Self-supervised learning of a facial attribute embedding from video

Olivia Wiles*
ow@robots.ox.ac.uk
A. Sophia Koepke*
koepke@robots.ox.ac.uk
Andrew Zisserman
az@robots.ox.ac.uk

Visual Geometry Group
University of Oxford
Oxford, UK

Abstract

We propose a self-supervised framework for learning facial attributes by simply watching videos of a human face speaking, laughing, and moving over time. To perform this task, we introduce a network, **F**acial **A**ttri**b**utes-**Net** (FAb-Net), that is trained to embed multiple frames from the same video face-track into a common low-dimensional space. With this approach, we make three contributions: first, we show that the network can leverage information from multiple source frames by predicting confidence/attention masks for each frame; second, we demonstrate that using a curriculum learning regime improves the learned embedding; finally, we demonstrate that the network learns a meaningful face embedding that encodes information about head pose, facial landmarks and facial expression – i.e. facial attributes – *without* having been supervised with any labelled data. We are comparable or superior to state-of-the-art self-supervised methods on these tasks and approach the performance of supervised methods.

1 Introduction

Babies and children are highly perceptive to the facial expressions of the people they interact with [14, 18]. The ability to understand and respond to changes in people's emotional state is similarly important for computer vision systems and affective systems when interacting with a human user. Thus being able to predict head pose and expression is of vital importance.

Recently, leveraging deep learning has led to state-of-the-art results on a variety of tasks such as emotion recognition and facial landmarks detection. Despite these advances, supervised methods require large amounts of labelled data which may be expensive or difficult to obtain in realistic, unconstrained settings, or necessitate assigning data to ill-defined categories. For example, categorising emotions with three human annotators leads to only 46% agreement [2], and labelling pose in the wild is notoriously difficult. Moreover, performing each task independently does not leverage the fact that detecting landmarks requires understanding pose and facial features, which in turn correspond to expression.

Consequently, we consider the following question: is it possible to learn an embedding of facial attributes that encodes landmarks, pose, emotion, etc. in a self-supervised manner without *any* hand labelling? The learned embedding can then be used for another task

* Equal contribution.

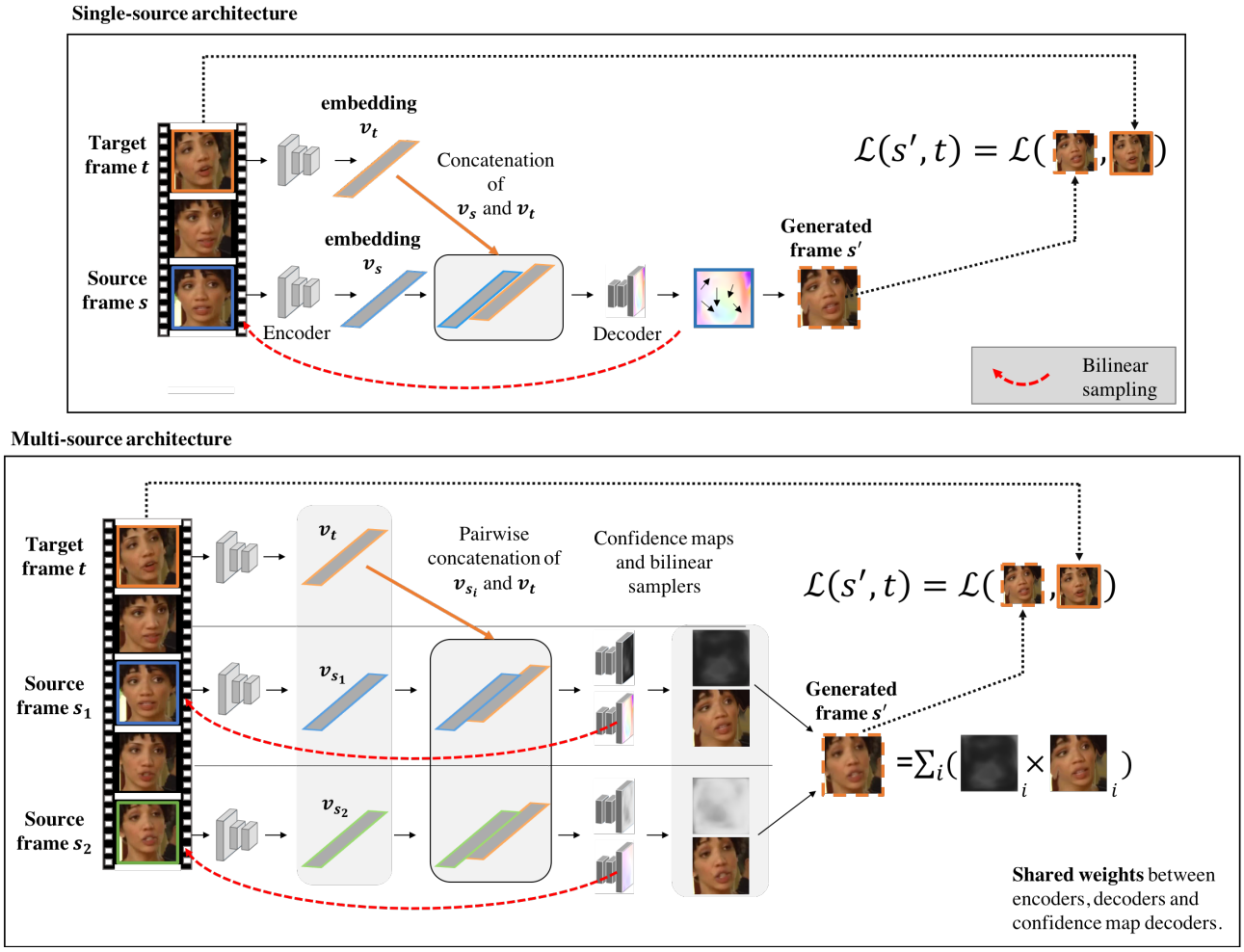


Figure 1: An overview of FAb-Net. In the single-source case (top), the encoder-decoder architecture takes one source frame and a target frame as inputs and learns to generate the target frame. The 256-dimensional outputs of the encoders – the source and target attribute embeddings – are concatenated and input to the decoder. The decoder predicts the point-to-point correspondence from the source frame to the target frame, and RGB values for the generated frame are then obtained from the source frame using a bilinear sampler. The network is trained with an $L1$ loss between the generated frame and the target frame. In the multi-source case (bottom), the decoder also predicts a confidence map, and the confidence maps are used to weight the contributions of the different source frames.

(e.g. landmark, pose, and expression prediction) using a linear layer. To do this, we contribute FAb-Net, a self-supervised framework for learning a low-dimensional face embedding for facial attributes (Section 3). We take advantage of video data which contains a large collection of images of the same person from different viewpoints and with varied expressions. Given only the embeddings corresponding to a source and target frame, the network is tasked to map the source frame to the target frame by predicting a flow field between them. This proxy task forces the network to distil the information required to compute the flow field (e.g. the head pose and expression) into the source and target embeddings. After explaining the setup for a single source frame in Section 3.1, we introduce our additional contributions: a method for leveraging multiple frames in order to improve the learned embedding in Section 3.2; and how a curriculum strategy for training FAb-Net in Section 4 can be used to improve performance.

The learned embedding is extracted and used for a variety of tasks such as landmark

detection, pose regression, and expression classification in Section 5 by simply learning a linear layer. Our results on these tasks are comparable or superior to other self-supervised methods, approaching the performance of supervised methods. These experiments verify the hypothesis that our self-supervised framework learns to encode facial attributes which are useful for a variety of tasks. Finally, the method is tested qualitatively by using the learned embedding to retrieve images with similar facial attributes across different identities.

2 Related Work

Self-supervised learning. Self-supervised methods such as [12, 38, 43, 62] require no manual labelling of the images; instead, they directly use image data to provide proxy supervision for learning good feature representations. The benefit of these approaches is that the features learned using large quantities of available data can be transferred to other tasks/domains which have less or even no annotated data. To provide further supervision from image data, the images themselves can be transformed via a synthetic warp or rotation and the network trained to recognise the rotation [19] or to learn equivariant pixel embeddings [40, 51, 52].

Of more direct relevance to our training framework are self-supervised frameworks that use video data ([1, 8, 11, 16, 17, 21, 22, 30, 34, 44, 55, 56, 58, 61]). Our approach builds in particular on those that use frame synthesis [8, 11, 22, 44, 56, 58], though for us synthesis is a proxy task rather than the end goal. Note, unlike [16, 30, 34, 55], we do not make use of the temporal ordering information inherent in a video; nor do we predict future frames conditioned on a number of past frames [44], or explicitly predict the motion between frames as a convolutional kernel [22, 58], or condition the generation on another modality (e.g. voice [8]). Instead, we treat the frames as an unordered set, and propose a simple formulation: that by embedding the source and target frames in a common space and conditioning the transformation from source to target frame on these embeddings, the learned embeddings must learn to encode the relevant modes of variation necessary for the transformation.

Concurrent to our work, Zhang *et al*. [64] and Jakab *et al*. [21] build on [51] by using the discovered landmarks to reconstruct the original image. However, unlike these works, we do not place any constraints on the learned representation – such as an explicit representation that encodes landmarks as heatmaps.

Supervised learning of face embeddings. Given *known* (labelled) attribute information, e.g. for pose or expression, the embedding can be learned by training in a supervised manner to directly predict the attribute [28, 32, 46], or to generate images (of faces, cars, or other classes) at a new, *known* pose, expression, etc. [13, 25, 53, 59, 68]. Another way of supervising a face embedding is to explicitly learn the parameters of a 3D morphable model (3DMM) [7]. As fitting a 3DMM is relatively expensive, [3, 50] learn this end-to-end using either landmarks or a photometric error as supervision. However, unlike our method, these methods require either ground truth labels or a morphable model which fixes the modes of variation and the embedding.

An interesting half-way point is weak supervision, where the learned object or face embedding is conditioned for instance on object labels [39] or weather/geo-location information [31] respectively. This requires additional meta-data, but results in embeddings that can represent attributes such as age and expression for faces or keypoints for objects.

3 Method

The aim is to train a network to learn an embedding that encodes facial attributes in a self-supervised manner, without any labels. To do this, the network is trained to generate a target frame from one or multiple source frames by learning how to transform the source into the target frame. The source and target frames are taken from the same face-track of a person speaking, i.e. the frames are of the same identity but with different expressions/poses. An overview of the architecture is given in Fig. 1 and further described for a single source frame in Section 3.1 and for multiple source frames in Section 3.2 (additional details are given in the supp. material).

3.1 Single-source frame architecture

The input to the network is a source frame s and a target frame t from the same face-track. These are passed through encoders with shared weights which learn a mapping f from the input frames to a 256-dimensional vector embedding (as shown in Fig. 1). The embeddings corresponding to the target and source frames are $v_t = f(t)$ and $v_s = f(s)$ respectively. The source and target embeddings are concatenated to give a 512-dimensional vector which is upsampled via a decoder. The decoder learns a mapping g from the concatenated embeddings to a bilinear grid sampler, which samples from the source frame to create a new, generated frame $s' = g(v_t, v_s)(s)$. Precisely, g predicts offsets $(\delta x, \delta y)$ for each pixel location (x, y) in the target frame; the generated frame s' at location (x, y) is obtained by sampling from the source frame s according to these offsets : $s'(x, y) = s(x + \delta x, y + \delta y)$. The network is trained to minimise the $L1$ loss between the generated and the target frame: $\mathcal{L}(s', t) = ||t - s'||_1$.

This setup enforces that the embeddings v_s and v_t represent facial attributes of the source and target frames respectively since the decoder maps from the source frame s to generate the frame s' (i.e. it uses pixel RGB values from the source frame s to create the generated s' – a similar formulation has been proposed concurrently to this work by [54]). As the decoder is a function of the target and source attribute embeddings, and the decoder is the only place in the network where information is shared, the target attribute embedding must encode information about expression and pose in order for the decoder to know where to sample from in the source frame and where to place this information in the generated frame.

3.2 Multi-source frames architecture

While using two frames for training enforces that the network learns a high-quality embedding, additional source frames can be leveraged to improve the learned embedding. This is achieved by also predicting a confidence heatmap – a 1 channel image – for each source frame via an additional decoder. The heatmaps denote how confident the network is of the flow at each pixel location – e.g. if the source frame has a very different pose than the target frame, the confidence heatmap would have low certainty. Moreover, it can express this for sub-parts of the image; if the mouth is closed in the source but open in the target frame, the confidence heatmap can express uncertainty in this region. The confidence heatmaps C_i are combined pixel-wise for each source frame s_i using a soft-max operation. For n source frames, the loss function to be minimised is given as $\mathcal{L} = ||t - \frac{\sum_{i=1}^{n} e^{C_i} * (g(v_t, v_{s_i})(s_i))}{\sum_{i=1}^{n} e^{C_i}}||_1$.

4 Curriculum Strategy

The training of the network is divided into stages, so that knowledge can be built up over time as the examples given become progressively more difficult, as inspired by [5, 27]. The loss computed by a forward pass is used to rank samples (i.e. source and target frame pairs) in the batch according to their difficulty in a manner similar to [33, 37, 48, 49]. However, these methods use only the most difficult samples, which was found to stop our network from learning. Similarly to [37], using progressively more difficult samples proved crucial for the strategy's success.

Given a batch size of N randomly chosen samples, i.e. source and target frame pairs, a forward pass is executed and the loss for each sample computed. The samples are ranked and sorted according to this loss. Initially the loss is back-propagated only on the samples in the batch which are in the 50th percentile (i.e. the $0.5N$ samples with the lowest loss computed by the forward pass). These are assumed to be easier samples. When the loss on the validation set plateaus, the subset to be back-propagated on is shifted by 10 (e.g. the samples in the 10th-60th percentile range). This is repeated 4 times until the samples being back-propagated on fall into the 40th-90th percentile range. At this point the curriculum strategy is terminated, as it is assumed that the samples in the 90-100th range are too challenging or may be problematic (e.g. there is a large shift in the background which is too challenging to learn).

5 Experiments

In this section, we evaluate the network and the learned embedding. In Section 5.1, the performance of using FAb-Net's learned representation is compared to that of state-of-the-art self-supervised and supervised methods on a variety of tasks: facial landmark prediction, head pose regression and expression classification. Section 5.2 discusses the benefit of using additional source frames, and Section 5.3 shows how the learned representation can be used for retrieving images with similar facial attributes.

Training. The model is trained on the VoxCeleb1 and VoxCeleb2 video datasets [9, 36]; we refer to the combined datasets as VoxCeleb+. The VoxCeleb+ dataset consists of videos of interviews containing more than 1 million utterances of around 7,000 speakers. The frames are extracted at 1 fps. The frames are cropped, resized to 256×256, and the identities are randomly split into train/val/test (with a split of 75/15/10).

The models are trained in PyTorch [42] using SGD, an initial learning rate of 0.001, and momentum 0.9. When using the curriculum strategy described in Section 4, the batchsize is $N = 32$, else $N = 8$. The learning rate is divided by a factor of 10 when the loss on the validation set plateaus. (If the curriculum strategy is used, the learning rate is updated only when the 40-90th percentile is considered.) This is repeated until the loss converges. Further details about the training can be found in the supp. material.

5.1 Using the embedding for regression and classification

First, we investigate the representation learned in our embedding and evaluate whether it indeed encodes facial attributes by challenging it to predict three different attributes: landmarks, pose, and expression.

Setup. Given a network trained on VoxCeleb+, a linear regressor or classifier is trained from the learned embedding to the output task. The linear regressor/classifier consists of two layers: batch-norm [20] followed by a linear fully connected layer with no bias. The regression tasks are trained using a MSE loss. The classification tasks are trained with a cross-entropy loss. The parameters of the encoder are fixed while the two additional layers are trained on the training set of the target dataset using Adam [23], a learning rate of 0.001, $\beta_1 = 0.9$ and $\beta_2 = 0.999$.

5.1.1 Baselines

Self-supervised. There are prior publications on using self-supervision for landmark prediction on the datasets we evaluate on, but none for predicting emotion on standard datasets. Consequently, we implement an autoencoder and a set of state-of-the-art self-supervised methods [19, 63] for object detection and segmentation. The baselines are trained using the same architecture as FAb-Net but with their associated loss functions and training objectives. For [63], the regression loss for both the L and ab channels is used. These models are trained on VoxCeleb+ until convergence, with the same training parameters and data augmentation as FAb-Net. More details are given in the supp. material.

VGG-Face descriptor. We additionally compare to the *VGG-Face descriptor* which is obtained from the 4096-dimensional FC7 features from a VGG-16 network trained on the VGG-Face dataset [41]. Contrary to popular belief, it has been recently shown that a network trained for identity does retain information about other facial attributes [10, 15]. We use the VGG-Face descriptor to learn a linear regression/classification layer to the desired attribute task. This provides a strong baseline, and the results obtained confirm the finding that a network trained for identity does indeed encode expression and to some extent also pose information. However, note that unlike our method, this face descriptor requires a large dataset of *labelled* face images for training.

5.1.2 Results

Facial landmarks. Facial landmark locations are regressed from the learned embedding and compared to state-of-the-art methods on MAFL [66] and the more challenging 300-W [47] datasets. The evaluation is performed as outlined in [51, 66], and the errors given in inter-ocular distance. For MAFL, 5 facial landmarks are regressed for 19k/1k train/test images. For 300-W, 68 landmarks are regressed for 3148/689 train/test images which are obtained (as described in [51]) from combining multiple datasets [4, 67, 70].

The results are reported in Table 1 and some qualitative results are visualised in Fig. 2 and Fig. 3. These results first demonstrate that fine-tuning with additional views and our curriculum strategy improve the embedding learned by FAb-Net. Second, these results show that our method performs competitively or better than state-of-the-art unsupervised landmark detection methods, better than the VGG-Face descriptor baseline and competitively with state-of-the-art supervised methods. This is achieved even though the other self-supervised methods [21, 51, 52, 64] are explicitly engineered to detect landmarks whereas our method is not. In addition to that, our method is able to bridge the domain gap between VoxCeleb+ and CelebA [32] (the other self-supervised methods that we compare to are pre-trained on CelebA).

Pose. The learned embedding is used for pose prediction and compared to a supervised method [26] and to using the VGG-Face descriptor. To perform the evaluation, the linear

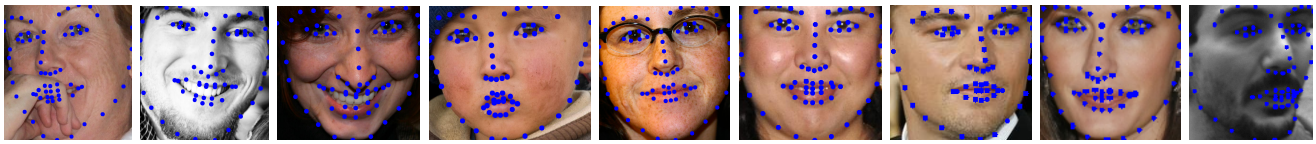
Method	300-W	MAFL
Self-supervised		
Trained on VoxCeleb+		
FAb-Net	6.31	3.78
FAb-Net w/ curric.	5.73	3.49
FAb-Net w/ curric., 3 source frames	**5.71**	3.44
Trained on CelebA		
Jakab *et al.* [21] (2018)	–	**3.08**
Zhang *et al.* [64] (2018)	–	3.15
Thewlis *et al.* [51] (2017)	9.30	6.67
Thewlis *et al.* [52] (2017)	7.97	5.83
Supervised		
Trained on CelebA		
MTCNN [65] (2014)	–	**5.39**
LBF [45] (2014)	6.32	–
CFSS [69] (2015)	5.76	–
cGPRT [29] (2015)	5.71	–
DDN [60] (2016)	5.65	–
TCDCN [66] (2016)	5.54	–
RAR [57] (2016)	**4.94**	–
VGG-Face descriptor [41]	11.16	5.92

Table 1: Landmark prediction error on 300-W and MAFL datasets. Lower is better.

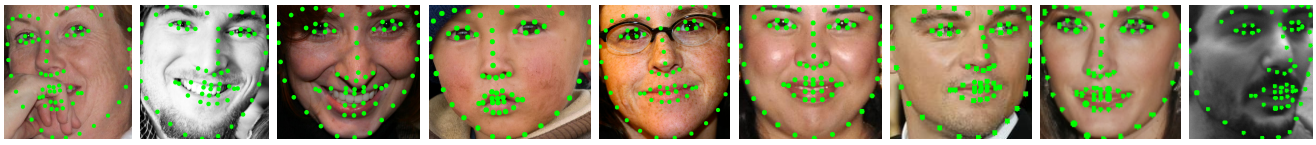
Method	Roll	Pitch	Yaw	MAE
Self-supervised				
FAb-Net	5.54°	7.84°	12.93°	8.77°
FAb-Net w/ curric.	5.33°	7.21°	11.34°	7.96°
FAb-Net w/ curric., 3 source frames	**5.14°**	**7.13°**	**10.70°**	**7.65°**
Supervised				
VGG-Face descriptor [41]	**8.24°**	8.36°	18.35°	11.65°
KEPLER [26] (2017)	8.75°	**5.85°**	**6.45°**	**7.02°**

Table 2: Pose prediction error on the AFLW test set from [26]. Lower is better.

Figure 2: Landmark prediction visualisation for FAb-Net on the MAFL dataset. A dot denotes ground truth and the cross FAb-Net's prediction. A failure case is shown to the right.



(a) Ground truth landmarks



(b) FAb-Net's predicted landmarks

Figure 3: Landmark prediction visualisation for FAb-Net on the 300-W dataset.

regression is trained from the given embedding to head pose labels using the AFLW dataset [24], but after leaving out the 1,000 images of the AFLW test set from [26]. As can be seen in Table 2, FAb-Net performs better in predicting the roll angle, and the MAE is comparable to [26] which is supervised with head pose labels. Furthermore, our embedding outperforms the VGG-Face descriptor which is trained on identities; i.e. our learned embedding encodes more information about head pose.

Expression. We evaluate the performance of our learned embedding for expression estimation on two datasets: AffectNet [35] and EmotioNet [6], which both contain over 900,000 images. These datasets are taken 'in-the-wild' as opposed to in a constrained environment. AffectNet contains 8 facial expressions (neutral, happy, sad, surprise, fear, disgust, anger, contempt) and EmotioNet contains 11 action units (AUs) (combinations of AUs correspond

	AUC for different AUs											
	1	2	4	5	6	9	12	17	20	25	26	avg.
Self-supervised												
FAb-Net	72.0	68.9	73.2	69.4	88.2	78.6	89.5	71.0	75.9	81.4	72.0	76.4
FAb-Net w/ curriculum	73.4	71.8	75.3	67.8	90.4	78.8	91.9	72.4	74.5	**83.7**	73.3	77.6
FAb-Net w/ curriculum, 3 source frames	**74.1**	**72.3**	**75.8**	**68.8**	**90.7**	**81.8**	**92.5**	**73.7**	**77.2**	83.6	**73.6**	**78.6**
Gidaris *et al*. [19]	68.6	64.0	72.8	70.0	83.9	78.1	83.8	68.4	72.6	73.1	67.2	72.9
SplitBrain [63]	65.5	59.8	66.7	60.8	71.8	65.8	73.3	64.5	57.4	68.1	61.1	65.0
Autoencoder	67.2	60.5	70.1	65.1	79.6	70.4	80.1	68.3	66.5	70.5	64.1	69.3
Supervised												
VGG-Face descriptor [41]	**81.8**	**83.0**	83.5	81.8	92.0	**90.9**	95.7	**80.6**	85.2	86.5	73.0	84.9
VGG-11 (from scratch)	74.7	77.2	**85.8**	**83.7**	**93.8**	89.7	**97.5**	78.3	**86.9**	**96.4**	**81.5**	**86.0**

Table 3: Expression classification results for state-of-the-art self-supervised and supervised methods on EmotioNet [6] for multiple facial action units (AUs). Higher is better for AUC.

	AUC								
	Neutral	Happy	Sad	Surprise	Fear	Disgust	Anger	Contempt	avg.
Self-supervised									
FAb-Net	70.0	87.6	68.8	75.5	76.5	70.0	73.2	71.2	74.2
FAb-Net w/ curric.	71.5	90.0	70.8	78.2	77.4	72.2	75.7	72.1	76.0
FAb-Net w/ curric., 3 source frames	**72.3**	**90.4**	**70.9**	**78.6**	**77.8**	**72.5**	**76.4**	**72.2**	**76.4**
Gidaris *et al*. [19]	67.8	84.9	69.0	73.9	75.7	69.8	71.5	68.7	72.7
SplitBrain [63]	63.9	74.7	64.2	61.3	68.3	58.6	68.2	62.8	65.3
Autoencoder	65.8	80.0	64.7	66.1	70.6	63.4	68.3	65.0	68.0
Supervised									
VGG-Face descriptor [41]	75.9	92.2	80.5	81.4	82.3	81.4	81.2	77.1	81.5
AlexNet [35]	–	–	–	–	–	–	–	–	**82**

Table 4: Expression classification results for state-of-the-art self-supervised and supervised methods on AffectNet [35]. Higher is better for AUC.

to facial expressions).

Both of these datasets were organised for challenges with a held out, unreleased test set. Therefore, the train set is subdivided into two subsets; one is used for training and the other for validation. The validation set of the original dataset is used to test the different models. The linear classifier for EmotioNet is trained with a binary cross-entropy loss for each AU, whereas for AffectNet, a cross-entropy loss is used. Both training datasets are highly imbalanced. As a result, the examples from the under-represented classes are re-weighted inversely proportionally to the class frequencies to penalise the loss more heavily for mis-classifying images of the under-represented classes.

The embedding learned by FAb-Net is compared to a number of self-supervised and supervised methods by measuring the Area Under the ROC curve (AUC). For each class (e.g. emotion or AU), the AUC is computed independently and the result is averaged over all classes. The results are reported in Table 3 and Table 4 for EmotioNet and AffectNet respectively showing that our network performs better than other self-supervised methods over both metrics when given the same training data. This is supposedly due to the fact that the network must learn to transform the source frame in order to generate the target frame. As parts of the face move together (e.g. an eyebrow raise or the lips when the mouth opens), the embedding must learn to encode information about facial features and thereby encode expression. Interestingly, the autoencoder performs well, presumably due to the restricted nature of this domain.

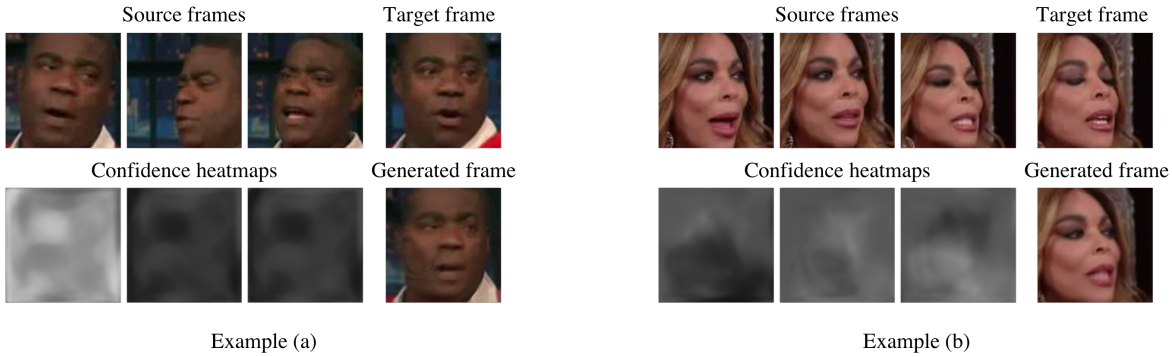


Figure 4: Confidence heatmaps learned by FAb-Net. Higher intensity corresponds to higher confidence. The network chooses the frames with most similar poses to draw from and ignores those with less similar poses (see Example (a)). In Example (b), the mouth has higher confidence in the third source frame allowing the network to re-construct the teeth that are present in the target frame. More examples can be found in the supp. material.

FAb-Net is also not far off supervised methods despite the domain shift; VoxCeleb+ consists only of people being interviewed (and consequently with mostly neutral/smiling faces), so it does not include the range/extremity of expressions found in AffectNet or EmotioNet. Finally, it can be observed that the VGG-Face descriptor trained to predict identities does surprisingly well at predicting emotion.

Discussion. FAb-Net has achieved impressive performance, as most self-supervised methods when transferred to another task have a large gap in comparison to supervised methods. There is no gap or a small gap for the smaller datasets (landmarks/pose) and the model approaches supervised performance for the larger datasets (expression).

5.2 What is the benefit of additional source frames?

The previous sections have shown that using additional source frames improves performance. This is at the expense of performing additional forward passes through the encoder (in this case two). Given enough GPU memory, these forward passes can be done in parallel, affecting only the memory requirements and not the computational speed.

Using multiple source frames is further investigated by visualising confidence heatmaps for a given set of source frames in Fig. 4. The confidence heatmaps allow images with more similar pose to be used for creating the generated frame. Furthermore, the network can focus on one frame for generating a part of the face (e.g. the mouth) and on another one for a different part.

5.3 Image retrieval

This section considers an application of the learned embedding: retrieving images with similar facial attributes (e.g. pose) but across different identities. To perform this task, a subset of 10,000 randomly sampled test images from VoxCeleb+ is obtained. For a given query image, all other images (the gallery) are ranked based on their similarity to the query image using the cosine similarity metric between the corresponding embeddings. For a given query image Q, the embedding x_q is extracted by performing a forward pass through the network. Similarly, the embedding x_i is extracted for each image I_i in the gallery. Each image I_i is then ranked according to the cosine similarity between x_q and x_i. If the network does indeed

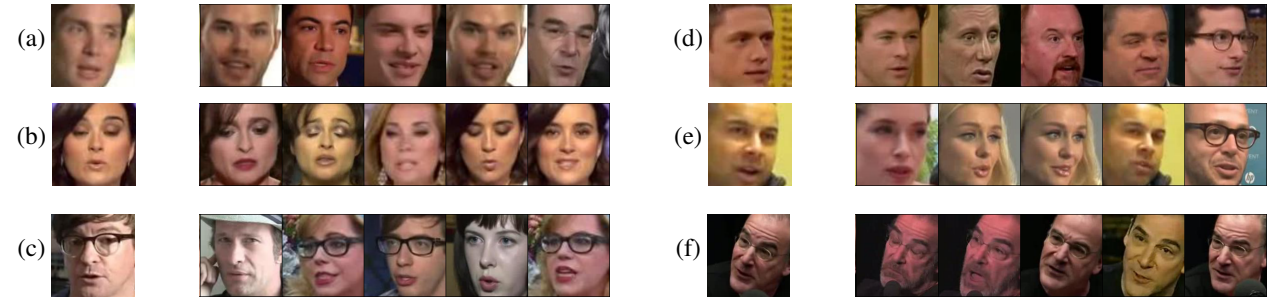


Figure 5: Retrieval results using the embedding learned by FAb-Net. The embedding captures similar visual attributes since gallery images with similar facial attributes are retrieved for a given query image. The retrieved images have similar pose to the query in all cases, and the expression similarity can be seen for example in (b) with the eyes shut, and (a) with the mouth slightly open. Please refer to the supp. material for additional examples.

encode salient information about facial attributes, the cosine similarity can be used to identify images with similar poses and facial attributes. For a set of query images, the results are visualised in Fig. 5. From these results it is again affirmed that our embedding encodes information about facial attributes, as the retrieved images have poses and expressions similar to those of the query images. Note, the embedding is largely unaffected by facial decorations (e.g. glasses) and identity, as these do not change within a face-track and so do not need to be learned in order to predict the transformation.

6 Conclusion

We have introduced FAb-Net: a self-supervised framework for learning facial attributes from videos. Our method learns about pose and expression by watching faces move and change over a large number of videos without *any* hand labels. The features of our trained network can then be used to predict pose, landmarks, and expression on other datasets (despite the domain shift) by just training a linear layer on top of the learned embedding. The features have been shown to be comparable or superior performance to self-supervised and supervised methods on a variety of tasks. This is impressive as generally the performance of self-supervised methods has been found to be worse than that of supervised methods, yet our method is indeed competitive/superior to supervised methods for pose regression and facial landmark detection, and approaches supervised performance on expression classification.

7 Acknowledgements

The authors would like to thank James Thewlis for helpfully sharing code and datasets. This work was funded by an EPSRC studentship and EPSRC Programme Grant Seebibyte EP/M013774/1.

References

[1] P. Agrawal, J. Carreira, and J. Malik. Learning to see by moving. In *Proc. ICCV*, 2015.

[2] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 279–283. ACM, 2016.

[3] A. Bas, P. Huber, W. A. P. Smith, M. Awais, and J. Kittler. 3D morphable models as spatial transformer networks. In *Proc. ICCV Workshop on Geometry Meets Deep Learning*, 2017.

[4] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. *IEEE PAMI*, 2013.

[5] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *Proc. ICML*, 2009.

[6] C. F. Benitez-Quiroz, R. Srinivasan, and A. M. Martinez. EmotioNet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proc. CVPR*, 2016.

[7] V. Blanz, S. Romdhani, and T. Vetter. Face identification across different poses and illumination with a 3D morphable model. In *Proc. AFGR*, 2002.

[8] J. S. Chung, A. Jamaludin, and A. Zisserman. You said that? In *Proc. BMVC.*, 2017.

[9] J. S. Chung, A. Nagrani, and A. Zisserman. VoxCeleb2: Deep speaker recognition. In *INTERSPEECH*, 2018.

[10] F. Cole, D. Belanger, D. Krishnan, A. Sarna, I. Mosseri, and W. T. Freeman. Synthesizing normalized faces from facial identity features. In *Proc. CVPR*, 2017.

[11] E. L. Denton and V. Birodkar. Unsupervised learning of disentangled representations from video. In *NIPS*, 2017.

[12] C. Doersch, A. Gupta, and A. A. Efros. Unsupervised visual representation learning by context prediction. In *Proc. ICCV*, 2015.

[13] A. Dosovitskiy, J. T. Springenberg, and T. Brox. Learning to generate chairs with convolutional neural networks. In *Proc. CVPR*, 2015.

[14] P. Ekman and H. Oster. Facial expressions of emotion. *Annual review of psychology*, 30(1):527–554, 1979.

[15] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *Proc. ACM SIGGRAPH*, 2018.

[16] B. Fernando, H. Bilen, E. Gavves, and S. Gould. Self-supervised video representation learning with odd-one-out networks. In *Proc. CVPR*, 2017.

[17] C. Gan, B. Gong, K. Liu, H. Su, and L. J. Guibas. Geometry guided convolutional neural networks for self-supervised video representation learning. In *Proc. CVPR*, 2018.

[18] F. C. Gerull and R. M. Rapee. Mother knows best: effects of maternal modelling on the acquisition of fear and avoidance behaviour in toddlers. *Behaviour research and therapy*, 40(3):279–287, 2002.

[19] S. Gidaris, P. Singh, and N. Komodakis. Unsupervised representation learning by predicting image rotations. In *Proc. ICLR*, 2018.

[20] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proc. ICML*, 2015.

[21] T. Jakab, A. Gupta, H. Bilen, and A. Vedaldi. Conditional image generation for learning the structure of visual objects. *arXiv preprint arXiv:1806.07823*, 2018.

[22] X. Jia, B. De Brabandere, T. Tuytelaars, and L. V. Gool. Dynamic filter networks. In *NIPS*, 2016.

[23] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *Proc. ICLR*, 2015.

[24] M. Koestinger, P. Wohlhart, P. M. Roth, and H. Bischof. Annotated Facial Landmarks in the Wild: A Large-scale, Real-world Database for Facial Landmark Localization. In *Proc. First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies*, 2011.

[25] T. D. Kulkarni, W. F. Whitney, P. Kohli, and J. Tenenbaum. Deep convolutional inverse graphics network. In *NIPS*, 2015.

[26] A. Kumar, A. Alavi, and R. Chellappa. KEPLER: keypoint and pose estimation of unconstrained faces by learning efficient H-CNN regressors. In *Proc. Int. Conf. Autom. Face and Gesture Recog.*, 2017.

[27] M. P. Kumar, B. Packer, and D. Koller. Self-paced learning for latent variable models. In *NIPS*. 2010.

[28] N. Kumar, A. Berg, P. Belhumeur, and S. Nayar. Describable visual attributes for face verification and image search. *IEEE PAMI*, 33(10):1962–1977, 2011.

[29] D. Lee, H. Park, and C. D. Yoo. Face alignment using cascade gaussian process regression trees. In *Proc. CVPR*, 2015.

[30] H.-Y. Lee, J.-B. Huang, M. Singh, and M.-H. Yang. Unsupervised representation learning by sorting sequences. In *Proc. ICCV*, 2017.

[31] Y. Li, R. Wang, H. Liu, H. Jiang, S. Shan, and X. Chen. Two birds, one stone: Jointly learning binary code for large-scale face image retrieval and attributes prediction. In *Proc. ICCV*, 2015.

[32] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proc. ICCV*, 2015.

[33] I. Loshchilov and F. Hutter. Online batch selection for faster training of neural networks. *ICLR Workshops*, 2016.

[34] I. Misra, C. L. Zitnick, and M. Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *Proc. ECCV*, 2016.

[35] A. Mollahosseini, B. Hasani, and M. H. Mahoor. AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 2017.

[36] A. Nagrani, J. S. Chung, and A. Zisserman. VoxCeleb: a large-scale speaker identification dataset. In *INTERSPEECH*, 2017.

[37] A. Nagrani, S. Albanie, and A. Zisserman. Learnable pins: Cross-modal embeddings for person identity. *arXiv preprint arXiv:1805.00833*, 2018.

[38] M. Noroozi and P. Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Proc. ECCV*, 2016.

[39] D. Novotny, D. Larlus, and A. Vedaldi. Anchornet: A weakly supervised network to learn geometry-sensitive features for semantic matching. In *Proc. CVPR*, 2017.

[40] D. Novotny, S. Albanie, D. Larlus, and A. Vedaldi. Self-supervised learning of geometrically stable features through probabilistic introspection. In *Proc. CVPR*, 2018.

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

[42] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in PyTorch. 2017.

[43] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In *Proc. CVPR*, 2016.

[44] V. Pătrăucean, A. Handa, and R. Cipolla. Spatio-temporal video autoencoder with differentiable memory. In *NIPS*, 2016.

[45] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 fps via regressing local binary features. In *Proc. CVPR*, 2014.

[46] E. M. Rudd, M. Günther, and T. E. Boult. Moon: A mixed objective optimization network for the recognition of facial attributes. In *Proc. ECCV*, 2016.

[47] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: Database and results. *Image and Vision Computing*, 47:3–18, 2016.

[48] A. Shrivastava, A. Gupta, and R. Girshick. Training region-based object detectors with online hard example mining. In *Proc. CVPR*, 2016.

[49] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, and F. Moreno-Noguer. Discriminative learning of deep convolutional feature point descriptors. In *Proc. ICCV*, 2015.

[50] A. Tewari, M. Zollhöfer, H. Kim, P. Garrido, F. Bernard, P. Perez, and C. Theobalt. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *Proc. ICCV*, 2017.

[51] J. Thewlis, H. Bilen, and A. Vedaldi. Unsupervised learning of object landmarks by factorized spatial embeddings. In *Proc. ICCV*, 2017.

[52] J. Thewlis, H. Bilen, and A. Vedaldi. Unsupervised learning of object frames by dense equivariant image labelling. In *NIPS*, 2017.

[53] L. Tran, X. Yin, and X. Liu. Disentangled representation learning gan for pose-invariant face recognition. In *Proc. CVPR*, 2017.

[54] C. Vondrick, A. Shrivastava, A. Fathi, S. Guadarrama, and K. Murphy. Tracking emerges by colorizing videos. In *Proc. ECCV*, 2018.

[55] X. Wang and A. Gupta. Unsupervised learning of visual representations using videos. In *Proc. ICCV*, 2015.

[56] O. Wiles, A. S. Koepke, and A. Zisserman. X2Face: A network for controlling face generation using images, audio, and pose codes. In *Proc. ECCV*, 2018.

[57] S. Xiao, J. Feng, J. Xing, H. Lai, S. Yan, and A. Kassim. Robust facial landmark detection via recurrent attentive-refinement networks. In *Proc. ECCV*, 2016.

[58] T. Xue, J. Wu, K. Bouman, and B. Freeman. Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks. In *NIPS*, 2016.

[59] J. Yang, P. Ren, D. Chen, F. Wen, H. Li, and G. Hua. Neural aggregation network for video face recognition. In *Proc. CVPR*, 2017.

[60] X. Yu, F. Zhou, and M. Chandraker. Deep deformation network for object landmark localization. In *Proc. ECCV*, 2016.

[61] A. R. Zamir, T. Wekel, P. Agrawal, C. Wei, J. Malik, and S. Savarese. Generic 3d representation via pose estimation and matching. In *Proc. ECCV*, 2016.

[62] R. Zhang, P. Isola, and A. A. Efros. Colorful image colorization. In *Proc. ECCV*, 2016.

[63] R. Zhang, P. Isola, and A. A. Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proc. CVPR*, 2017.

[64] Y. Zhang, Y. Guo, Y. Jin, Y. Luo, Z. He, and H. Lee. Unsupervised discovery of object landmarks as structural representations. In *Proc. CVPR*, 2018.

[65] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *Proc. ECCV*, 2014.

[66] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Learning deep representation for face alignment with auxiliary attributes. *IEEE PAMI*, 2016.

[67] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin. Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In *ICCVW*, 2013.

[68] T. Zhou, S. Tulsiani, W. Sun, J. Malik, and A. A. Efros. View synthesis by appearance flow. In *Proc. ECCV*, 2016.

[69] S. Zhu, C. Li, C. C. Loy, and X. Tang. Face alignment by coarse-to-fine shape searching. In *Proc. CVPR*, 2015.

[70] X. Zhu and D. Ramanan. Face detection, pose estimation and landmark localization in the wild. *Proc. CVPR*, 2012.