

# ICface: Interpretable and Controllable Face Reenactment Using GANs

Soumya Tripathy  
Tampere University  
soumya.tripathy@tuni.fi

Juho Kannala  
Aalto University of Technology  
juho.kannala@aalto.fi

Esa Rahtu  
Tampere University  
esa.rahtu@tuni.fi

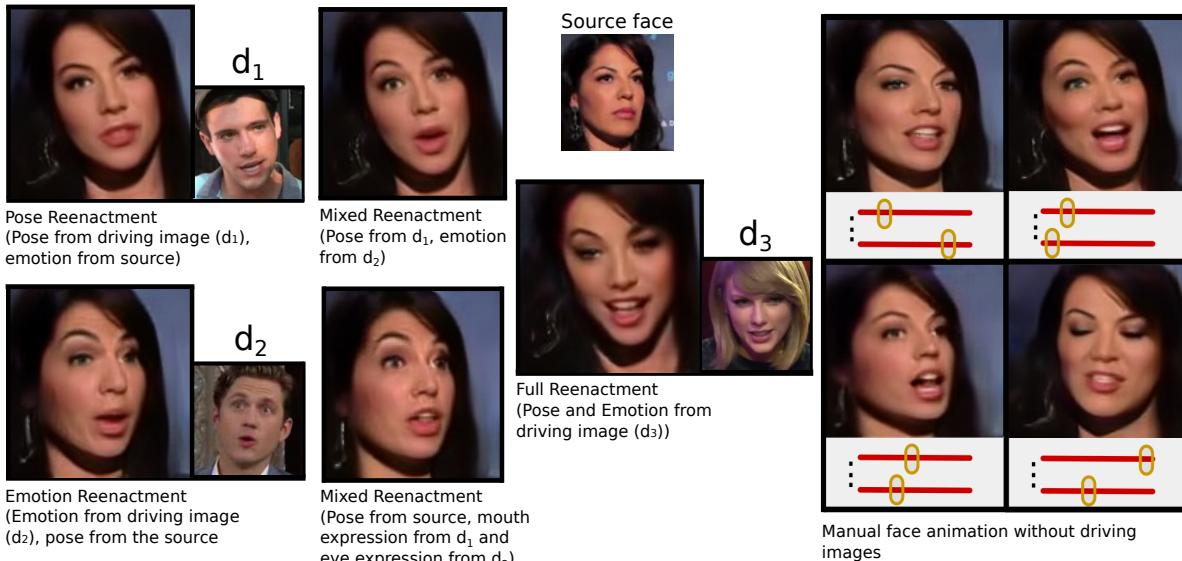


Figure 1: Given a single *source* face image and a set of *driving* attributes, our model is able to generate a high quality facial animation. The driving attributes can be selectively specified by one or more driving face images or controlled via interpretable continuous parameters.

## Abstract

*This paper presents a generic face animator that is able to control the pose and expressions of a given face image. The animation is driven by human interpretable control signals consisting of head pose angles and the Action Unit (AU) values. The control information can be obtained from multiple sources including external driving videos and manual controls. Due to the interpretable nature of the driving signal, one can easily mix the information between multiple sources (e.g. pose from one image and expression from another) and apply selective post-production editing. The proposed face animator is implemented as a two stage neural network model that is learned in self-supervised manner using a large video collection. The proposed Interpretable and Controllable face reenactment network (ICface) is compared to the state-of-the-art neural network based face animation techniques in multiple tasks. The results indicate that ICface produces better visual quality, while being more versatile than most of the comparison methods. The introduced model could provide a lightweight and easy to use tool for multitude of advanced image and video editing tasks. The program code will be publicly available upon the acceptance of the paper.*

## 1. Introduction

The ability to create a realistic animated video from a single face image is a challenging task. It involves both rotating the face in 3D space as well as synthesising detailed deformations caused by the changes in the facial expression. A lightweight and easy-to-use tool for this type of manipulation task would have numerous applications in animation industry, movie post-production, virtual reality, photography technology, video editing and interactive system design, among others.

Several recent works have proposed automated face manipulation techniques. A commonly used procedure is to take a *source* face and a set of desired facial attributes (e.g. pose) as an input and produce a face image depicting the source identity with the desired attributes. The source face is usually specified by one or more example images depicting the selected person. The facial attributes could be presented by categorical variables, continuous parameters or by another face image (referred as a *driving* image) with desired pose and expression.

Traditionally, face manipulation systems fit a detailed 3D face model on the source image(s) that is later used to render the manipulated outputs. If the animation is driven by

another face image, it must also be modelled to extract the necessary control parameters. Although these methods have reported impressive results (see e.g. Face2Face [27],[17]), they require complex 3D face models and considerable efforts to capture all the subtle movements in the face.

Recent works [32, 34] have studied the possibility to bypass the explicit 3D model fitting. Instead, the animation is directly formulated as an end-to-end learning problem, where the necessary model is obtained implicitly using a large data collection. Unfortunately, such implicit model usually lacks interpretability and does not easily allow selective editing or combining driving information from multiple sources. For example, it is not possible to generate an image which has all other attributes from the driving face, except for an extra smile on the face. Another challenge is to obtain expression and pose representation that is independent of the driving face identity. Such disentanglement problem is difficult to solve in a fully unsupervised setup and therefore we often see that the identity specific information of the driving face is "leaking" to the generated output. This may limit the relevant use cases to a few identities or to faces with comparable size and shape.

In this paper, we propose a generative adversarial network (GAN) based system that is able to reenact realistic emotions and head poses for a wide range of source and driving identities. Our approach allows further selective editing of the attributes (e.g. rotating the head, closing the eyelid etc.) to produce novel face movements which were not seen in the original driving face. The proposed method offers extensive human interpretable control for obtaining more versatile and high quality face animation than with the previous approaches. Figure 1 depicts a set of example results generated by manipulating a single source image with different mixtures of driving information.

The proposed face manipulation process consists of two stages: 1) extracting the facial attributes (emotion and pose) from the given driving image, and 2) transferring the obtained attributes to the source image for producing a photorealistic animation. We implement the first step by representing the emotions and facial movements in terms of Action Units (AUs) [10] and head pose angles (pitch, yaw and roll). The AU activations [10] aim at modelling the specific muscle activities and each combination of them can produce different facial expression [10, 25]. Our main motivation is that such attributes are relatively straightforward to extract from any facial image using publicly available software and this representation is fairly independent of the identity specific characteristics of the face.

We formulate the second stage of the face animation process using a conditional generative model on the given source image and facial attribute vector. In order to eliminate the current expression of the source face, we first map the input image to a neutral state representing frontal

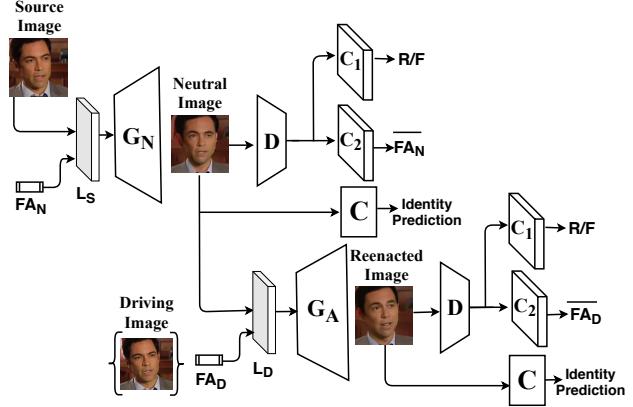


Figure 2: The overall architecture of the proposed model (ICface). During the training, two frames are selected from the same video and denoted as source and driving images. The generator  $G_N$  takes the source image and neutral facial attributes ( $FA_N$ ) as input and produces the source identity with central pose and neutral expression (neutral image). In the second phase, the generator  $G_A$  takes the neutral image and attributes extracted from the driving image ( $FA_D$ ) as an input and produces an image with the source identity and driving image's attributes. The generators are trained using multiple loss functions implemented using the discriminator  $D$  (see Section 3 for details). In addition, since the driving and source images have the same identity, a direct pixel based reconstruction loss can also be utilized. Note that this is assumed to be true only during training and in the test case the identities are likely to be different.

pose and zero AU values. Afterwards, the neutral image is mapped to the final output depicting the desired combination of driving attributes (e.g. obtained from driving faces or defined manually). As a result, we obtain a model called **Interpretable and Controllable face reenactment network (ICface)**.

We make the following three contributions. i) We propose a data driven and GAN based face animation system that is applicable to a large number of source and driving identities. ii) The proposed system is driven by human interpretable control signals obtainable from multiple sources such as external driving videos and manual controls. iii) We demonstrate our system in multiple tasks including face reenactment, facial emotion synthesis, and multi-view image generation from single-view input. The presented results outperform several recent (possibly purpose-built) state-of-the-art works.

## 2. Related work

The proposed approach is mainly related to face manipulation methods using deep neural networks and adversarial generative networks. Therefore, we concentrate on reviewing the most relevant literature under this scope.

## 2.1. Face Manipulation by Generative Networks

Deep neural networks are very popular tools for controlling head pose, facial expressions, eye gaze, etc. Many works [11, 33, 18, 31] approach the problem using supervised paradigm that requires plenty of annotated training samples. While such data is expensive to obtain, the recent literature proposes several unsupervised and self-supervised alternatives [37, 7, 8, 20].

In [26], the face editing was approached by decomposing the face into a texture image and a deformation field. After decomposition, the deformation field could be manipulated to obtain desired facial expression, pose, etc. However, this is a difficult task, partially because the field is highly dependent on the identity of the person. Therefore, it would be hard to transfer attributes from another face image.

Finally, X2Face [32] proposes a generalized facial reenactor that is able to control the source face using driving video, audio, or pose parameters. The transferred facial features were automatically learned from the training data and thus lack clear visual interpretation (e.g. close eyes or smile). The approach may also leak some identity specific information from the driving frames to the output. X2Face seems to work best if the driving and source images are from the same person.

## 2.2. Face Manipulation with GANs

The conditional variant of the Generative Adversarial Network (GAN) [21, 13] have received plenty of attention in image to image domain translation with paired data [14], unpaired data [38], or even both [30]. Similar GAN based approaches are widely used for facial attribute manipulation in many supervised and unsupervised settings [9, 6, 28, 24]. The most common approach is to condition the generator on discrete attributes such as blond or black hair, happy or angry, glasses or no glasses and so on. A recent work [25] proposed a method called GANimation that was capable of generating a wide range of continuous emotions and expressions for a given face image. They utilized the well known concept of action units (AUs) as a conditioning vector and obtained very appealing results. Similar results are achieved on portrait face images in [12, 1]. Unfortunately, unlike our method, these approaches are not suitable for full reenactment, where the head pose has to be modified. Moreover, they add the new attribute directly on to the existing expression of the source image which can be problematic to handle.

In addition to the facial expression manipulation, the GANs are applied to the full face reenactment task. For instance, the CycleGAN [36] could be utilised to transform expressions and pose between image pair (see examples in [32]). Similarly, the ReenactGAN [34] is able to perform face reenactment, but only for a limited number of source identities. In contrast, our GAN based approach generalizes

to a large number of source identities and driving videos. Furthermore, our method is based on interpretable and continuous attributes that are extracted from the driving video. The flexible interface allows user to easily mix the attributes from multiple driving videos and edit them manually at will. In the extreme case, the attributes could be defined without any driving video at all.

## 3. Method

The goal of our method is to animate a given *source* face in accordance to the facial attribute vector ( $FA_D = [\mathbf{p}^T, \mathbf{a}^T]^T$ ) that consists of the head pose parameters  $\mathbf{p}$  and action unit (AU) activations  $\mathbf{a}$ . More specifically, the head pose is determined by three angles (pitch, yaw, and roll) and the AUs represent the activations of 17 facial muscles [10]. In total, the attribute vector consists of 20 values determining the pose and the expression of a face. In the following, we will briefly outline the workflow of our method. The subsequent sections and Figure 2 provide further details of the architecture and training procedure. The specific implementation details are found in the supplementary material.

In the first stage, we concatenated the input image (size  $W \times H \times 3$ ) with the neutral facial parameters  $FA_N = [\mathbf{p}_N, \mathbf{0}]$ , where  $\mathbf{p}_N$  refers to the central pose. This is done by first spatially replicating the attribute vector and then channel-wise concatenating the replicated attributes ( $W \times H \times 20$ ) with the input image. The resulting representation  $L_S$  ( $W \times H \times 23$ ) is subsequently fed to the neutralisation network  $G_N$  that aims at producing a frontal face image ( $W \times H \times 3$ ) depicting the source identity with neutral facial expression.

In the second stage, we concatenated the obtained neutral (source) face image with the *driving* attribute vector  $FA_D$  that determines the desired output pose and AU values. The concatenation is done in similar fashion as in the first stage. In our experiments, we used OpenFace [3, 4] to extract the pose and AUs when necessary. The concatenated result  $L_D$  is passed to the generator network  $G_A$  that produces the final animated output ( $W \times H \times 3$ ) depicting the original *source* identity with the desired facial attributes  $FA_D$ .

## 3.1. Architecture

Our model consists of four different sub-networks: Neutraliser, generator, discriminator and identity preserving network. Their structures are briefly explained as follows:

**Neutralizer ( $G_N$ ) :** The neutralizer is a generator network that transforms the input representation  $L_s$  into a canonical face that depicts the same identity as the input and has central pose with neutral expression. The architecture of the  $G_N$  network consists of strided convolution, residual blocks and deconvolution layers. The overall structure is inspired by the generator architecture of CycleGAN [36].

**Generator ( $G_A$ ) :** The generator network transforms input representation  $L_D$  of the neutral face into the final reenacted output image. The output image is expected to depict the *source* identity with pose and expression defined by the *driving* attribute vector  $FA_D$ . The architecture of the  $G_A$  network is similar to that of  $G_N$ .

**Discriminator ( $D$ ) :** The discriminator network performs two tasks simultaneously: i) it evaluates the realism of the neutral and reenacted images through  $C_1$ ; ii) it predicts the facial attributes ( $FA_N$  and  $FA_D$ ) through  $C_2$ . The blocks  $C_1$  and  $C_2$  consist of convolution block with sigmoid activation. The overall architecture of  $D$  is similar to the Patch-GANs [36] consisting of strided convolution and activation layers. The same discriminator with identical weights is used for  $G_N$  and  $G_A$ .

**Identity Preserving Network ( $C$ ) :** We have used a pre-trained network called LightCNN [35] as  $C$  and kept the weights fixed for the whole training process. It provides the identity features for both generated and source faces which are used in the training as identity preserving loss.

### 3.2. Training the Model

Following [32], we train our model using VoxCeleb [23] dataset that contains short clips extracted from interview videos. Furthermore, Nagrani et al. [22] provide face detections and face tracks for this dataset, which is utilised in this work. As in [32], we extract two frames from the same face track and feed one of them to our model as a *source* image. Then, we extract the pose and AUs from the second frame and feed them into our model as *driving* attributes  $FA_D$ . Since both frames originate from the same face track and depict the same identity, the output of our model should be identical to the second frame and it can be treated as a pixel-wise ground truth in the training procedure. All the loss functions are described in the following paragraphs.

**Adversarial Loss( $\mathcal{L}_{adv}$ ) :** The adversarial loss is a crucial component for obtaining the photorealistic output images. The generator  $G_N$  maps the feature representation  $L_S$  into domain of real images  $X$ . Now if  $x \in X$  is a sample from the training set of real images, then the discriminator has to distinguish between  $x$  and  $G_N(L_S)$ . The corresponding loss function can be expressed as:

$$\mathcal{L}_{adv}(G_N, D) = \mathbb{E}_x [\log D(x)] + \mathbb{E}_{L_S} [\log(1 - D(G_N(L_S)))] \quad (1)$$

Similar loss function can also be formulated for  $G_A$  and  $D$  and it would be represented as  $\mathcal{L}_{adv}((G_A, D))$ .

**Facial attribute reconstruction loss( $\mathcal{L}_{FA}$ ) :** The generators  $G_N$  and  $G_A$  are aiming at producing photorealistic face images, but they need to do this in accordance with the facial attribute vectors  $FA_N$  and  $FA_D$ , respective. To this end, we extend the discriminator to regress the facial attributes from the generated images and compare them to the

given target values. The corresponding loss function  $\mathcal{L}_{FA}$  is expressed as:

$$\begin{aligned} \mathcal{L}_{FA} = & \mathbb{E}_x [\|C2(D(x)) - FA_D\|_2^2] + \mathbb{E}_{L_D} [\|C2(D(G_A(L_D))) \\ & - FA_D\|_2^2] + \mathbb{E}_{L_S} [\|C2(D(G_N(L_S))) - FA_N\|_2^2] \end{aligned} \quad (2)$$

where  $x \in X$  is the driving image with attributes  $FA_D$ .

**Identity classification loss( $\mathcal{L}_I$ ) :** The goal of our system is to generate an output image retaining the identity of the source person. To encourage this, we have used a pretrained LightCNN [35] network to compare the features of the generated image ( $g$ ) and source image ( $s$ ). Specifically we have compared the features of last pooling layer ( $f^p$ ) and fully connected layer's ( $f^{fc}$ ) of LightCNN as follows:

$$\mathcal{L}_I = \|C(f^p(s)) - C(f^p(g))\|_1 + \|C(f^{fc}(s)) - C(f^{fc}(g))\|_1 \quad (3)$$

**Reconstruction loss( $\mathcal{L}_R$ ) :** Due to the specific training procedure described above, we have access to the pixel-wise ground truth of the output. We take advantage of this by applying L1 loss between the output and the ground truth. Furthermore, we stabilize the training of the  $G_N$  by using generated images from  $G_A$  with neutral attributes as a pseudo ground truth. The corresponding loss function is defined as

$$\mathcal{L}_R = \mathbb{E}_x [\|x - G_A(L_D)\|_1] + \mathbb{E}_{L_S} [\|G_N(L_S) - G_A(L_S)\|_1] \quad (4)$$

**The complete loss function:** The full objective function of the proposed model is obtained as a weighted combination of the individual loss functions defined above. The corresponding full loss function, with  $\lambda_i$  as regularization parameters, is expressed as

$$\mathcal{L} = \mathcal{L}_{adv}(G_N, D) + \mathcal{L}_{adv}(G_A, D) + \lambda_1 \mathcal{L}_{FA} + \lambda_2 \mathcal{L}_I + \lambda_3 \mathcal{L}_R \quad (5)$$

## 4. Experiments

In all the following experiments, we use a single ICface model that is trained using the publicly available VoxCeleb video dataset [23]. The video frames are extracted using the preprocessing techniques presented in [22] and resized to  $128 \times 128$  for further processing. We used 75% of the data for training and the rest for validation and testing. Each component of  $FA$  is normalized to the interval  $[0, 1]$ . The neutral attribute vector  $FA_N$  contains central head pose parameters  $[0.5, 0.5, 0.5]$  and zeros for the AUs. More architectural and training details are provided in the supplementary material.

### 4.1. Face Reenactment

In face reenactment, the task is to generate a realistic face image depicting a given *source* person with the same pose and expression as in a given *driving* face image. The *source* identity is usually specified by one or more example face images (one in our case). Figure 3 illustrates several reenactment outputs using different *source* and *driving* images.



Figure 3: Qualitative results for the face reenactment on VoxCeleb [23] test set. The images illustrate the reenactment result for four different source identities. For each source, the results correspond to: ICface (first row), DAE [26] (second row), X2Face [32] (third row), and the driving frames (last row). The performance differences are best illustrated in cases with large differences in pose and face shape between source and driving frames.

We compare our results with two recent methods: X2Face [32] and DAE [26]. We further refer to the supplementary material for additional reenactment examples.

The results indicate that our model is able to retain the *source* identity relatively well. Moreover, we see that the facial expression of the *driving* face is reproduced with decent accuracy. Recall that our model transfers the pose and expression information via three angles and 17 AU activations. Our hypothesis was that these values are adequate at presenting the necessary facial attributes. This assumption is further supported by the result in Figure 3. Another important aspect in using pose angles and AUs, was the fact that they are independent of the identity. For this reason, the *driving* identity is not “leaking” to the output face (see also the results of the other experiments). Moreover, our model neutralises the *source* image from its prior pose and expression which helps in reenacting new facial movements from *driving* images. We assess this further in Section 4.3.

*Comparison to X2Face [32]:* X2Face disentangles the

identity (texture and shape of the face) and facial movements (expressions and head pose) using the **Embedding** and **Driving** networks, respectively. It is trained unsupervisedly, which make it difficult to prevent all movement and identity leakages through the respective networks. These type of common artefacts are visible in some of the examples in Figure 3. We further note that the X2Face results are produced using three *source* images unlike ICface with single source. Additionally, the adversarial training of our system seems to lead to more vivid and sharp results than X2Face. To further validate this, we quantitatively compare the quality of generated images by both the methods. The quality is assessed in terms of image degradations like smoothing, motion blur, etc. by using two pre-trained networks, CNNIQA [16] and DeepIQA [5], proposed for Non-Reference (NR) and Full-Reference (FR) Image Quality Assessment (IQA). For the FR-IQA, the source image is used as the reference image. The mean quality scores over all the test images for X2face and ICface are presented in Table 1.

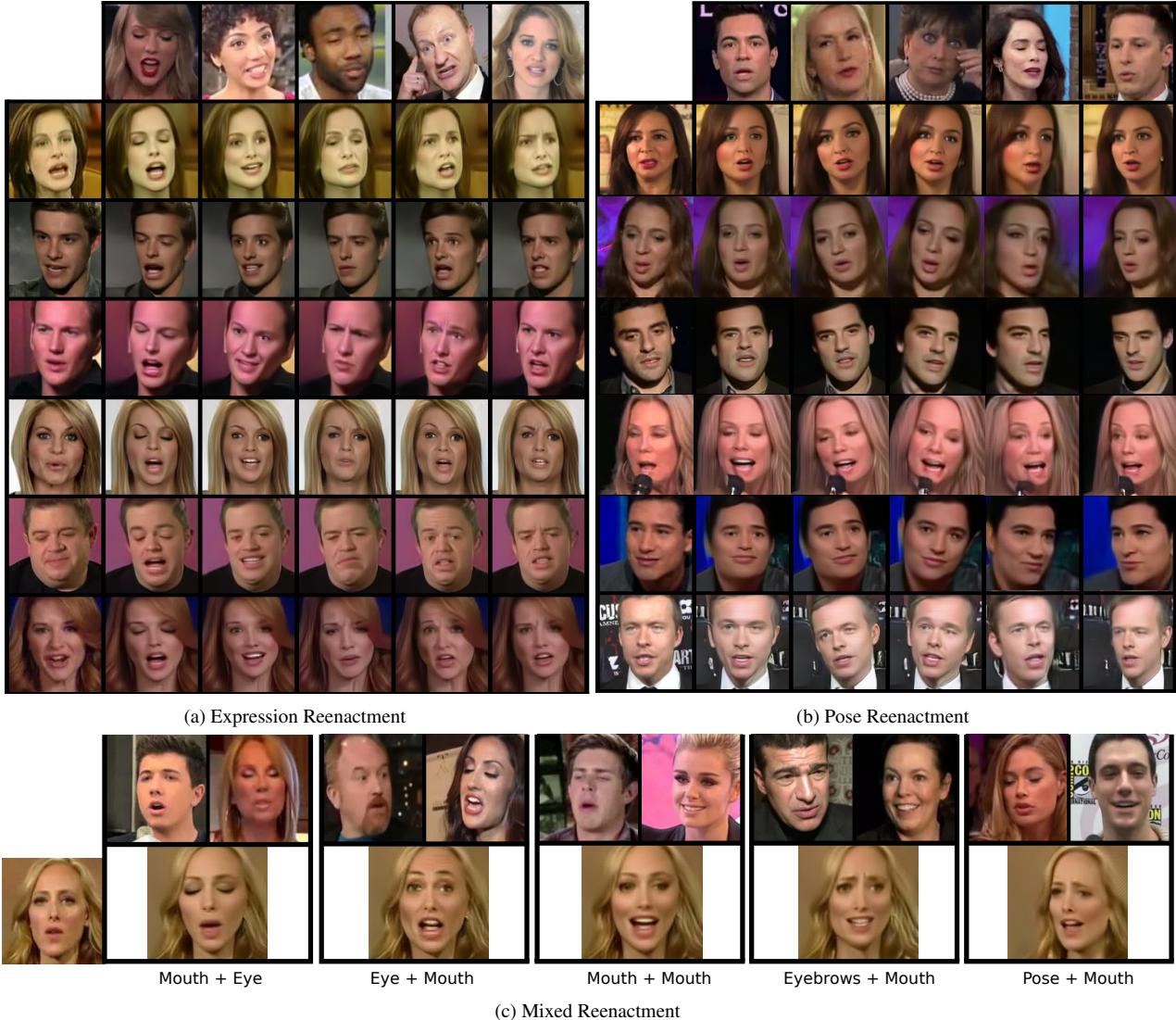


Figure 4: Results for selective editing of facial attributes in face reenactment. (a-b) illustrate emotion and pose reenactment for various source images (extreme left column) and driving images (top row). (c) illustrates mixed reenactment by combining various attributes from source (extreme left) and two driving images (top row). The proposed method produces good quality results and provides control over the animation process, unlike other methods. More results are in the supplementary material.

Method	CNNIQA(NR)	DeepIQA (FR)
X2face / ICface	28.89 / <b>25.02</b>	39.49 / <b>33.08</b>

Table 1: Image Quality Assesment scores (Lower is better.)

The lower scores for ICface signify that the reenacted faces are less distorted than X2Face in comaprision with an ideal imaging model or given reference image.

*Comparison to DAE [26]:* DAE proposed a special autoencoder architecture to disentangle the appearance and facial movements into texture image and deformation fields. We trained their model on VoxCeleb dataset [23] using the publicly available codes from the original authors. For reenactment, we first decomposed both the *source* and *driving*

images into corresponding appearances and deformations. Then we reconstructed the output face using the appearance of *source* image and the deformation of *driving* image. The obtained results are presented in Figure 3. The DAE often fails to transfer the head poses and identity accurately. The head pose related artefact is best observed when the pose difference between the *source* and *driving* is large. These challenges might be related to the fact that the deformation field is not free from the identity specific characteristics.

## 4.2. Controllable Face Reenactment

The pure face reenactment animates the *source* face by copying the pose and expression from the *driving* face. In practice, it might be difficult to obtain any single *driving*



Figure 5: Results for multi-view face generation from a single view. In each block, the first row corresponds to CR-GAN [28] and the second row corresponds ICface. It is to be noted that each block contains the same identity with different crop sizes as both methods are trained with different image crops. Proposed architecture produces semantically consistent facial rotations by preserving the identity and expressions better than the CR-GAN [28]. The last two rows correspond to multi-view images generated from ICface by varying pitch and roll respectively which is not possible in CR-GAN [28].

Measures	X2face	GANimation	DAE	ICface (No-Neutralizer)	ICface
Accuracy(%)	64.76	61.77	59.86	59.88	62.86
F-score	0.4476	0.3944	0.3734	0.3759	0.4185

Table 2: Comparison of Action units classification measures (Higher is better)

image that contain all the desired attributes. The challenge is further emphasised if one aims at generating an animated video. Moreover, even if one could record the proper *driving* frames, one may still wish to perform post-production editing on the result. This type of editing is hard to implement with previous methods like X2Face [32] and DAE [26] since the facial representation is learned implicitly and it lacks a clear interpretability. Instead, the head pose angles and AUs, utilised in our approach, provide human interpretable and easy-to-use interface for selective editing. Moreover, this presentation allows to mix attributes from different *driving* images in controlled way. Figures 1 and 4 illustrate multiple examples, where we have mixed the *driving* information from different sources. The supplementary material contains further example cases.

### 4.3. Facial expression Manipulation

In this experiment, we concentrate on assessing how the proposed model can be used to transfer only the expression from the *driving* face to the *source* face. We compare our results to GANimation [25] that is a purpose-built method for manipulating only the facial expression (i.e. it is not able to modify the pose). Figure 6 illustrates example results for the

proposed ICface and the GANimation. Note that in Figure 6 the head pose of ICface is kept constant as that of the source image only for better illustration. The latter method seems to have challenges when the *source* face has an intense prior expression that is different from the *driving* expression. In contrast, our model neutralised the *source* face before applying the *driving* attributes. This approach seems to lead in better performance in the cases where *source* has intense expression. We have further validated this qualitatively by calculating the facial action consistency. This is obtained by comparing the presence and absence of AUs in both the driving and the reenacted images. We used a pre-trained network [2] to predict the presence of 17 action units in a yes or no fashion for the images generated by X2face, DAE, ICface and GANimation. Then we calculated the balanced accuracy [15] and F-score [15] of detecting the presence and absence of AUs in the generated images in comparison to those of the driver images. These measures are chosen as the absent AU counts are significantly more than active AUs. The values are listed in Table 2 and ICface achieves higher or comparable scores than others even after using only 17 parameters for representing all the facial activities.

### 4.4. Multiview Face generation

Another interesting aspect in face manipulation is the ability to change the viewpoint of a given *source* face. Previous works have studied this as an independent problem. We compare the performance of our model in this task with respect to the recently proposed Complete-Representation GAN (CR-GAN) [28] method. The CR-GAN is a purpose-

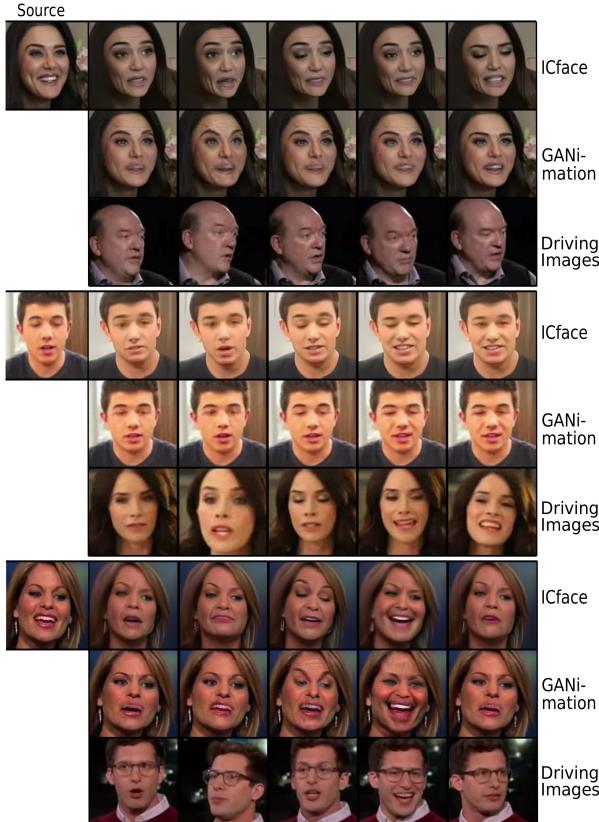


Figure 6: The results for manipulating emotion in the face images. For each source image, the first row is generated using ICface, the second row using GANimation [25] and the third row contains the driving images. As ICface first neutralises the source image, it is evident that it produces better emotion reenactment when the source has initial expressions (first and third row).

built method for producing multi-view face images from a single view input. The model is further restricted to consider only the yaw angle of the face. The results in Figure 5 were obtained using the CR-GAN implementation from the original authors [19]. Their implementation was trained on CelebA [19] dataset, and therefore we used CelebA [19] test to produce these examples. We note that we did not re-train or fine-tune our model on CelebA. The results indicate that our model is able to perform facial rotation with relatively small semantic distortions. Moreover, last two rows of Figure 5 depict rotation along pitch and roll axis which is not achievable with the CR-GAN. We believe that our two-stage based approach (*input* → *neutral* → *target*) is well suited for this type of rotation tasks.

#### 4.5. Identity disentanglement from face attributes

Finally, in Figure 7, we demonstrate the performance of our neutraliser network. The neutraliser was trained to produce a template face from the single *source* image with frontal pose and no expression. We believe that the effective



Figure 7: The results for generating neutral face from a single source image. The proposed method produces good image quality even with extreme head poses (third row).

neutralisation of the input face is one of the key reasons why our system produces high quality results in multiple tasks. To verify that, We repeated the experiments by removing neutralization part from our model and the performance decreases as given in Table 2 (Fourth column). Figure 7 also illustrates the neutral images (or texture image) produced by the baseline methods<sup>1</sup>. One of these is DR-GAN [29] that is a purpose-built face frontalisation method (i.e. it does not change the expression). The ICface successfully neutralises the face while keeping the identity intact even if the *source* has extreme pose and expression.

## 5. Conclusion

In this paper, we proposed a generic face animator that is able to control the pose and expression of a given face image. The animation was controlled using human interpretable attributes consisting of head pose angles and action unit activations. The selected attributes enabled selective manual editing as well as mixing the control signal from several different sources (e.g. multiple *driving* frames). One of the key ideas in our approach was to transform the *source* face into a canonical presentation that acts as a template for the subsequent animation steps. Our model was demonstrated in numerous face animation tasks including face reenactment, selective expression manipulation, 3D face rotation, and face frontalisation. In the experiments, the proposed ICface model was able to produce high quality results for a variety of different *source* and *driving* identities. The future work includes further increasing the resolution of the output images and further improving the performance with extreme poses having a few training samples

<sup>1</sup>It is not a direct comparison to X2face [32] as it never aims at generating a neutral image. It only illustrates the intermediate results for X2face.

## References

- [1] H. Averbuch-Elor, D. Cohen-Or, J. Kopf, and M. F. Cohen. Bringing portraits to life. *ACM Trans. Graph.*, 36(6):196:1–196:13, Nov. 2017.
- [2] T. Baltrušaitis, M. Mahmoud, and P. Robinson. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 6, pages 1–6. IEEE, 2015.
- [3] T. Baltrušaitis, A. Zadeh, Y. Lim, and L. Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2018)*, pages 59–66, Los Alamitos, CA, USA, may 2018. IEEE Computer Society.
- [4] T. Baltrušaitis, M. Mahmoud, and P. Robinson. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 06, pages 1–6, May 2015.
- [5] S. Bosse, D. Maniry, K. Miller, T. Wiegand, and W. Samek. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on Image Processing*, 27(1):206–219, Jan 2018.
- [6] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS’16, pages 2180–2188, USA, 2016. Curran Associates Inc.
- [7] Y.-C. Chen, H. Lin, M. Shu, R. Li, X. Tao, X. Shen, Y. Ye, and J. Jia. Facelet-bank for fast portrait manipulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3541–3549, 2018.
- [8] J. S. Chung, A. Jamaludin, and A. Zisserman. You said that? In *British Machine Vision Conference*, 2017.
- [9] H. Ding, K. Sriharan, and R. Chellappa. Exprgan: Facial expression editing with controllable expression intensity. *AAAI*, 2018.
- [10] P. Ekman and W. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto, 1978.
- [11] Y. Ganin, D. Kononenko, D. Sungatullina, and V. Lempitsky. Deepwarp: Photorealistic image resynthesis for gaze manipulation. In *European Conference on Computer Vision*, pages 311–326. Springer, 2016.
- [12] J. Geng, T. Shao, Y. Zheng, Y. Weng, and K. Zhou. Warp-guided gans for single-photo facial animation. *ACM Trans. Graph.*, 37(6):231:1–231:12, Dec. 2018.
- [13] 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.
- [14] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, 2017.
- [15] Y. Jiao and P. Du. Performance measures in evaluating machine learning based bioinformatics predictors for classifications. *Quantitative Biology*, 4(4):320–330, Dec 2016.
- [16] L. Kang, P. Ye, Y. Li, and D. Doermann. Convolutional neural networks for no-reference image quality assessment. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1733–1740, June 2014.
- [17] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, N. Nießner, P. Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt. Deep Video Portraits. *ACM Transactions on Graphics 2018 (TOG)*, 2018.
- [18] T. D. Kulkarni, W. F. Whitney, P. Kohli, and J. Tenenbaum. Deep convolutional inverse graphics network. In *Advances in neural information processing systems*, pages 2539–2547, 2015.
- [19] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.
- [20] R. Mechrez, I. Talmi, and L. Zelnik-Manor. The contextual loss for image transformation with non-aligned data. In *ECCV*, 2018.
- [21] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.
- [22] A. Nagrani, S. Albanie, and A. Zisserman. Seeing voices and hearing faces: Cross-modal biometric matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- [23] A. Nagrani, J. S. Chung, and A. Zisserman. Voxceleb: a large-scale speaker identification dataset. In *INTERSPEECH*, 2017.
- [24] G. Perarnau, J. van de Weijer, B. Raducanu, and J. M. Álvarez. Invertible Conditional GANs for image editing. In *NIPS Workshop on Adversarial Training*, 2016.
- [25] A. Pumarola, A. Agudo, A. Martinez, A. Sanfeliu, and F. Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.
- [26] Z. Shu, M. Sahasrabudhe, R. Alp Güler, D. Samaras, N. Paragios, and I. Kokkinos. Deforming autoencoders: Unsupervised disentangling of shape and appearance. In *Computer Vision – ECCV 2018*, pages 664–680. Springer International Publishing, 2018.
- [27] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. Face2Face: Real-time Face Capture and Reenactment of RGB Videos. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2016.
- [28] Y. Tian, X. Peng, L. Zhao, S. Zhang, and D. N. Metaxas. Cr gan: Learning complete representations for multi-view generation. In *IJCAI*, 2018.
- [29] L. Tran, X. Yin, and X. Liu. Disentangled representation learning gan for pose-invariant face recognition. In *In Proceeding of IEEE Computer Vision and Pattern Recognition*, Honolulu, HI, July 2017.
- [30] S. Tripathy, J. Kannala, and E. Rahtu. Learning image-to-image translation using paired and unpaired training samples. In *ACCV*, 2018.

- [31] P. Upchurch, J. R. Gardner, G. Pleiss, R. Pless, N. Snavely, K. Bala, and K. Q. Weinberger. Deep feature interpolation for image content changes. In *CVPR*, pages 6090–6099, 2017.
- [32] O. Wiles, A. Koepke, and A. Zisserman. X2face: A network for controlling face generation by using images, audio, and pose codes. In *European Conference on Computer Vision*, 2018.
- [33] D. E. Worrall, S. J. Garbin, D. Turmukhambetov, and G. J. Brostow. Interpretable transformations with encoder-decoder networks. In *The IEEE International Conference on Computer Vision (ICCV)*, volume 4, 2017.
- [34] W. Wu, Y. Zhang, C. Li, C. Qian, and C. C. Loy. Reenactgan: Learning to reenact faces via boundary transfer. In *ECCV*, 2018.
- [35] X. Wu, R. He, Z. Sun, and T. Tan. A light cnn for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, 13:2884–2896, 2015.
- [36] R. Xu, Z. Zhou, W. Zhang, and Y. Yu. Face transfer with generative adversarial network. *arXiv preprint arXiv:1710.06090*, 2017.
- [37] R. A. Yeh, Z. Liu, D. B. Goldman, and A. Agarwala. Semantic facial expression editing using autoencoded flow. 2016.
- [38] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.