The Universal Face Encoder: Learning Disentangled Representations Across Different Attributes (Supplementary Text)

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

Table 1. Encoder E architecture (input size is 128×128×3)

Layer	Filter/Stride	α_{in}/α_{out}	# blocks	# filters
conv1	3×3/2	-	-	8
IRB1	3×3/2	2/1	5	16
IRB2	3×3/2	4/1	1	24
IRB3	3×3/1	2/1	6	32
IRB4	3×3/2	4/1	1	64
IRB5	3×3/1	2/1	2	96
conv2	1×1/1	-	-	128
GAP1	-	-	-	-
fc1	128	-	-	-
fc2	128	-	-	-
fc3	128	-	-	-
fc4	128	-	-	-
fc5	128	-	-	-

1. Detailed Encoder Architecture

Here we describe in detail the architecture of the encoder E module in the UFE framework. As discussed in Section 3 of the main text, E takes as input a 128×128×3 input and passes it through a convolution block before followed by 5 inverted residual blocks [5] with varying stride length, and expansion (α_{in}) and contraction (α_{out}) factors. All blocks use batch normalization and leaky *ReLU* activation. This set of residual maps is sampled through a 2D global average pooling layer, and then fed to (k+1) dense layers for feature extraction for k labeled attribute subspaces, and additional one for unlabeled others. To maintain a [0,1] range of the feature output, we apply the sigmoid function (σ) as activation for each dense layer. The decoder architecture is similar to the StyleGAN2 network shared in [4][1], without progressive growing.

The detailed layers of E are listed in Tables 1. We represent convolution blocks, dense layers, inverted residual blocks and global average pooling as 'conv', 'fc', 'IRB', and 'GAP' respectively in the table.

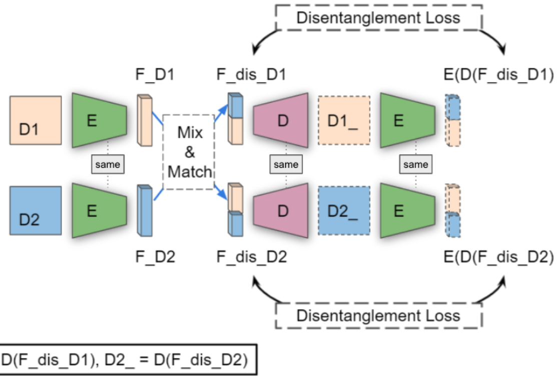

Figure 1. Disentanglement loss (L_{dis}) for cross dataset training: the encoder E is fed $D1$ and $D2$ from the two datasets and generates F_D1 and F_D2 respectively. A combination pair, F_dis_D1 and F_dis_D2, is produced by mixing different attributes from the two sets, and fed to the decoder D to synthesize $D1\_$ and $D2\_$ respectively. The disentanglement loss is computed as L1$(E(D1_), F_dis_D1) +$ L1$(E(D2_), F_dis_D2)$, for all such generated pairs from the two batches.

2. Cross Dataset Training: Leveraging the Disentanglement Loss

For cross dataset training with two datasets having disjoint labels, we apply the disentanglement loss (L_{dis} from the main text) to bridge the representations from the two sample sets. For the supervised training component, separate batches are taken from the two datasets and the UFE is trained using L_{cls}, L_{con} and L_{rec}. To connect the two dataset however we prepare couples of of combination sets of features from sample pairs coming from the two batches.

For example, the identity, lighting, expression and pose (age, gender and eyeglasses) features are labeled in Multi-PIE [2] (FFHQ [3]). However, the missing labels for [age, gender, eyeglasses] can be added by combining them from

[1] Available here: https://github.com/NVlabs/stylegan2

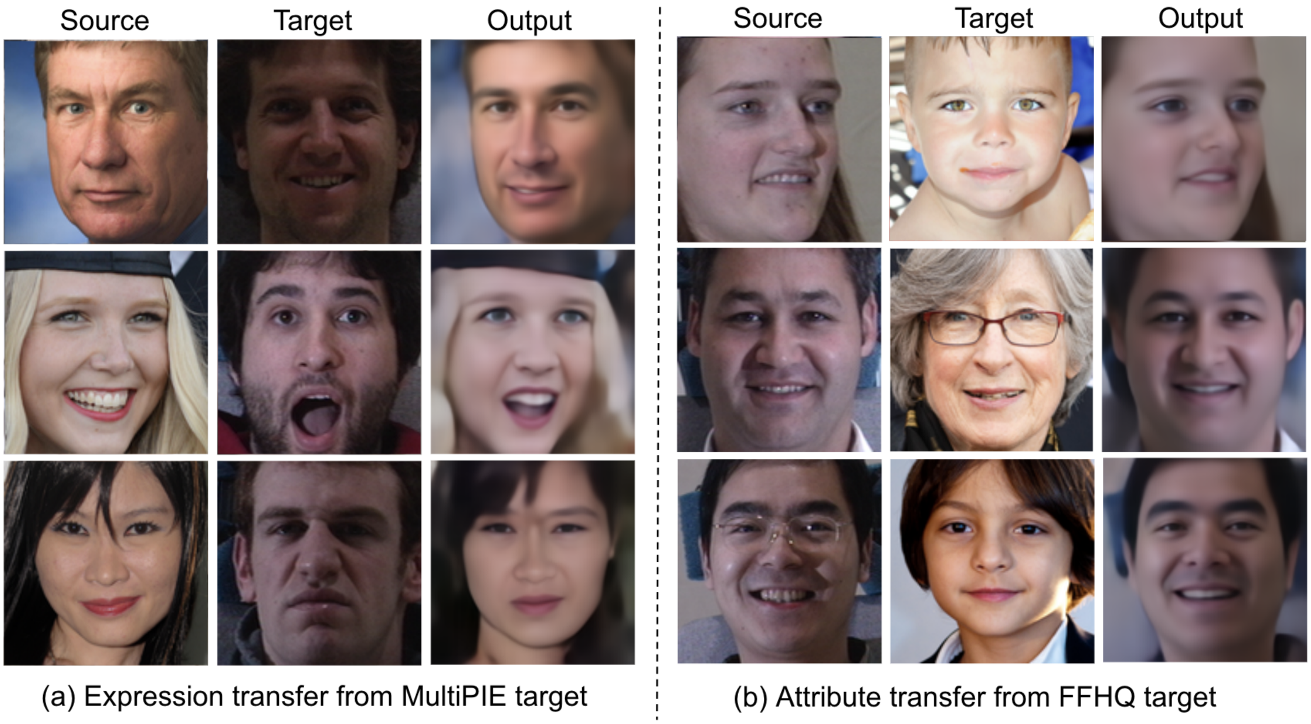
Source
Target
Output
Source
Target
Output
(a) Expression transfer from MultiPIE target
(b) Attribute transfer from FFHQ target

Figure 2. (a) Expression transfer to FFHQ [3] source from MultiPIE [2] target, and (b) (top – bottom) age, gender and eyeglasses transfer to MultiPIE from FFHQ. As shown, the UFE generated features can transfer across image pairs that have subjects looking in the opposite directions. While expression, age and spectacle changes are easily noticeable, we find gender changes to manifest very subtly, *e.g*. by removing facial hair, as they are correlated with identity.

a corresponding FFHQ sample. Similarly, the UFE can learn [identity, pose, expression, lighting] from corresponding MultiPIE features. Additionally, combining with multiple FFHQ samples can generate many possible [age, gender, eyeglasses] variations for the same MultiPIE identity, thus providing a catalog of future training set, allowing for an extensive cross dataset training. The idea is also illustrated in Figure 1.

3. More Cross Dataset Composites: Analyzing Limitations

Here, we present more qualitative results from the cross dataset experiment featuring the MultiPIE [2] and FFHQ [3] datasets. We transfer expression features from the former while any one of age, gender and eyeglasses from the latter. The results can be seen in Figure 2. While expression, age and eyeglass changes are easy to spot, we find gender changes to be very subtle, visible as gradual disappearance of facial hair and eyebrow reshaping. We attribute this to the correlation between the gender and identity subspaces in the training data itself, as they are heavily inter-related but not labeled as a pair in either of the dataset. MultiPIE has identity labels while FFHQ has gender. Such a mismatch makes the model for conservative in its representation and manifests in small pixel changes when decoded.

On analyzing some of the finer annotations within the In-house dataset samples, we found the features for *mustache*, *beard* and *gender* to be correlated with identity labels. Moreover, these three attributes are well correlated with each other as well. This is mainly due to the distribution of these features mainly being skewed to one class (*e.g. mustache* present for *male*) and no positive real sample being present for the opposite instance. Hence, the UFE never learns these cases in training, and consequently manages to shift gender attributes slightly when the identity is fixed.

4. In-house Test Data Distribution

Here, we discuss the non-uniformity in the In-house dataset presented in the main text. The samples collected for the different expression classes are heavily skewed towards 2 – 3 buckets, and even vary across the CC and RVM camera angles. In fact, the *Fear* class has no samples in the CC test set. The training set has a similar distribution and presents a considerable challenge for the UFE to learn meaningful representations. Additionally, since the sparse expressions are generally present for only a few subjects, identity and expression features become more entangled during training, as evident from the AFFDEX 2.0 [1] results from the main pa-

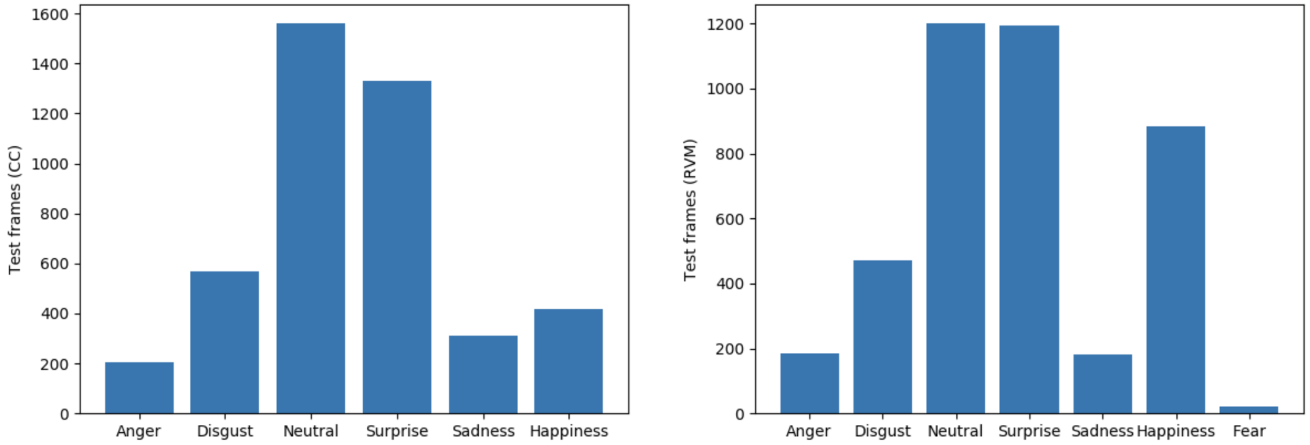

Figure 3. Inhouse test split distribution: the dataset is non-uniform in terms of class wise cardinality, making feature learning difficult for the UFE.

per. Alongside the contrastive objective L_{con} in UFE training, we find the disentanglement loss L_{dis} to be helpful in mitigating this issue and allows oversampling on the under represented samples for synthetic feature creation.

References

[1] 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. 2

[2] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *IVC*, 28(5):807–813, 2010. 1, 2

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

[4] 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

[5] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L-C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, 2018. 1