.J.P. van der Maaten and G.E. Hinton. Visualizing high-dimensional data using t-sne. *JMLR*, 9:2579–2605, 2008. 1, 6

[49] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *NeurIPS*, 2014. 2

[50] J-Y. Zhu, T. Park, P. Isola, and AA. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 2