StyleGAN-based heatmap generator for face alignment with limited training data

Martin Dornier, Philippe-Henri Gosselin, Christian Raymond, Yann Ricquebourg, Bertrand B. Coüasnon

Highlights

StyleGAN-based heatmap generator for face alignment with limited training data

Martin Dornier, Philippe-Henri Gosselin, Christian Raymond, Yann Ricquebourg, Bertrand Coüasnon

- We transform a StyleGAN generator into a facial landmark heatmap generator.
- Our face alignment model can be trained with limited training data.
- We beat state-of-the-art in the low training data setting.

StyleGAN-based heatmap generator for face alignment with limited training data

Martin Dornier[a,b,*], Philippe-Henri Gosselin[a], Christian Raymond[b], Yann Ricquebourg[b], Bertrand Coüasnon[b]

[a] *InterDigital, 975 Av. des Champs Blancs, 35510, Cesson-Sévigné, France*
[b] *Univ. Rennes, CNRS, IRISA, France, 263 avenue du Général Leclerc, 35042, Rennes, France*

Abstract

While the performance of face alignment models has been improving over the years, they still need large, annotated datasets during their training to perform well. In this paper, we propose a new architecture to perform face alignment with limited training data. Our model is based on StyleGAN, a popular architecture in the image generation domain, and takes advantage of its strong generative power to generate accurate facial landmark heatmaps of real face images, using only a small amount of training data. Even when trained down to only 50 samples, our model can still predict accurate facial landmarks. It exceeds state-of-the-art on several face alignment datasets in the low training data regime.

Keywords: face alignment, semi-supervised training, active learning, transfer learning

*Corresponding author:
Email address: martin.dornier@interdigital.com (Martin Dornier)

1. Introduction

The success of Deep Learning comes mainly from its ability to automatically learn optimal features for the target task instead of relying on hand-crafted features such as SIFT [1] or HOG [2]. A subdomain of Machine Learning is Representation Learning; the goal, in this case, is to learn useful features, e.g., interpretable or that can be used for transfer learning. One application of representation learning is reducing the number of annotated samples needed to train a model on a downstream discriminative task, such as classification or regression. Popular representation learning methods make use of auto-encoders [3, 4, 5], contrastive learning [6, 7, 8], or masked inputs [9, 10]. While some methods also use Generative Adversarial Networks (GANs) [11, 12, 13, 14, 15], we think that this idea is very promising and has not been fully explored, especially with the newest GANs architectures.

Among GANs, StyleGAN is a popular approach thanks to its disentangled latent space that can be distilled into multiple styles [16, 17, 18]. Each style controls a semantic characteristic of the image such as color scheme, object shape, or background. Because of its lack of an inference module, many approaches propose to *invert* StyleGAN in order to project real images into its latent space and perform semantic editing of these images [19, 20, 21, 15, 22, 23, 24, 14, 25]. However, most of these works focus on image-to-image translation tasks, such as attribute editability, super-resolution, or inpainting. While some methods have studied the use of StyleGAN for regression tasks [14, 15], they usually focus on semantic attribute regression such as face pose or age estimation [15] or don't evaluate in depth how well their method performs compared to modern supervised methods or

semi-supervised methods in the setting of limited training data.

In this work, we study the use of StyleGAN for a challenging regression task: facial landmark prediction (also called face alignment) with limited training data. Face alignment tries to localize some pre-defined facial anatomical keypoints (such as the corners of the mouth, the eyes, the boundaries of the face, ...) in a face image. Many downstream tasks rely on the predicted landmarks such as face swapping or facial expression recognition.

Through the years, the performance of face alignment models has increased, especially since the rise of Deep Learning, although most of the recent improvements come more from specifically designed training schemes like complex training losses [26, 27] rather than better network architectures.

Also, because labeling facial keypoints is time-consuming and can be challenging on images with large poses or occlusions, non-synthetic facial landmark datasets are usually relatively small (a few thousand samples) compared to other computer vision tasks such as image classification, making the trained models prone to overfitting. In this paper, we study whether it is possible to train a facial landmark detector with only a few samples and still get good accuracy and generalization. To tackle this challenge, we propose a new architecture. We transform a pre-trained StyleGAN generator, through the use of additional layers, into a facial landmark heatmap generator which can be trained with limited training data. Paired with a StyleGAN encoder, we can perform face alignment on real images.

With this new framework, we achieve competitive results on multiple facial landmark datasets, even on images very different from the StyleGAN generative distribution, and beat state-of-the-art in the low-data regime. Our

contributions are as follows:

- We propose a new architecture derived from StyleGAN to perform face alignment with limited training data. On multiple 2D and 3D face alignment datasets, our method achieves competitive results in the fully supervised setting and systematically beats state-of-the-art in the low training data setting.
- We study the use of active learning to improve our results even further.

The rest of our paper is organized as follows: first, we present in Section 2 the StyleGAN architecture and existing works on StyleGAN inversion. We also introduce the face alignment task and sum up methods dealing with limited training data for this task. For the last part of this section, we explain the active learning principle and list several existing methods and applications. In Section 3, we present our proposed framework for face alignment with limited training data. Section 4 reports the results of our different experiments. Finally, Section 5 concludes this paper.

2. Related Work

This section relates works related to our method, based on StyleGAN, and its purpose: face alignment with limited training data. The active learning principle is also explained and different existing methods are presented.

2.1. StyleGAN

Generative Adversarial Networks (GANs) [28] have greatly improved the quality of image generation over the last few years. StyleGAN [16] differs

from previous GAN architectures by its generative process. Instead of starting from Gaussian noise $z \in \mathcal{Z}$ (the latent representation) and progressively increasing the spatial dimensions through the network layers, z is first projected to an intermediate latent space $\mathcal{W}$ via a non-linear mapping network $f : \mathcal{Z} \to \mathcal{W}$ which produces an intermediate latent code $w \in \mathcal{W}$. The input of the generator is a constant learned vector $c1$ and at each layer, w is transformed by an affine transformation (different for each layer i) into a style vector y_i and injected into the current feature map via an AdaIn [29] operation. Each style vector controls a specific aspect of the generated image, style vectors corresponding to low-resolutions control high-level attributes such as, for face images, pose, face shape, or hairstyle while high-resolution style vectors control fine-grained aspects such as the color scheme or microstructure.

2.2. StyleGAN Inversion

To semantically edit a *synthetic* image generated by StyleGAN, we can modify some of its style vectors. To modify in the same way a *real* image, we first need to approximate its StyleGAN latent vector, this is called StyleGAN inversion. StyleGAN inversion methods are divided into three families. The *optimization-based* methods iteratively refine a latent code by minimizing the reconstruction error [20, 21]. The *encoder-based* methods train an encoder to predict the latent code [15, 22, 23, 24, 14, 25]. Finally, *hybrid* methods train an encoder to predict an initial latent code which is refined through optimization [19]. Optimization-based and hybrid-based methods usually have better reconstruction errors but are much slower than encoder-based.

Rather than predicting the true latent code $z \in \mathcal{Z}$ or the intermediate latent code $w \in \mathcal{W}$, most methods predict a code for each style: $w^+ =$

$(w_1, w_2, ..., w_n) \in \mathcal{W}^+$, n being the number of styles [15, 22, 23, 24]. This gives more flexibility and improves the reconstruction error. Some methods predict along w^+ a feature map $f \in \mathcal{F}$ which replaces the first layers of the generator [21, 25]. This feature map improves the reconstruction error but also makes it possible to encode images that do not follow the training dataset alignment (e.g., FFHQ [16]) and Celeba-HQ [30] always have the face centered in the image and eyes at the same level). For example, translated or rotated images can still be faithfully encoded and reconstructed.

2.3. StyleGAN for regression

While the most common use cases for StyleGAN inversion are image-to-image translation tasks, such as attribute edition or super-resolution [19, 20, 21, 22, 23, 24, 25], it can also be used for discriminative tasks like classification or regression. GHFeat [14] proposes a StyleGAN encoder which predicts the style vectors $\{y_i\}$. Then for each discriminative task, such as face verification and face alignment, a fully connected layer with the predicted style vectors as input is trained for the task. However, compared to modern supervised methods, its results are quite poor, and it does not study how well its method would perform when training with limited data. LARGE [15] does not propose a new StyleGAN encoder architecture but notices that the linear directions in the latent space which affect image attributes are also linear in terms of the magnitude of change, so it is possible to create a linear regression model using a few calibration samples. Unfortunately, its method is restricted to image attribute regression, such as age or head pose estimation, and can't be applied for more complex tasks like facial landmark prediction.

2.4. Face Alignment

Rather than directly predicting the position of the landmarks in the image, most of the recent 2D face alignment models predict facial landmark heatmaps [31], one for each landmark. The landmark position is then inferred from the best local maximum of the heatmap.

In the case of 3D face alignment, we try to detect the true anatomical position of the keypoints, for example, the landmarks on the outline of the face remains at the same anatomical position even if there are occluded (the reader can compare the Figure 2 and 4 for a visual explanation). For this task, methods can be divided into two families, those that attempt to detect the landmarks directly [32] or those that aim to fit a 3D face model and then obtain the landmarks from this model [33, 34, 35]. Wu et al. [36] learn both tasks at the same time.

Semi-supervised methods try to alleviate the problem of facial landmark annotations we explained in the introduction of this paper. To do so they use annotated and non-annotated data during the training. Honori et al. [37] impose equivariance of predicted landmarks to geometric transformations. TS^3 [38] uses pseudo-labeling with a teacher-student method. Some methods are based on transfer learning: 3FabRec [3] trains an auto-encoder to reconstruct face images and then modifies its decoder to generate facial landmark heatmaps, SCAF [39] improves 3FabRec by adding skip-connections to the auto-encoder and using active learning. FaRL [8] uses masked image modeling and image-text contrastive learning on a large text/image pair dataset to pre-train a network and then use it for several facial downstream tasks including face alignment. Transfer learning is an interesting principle to train

with limited annotated data but the whole pipeline can be computationally expensive and we think the carbon footprint for training a model should be taken into account. In the case of 3FabRec and SCAF, the self-supervised training of the encoder is done using millions of face images and lasts for days. Also, its sole purpose is the face alignment task. FaRL, in its principle, is more efficient because the pre-trained network can be used for different facial downstream tasks such as face parsing or facial attribute recognition and not only facial alignment. But its architecture is based on a visual Transformer [40], a heavy architecture, and trains on a dataset of 20 million samples using 32 Nvidia V-100 GPUs so the pre-training is very expensive.

Our approach is also based on transfer learning but we don't need to do any dedicated pre-training prior to our face alignment training. Rather, we promote StyleGAN pre-trained generators and encoders, which have been trained for a completely different purpose: face image generation for generators and face attribute modification for encoders. It makes the carbon footprint for the redaction of this paper very low compared to the previously discussed transfer learning methods.

Another solution to tackle the challenge of facial landmark annotations is to use synthetic samples. Qian et al. [41] train a network to produce multiple images with different styles but the same face pose from an input image to increase the training dataset size. Wood et al. [42] use a fully synthetic dataset created with computer graphics [43] to train a facial landmark detector. With computer graphics, it is possible to generate a lot of samples with perfect annotations, but generating such a dataset requires much computation power and there is always a domain gap between the generated

images and real face images which may deteriorate the performance of the model on real test images. Also, if you need a new kind of annotations (additional landmarks for example), you need to generate again the whole dataset. While Wood et al. [42] obtain fair results on the face alignment dataset 300-W [44], it is hard to determine how well their model would perform on more challenging face alignment datasets such as WFLW [45].

2.5. Active learning

In academic research, most of the time, scientists randomly choose the labeled samples from the whole labeled training set to demonstrate the effectiveness of their model in the few-shot setting. Yet, at first, this labeled training set does not exist for real-world applications, and one must choose the samples to annotate among an unlabeled training dataset.

The goal of active learning is to select the most valuable samples to annotate rather than sampling them randomly from the initial unlabeled dataset. Indeed, in the case of random sampling, very similar or easy samples might be sampled even though they won't help the model to improve during the training. Active learning tries to prevent this kind of situation. When annotating samples is very time-consuming, such as annotating facial landmarks, active learning can be very valuable.

Active learning is an iterative procedure: from a non-annotated dataset $\mathcal{U}_N$, an initial set $\mathcal{L}^0$ is annotated and then the model is trained on $\mathcal{L}^0$. Once the training is over, all the remaining unlabeled samples from $\mathcal{U}_N$ are ranked using an *acquisition function* which depends on the model predictions, the K best samples are annotated and added to $\mathcal{L}^0$ giving a new annotated set $\mathcal{L}^1$. Then, the model is trained again from scratch on this new annotated set and

$\mathcal{U}_N$ samples are ranked again, etc. This scheme repeats until the exhaustion of the annotation budget.

The crucial part of active learning is the choice of the acquisition function. The acquisition functions can be divided into two families even though works combine both approaches [46]. The first family follows the "uncertainty sampling" principle [47, 48, 39], where the acquisition function tries to choose the samples where the model is the least confident in its predictions; the acquisition function mimics the training loss. The second family of acquisition functions is based on "diversity sampling" [49], this time, the acquisition function tries to select samples that represent the diversity of the unlabeled dataset and tries to avoid selecting samples too similar.

For works related to our task, Yoo et al. [48] compute the spatial entropy of body part heatmaps to assess their quality. SCAF [39] proposes a new acquisition function: the Negative Neighborhood Magnitude which computes the sum of heatmap pixel values in a window centered on the predicted landmark position, and uses it for the face alignment task.

3. Method

3.1. Overview

Noticing the quality of StyleGAN [16] for image generation, we propose to modify its generator so that it generates facial landmarks heatmaps instead. To apply it to real images, we also need to project them into StyleGAN latent space. This leads to our proposed framework: from an input face image, we first approximate its StyleGAN latent vector using a pre-trained Feature-Style encoder [25], then we use the modified StyleGAN generator to predict

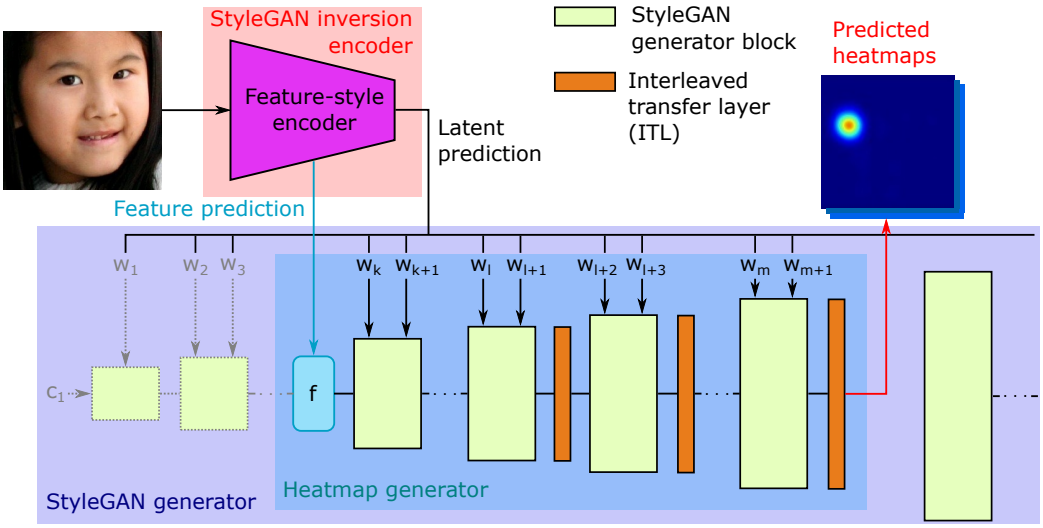

Figure 1: Our network architecture. We use a pre-trained Feature-style encoder and StyleGAN2 generator. Additional convolution layers are interleaved with some generator convolution blocks and are trained to predict landmark heatmaps.

facial landmark heatmaps from this latent vector. The generator is modified in a way similar to 3FabRec [3]: additional convolution layers are interleaved with the generator layers and trained to generate facial landmark heatmaps. Our full architecture can be seen in Fig. 1.

3.2. StyleGAN inversion encoder

We need to choose a StyleGAN inversion model for our architecture. We don't consider optimization-based and hybrid models because their computational time would lead to intractable training time. FFHQ [16], the training dataset of most StyleGAN generators for face images contains only aligned images, the face is always perfectly centered, the eyes are always at the same horizontal position and the zoom is also the same for all images. Many StyleGAN encoders can only faithfully reconstruct images that follow this

alignment. But images from face alignment datasets do not follow this alignment, it depends on the face detector used or the provided ground truth bounding boxes, so we need a StyleGAN encoder that can still reconstruct non-aligned images. That's why we choose Feature-Style encoder [25] which outputs the extended latent vector w^+ but also a feature map f which makes it possible to handle non-aligned images.

3.3. Modified StyleGAN generator

3FabRec [3] and then SCAF [39] showed that the decoder of an auto-encoder trained to reconstruct face images can be modified to predict facial landmark heatmaps. Indeed, during the generation of a face image from a latent representation, generator layers must extract (among other face attributes) face shape information from the latent vector to generate the image. Facial landmark heatmaps can be seen as a face image where only information about face shape has been kept. So, we can try to perform some kind of style transfer by modifying the already trained generator to generate not RGB face images but facial landmark heatmaps instead. While 3FabRec and SCAF use a small inverted ResNet-18 as a decoder, we propose to use a more advanced network: a StyleGAN generator. We detail this architecture in the following paragraphs.

As explained in Section 2.1, the StyleGAN generator starts from a constant vector c_1 followed by StyleGAN blocks which progressively increase the resolution. Each block is composed of an upsampling operation followed by two convolutions equipped with an AdaIn [29] operation to inject the layer-specific style vector. From an image, the Feature-Style encoder predicts 2 codes: the feature code f and the extended latent vector $w^+ = (w_1, w_2, ...)$.

During image generation, the first layers of the generator (five in this paper) are replaced by the feature code f. Then, each latent vector w_i (except for the first ones) from w^+ is transformed into a style vector by an affine transformation and injected into the corresponding StyleGAN layer through AdaIN. The last block outputs the reconstructed image.

We modify this architecture to predict facial landmark heatmaps. From a chosen generator block until the landmark heatmap resolution is reached, in a similar way to 3FabRec and SCAF, we interleave the generator blocks with Interleaved Transfer Layers (ITLs) [3]. An ITL is a 3×3 convolutional layer located directly after a StyleGAN block. It takes as input the output of the StyleGAN block and outputs a feature map with the same spatial dimensions and number of channels of the input, except for the last ITL. This last ITL generates the landmark heatmaps so this time the number of channels is equal to the number of landmarks. Fig. 1 shows the full architecture. During the supervised training, the StyleGAN blocks parameters are frozen, but not the encoder ones which are fine-tuned while the ITLs are trained.

3.4. Active learning

Optionally, we use active learning to select our samples when training with limited data. We use the Negative Neighborhood Magnitude (NNM) [39] as the acquisition function. For each predicted heatmap $\tilde{H}$, the NNM sums the heatmap pixels values in a square window W_i centered on the predicted landmark position $\tilde{l_i}$ (the argmax of the heatmap in our case). To finish the NNM computation, all the heatmaps sums are then summed again and we take negative so that the less confident the model is, the greater NNM is.

$$NNM(\tilde{H}) = -\sum_{i=1}^{L} \sum_{u,v \in W_i} \tilde{H}_i(u,v) \tag{1}$$

This acquisition function selects heatmaps with low magnitude around the predicted landmark positions, it assumes that when the model is not confident about its prediction, the heatmap magnitude will be low so adding this sample to the training set would improve the model performance. As in SCAF [39], we discard the top-10% NNM samples from the potential candidates to avoid selecting outliers.

4. Experiments

4.1. 2D Face alignment datasets

We evaluated our method on three 2D face alignment datasets commonly used in the literature.

AFLW: This dataset [50] contains 24,386 face images annotated with 21 landmarks. Following usual practice [38, 3], we ignore the landmarks of the ears and use 20,000 training images and 4,386 testing images. We evaluate our model on the *Full* test set and the *Frontal* test set, a subset that contains only images with a frontal view.

300-W: This dataset [44] contains face images annotated using 68 landmarks. Following the common splits [38, 3], the training set contains 3,148 images while the *Full* test set contains 689 images and is divided into a *Common* test set of 554 images and a *Challenging* test set of 135 images.

WFLW: This dataset introduced in [45] contains 7,500 training images and 2,500 testing images annotated with 98 landmarks. The test set is split

into several (partially overlapping) subsets, each focused on a specific characteristic: pose, expression, illumination, make-up, occlusion, or blur.

4.2. 3D Face alignment datasets

We also evaluated our architecture on the 3D face alignment task to see if it can also generate accurate 3D facial landmark heatmaps. For this task, we use two different datasets, one for training and another for evaluation. By 3D landmarks we actually mean the 2D projections of the 3D landmarks because to predict the full 3D landmarks we would need to modify our architecture.

300-W-LP: This is a synthetic dataset [33] created from 300-W images using the profiling method of [33] to render its faces into larger poses. This dataset contains 122,450 face images with a face pose yaw angle ranging from -90° to 90°. 68 2D and 3D landmarks annotations are provided for each face. We train our model on this dataset to predict the 3D landmarks.

AFLW2000-3D: This dataset [33] was constructed by re-annotating the first 2,000 images of AFLW annotated with 68 3D landmarks consistent with the ones of 300-W-LP. The face pose also ranges from -90° to 90°. We use this dataset to evaluate our models trained on 300-W-LP. The dataset can be divided into 3 subsets according to the absolute face pose: a subset with almost frontal faces ([0°, 30°]), another one with medium poses ([30°, 60°]) and the last one with profile views ([60°, 90°]).

4.3. Evaluation

As metrics, we use the usual Normalized Mean Error (NME) and Area Under Curve (AUC) of the Cumulative Error Distribution (CED) to compare our model to other methods. The NME is defined as:

$$NME(\%) = \frac{1}{N}\sum_{i=1}^{N}\frac{||s_i - \tilde{s}_i||}{d} * 100 \quad (2)$$

where s_i and $\tilde{s}_i$ are the ground truth and predicted location of landmark i, N the number of landmarks and d a normalization distance.

For 300-W and WFLW, we use the distance between the outer eye corners as the normalization distance for the NME ($\text{NME}_{\text{inter-ocular}}$). For AFLW, because of the large number of profile faces, we report both the NME normalized with the diagonal of the ground truth bounding box (NME_{diag}) or the square root of the ground truth bounding box area ($\sqrt{w_{bbox} * h_{bbox}}$) ($\text{NME}_{\text{box}}$). We also report the AUC at 7% NME_{box} ($\text{AUC}^{7}_{\text{box}}$) for this dataset. For AFLW2000-3D, as the normalization distance, we also use the square root of the bounding box area. Because no bounding box is provided for this dataset, it is computed from the ground truth 3D landmarks.

4.4. Architecture and training parameters

The StyleGAN inversion encoder is a Feature-style encoder [25] pretrained on FFHQ [16]. The generator is a StyleGAN2 generator [17] pretrained on the same dataset. Input images are resized to 256×256 pixels. We use 5 ITLs (see Section 4.7.2 for details) and output heatmaps of size 128×128 pixels.

Our models are implemented with PyTorch [51]. We train all our models for 200,000 training steps using a batch size of 8 on a Nvidia V100 GPU with 16 GB of memory. For the training loss, we use a standard MSE loss between the predicted and ground truth heatmaps. As optimizer, we use Adam [52] ($\beta_1 = 0.9$, $\beta_2 = 0.99$) with an initial learning rate of 0.0001

for the ITLs and 0.00002 for the pre-trained encoder. Both learning rates are decayed by a factor of 0.995 every 10 epochs. We use random vertical flip (p=50%), rotation (±30°), translation (±4%), scaling (±5%), occlusions (p=50%, PyTorch default settings for the bounding box size), Gaussian blur (p=20%), brightness (p=45%, ±80%) and contrast changes (p=45%, 30%-200%) as data augmentations.

When training with limited data (without active learning), the training samples are chosen randomly from among the full training set before each run. We report the means and standard deviations over 5 runs for all training sizes except 50, for which we use 10 runs. During active learning, the initial training set contains 10 random samples. K, the number of samples added after training depends on the final training set size. We make it large enough so that there are no more than 5 trainings in total.

4.5. Results on 2D face alignment

4.5.1. Comparison with fully supervised methods

Table 1 shows comparisons of our method with state-of-the-art (SOTA) on 300-W and WFLW datasets when training with the full training dataset (except for Wood et al. [42] which is trained on a synthetic dataset). Our method achieves results comparable to the SOTA of 2019 but is surpassed by current SOTA methods. For AFLW (see Table 2), we are second on all metrics, just behind another semi-supervised method [8]. We hypothesize that these better results, compared to 300-W and WFLW are caused by the fact that AFLW is relatively easier (only 19 landmarks) and many images are closer to the ones found in FFHQ, the training dataset for the StyleGAN encoder and generator. While these results are interesting, our main goal

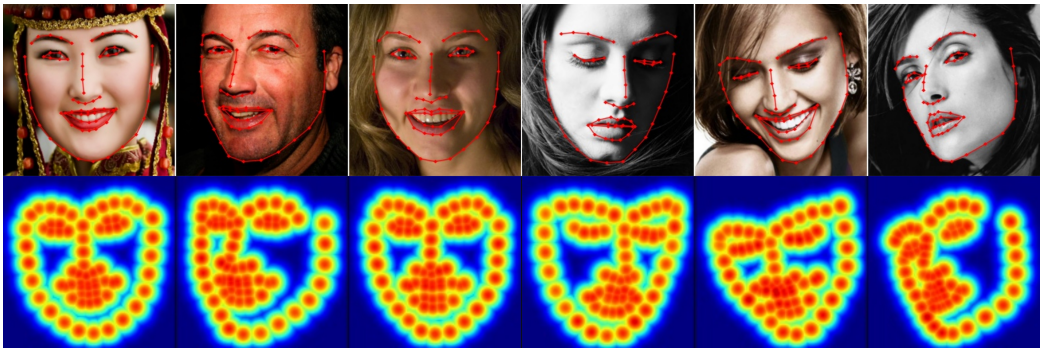

Figure 2: Our model predictions on some images of 300W Full test set. Top row shows landmark predictions and bottom row shows heatmap predictions.

with our method is to train the model with limited training. These results are discussed in the next subsection. Figure 2 shows predictions of our model on some images of the 300-W Full test.

4.5.2. Comparison with semi-supervised methods

We compare our method to other semi-supervised methods training with limited data. On 300-W (Table 4), we surpass other methods for all limited training sizes ranging from 20% (630 samples) of the full training set (3,148 samples in total) to only 50 samples. When training with this last training set size, we get comparable or better results than other methods training with 20% of the dataset. Also, the performance of our algorithm does not degrade significantly when the training set size goes down, even for 50 samples. For the more challenging dataset WFLW (see Table 3), we also surpass other methods when training on limited data.

Methods reported their results on AFLW with the NME_{diag} or NME_{box} so we computed both metrics (see Table 5). For the NME_{diag}, we surpass other semi-supervised methods on all training set sizes on both the Full and

	300-W			WFLW
Method	Com.	Chal.	Full	Full
SAN [53]	3.34	6.60	3.98	5.22
LAB [45]	2.98	5.19	3.49	5.27
AVS [41]	3.21	6.49	3.86	4.39
DeCaFa [54]	2.93	5.26	3.39	4.62
AWing [26]	2.72	4.52	3.07	4.36
LUVLi [55]	2.76	5.16	3.23	4.37
HiH [56]	2.93	5.00	3.36	4.18
SHR-FAN [57]	2.61	**4.13**	2.94	**3.72**
ADNet [27]	**2.53**	4.58	**2.93**	4.14
FaRL [8]	2.56	4.45	**2.93**	3.96
Wood et al.* [42]	3.03	4.80	3.38	-
Ours	2.97	5.30	3.42	4.62
Ours (standard deviations)	0.01	0.09	0.02	0.03

Table 1: $\text{NME}_{\text{inter-ocular}}$ (%) (↓) on the 300-W Common, Challenging and Full test sets, and on the WFLW Full test set. *Trained on a synthetic dataset.

AFLW dataset				
	NME_{diag} (%)↓		NME_{box} (%)↓	$\text{AUC}^{7}_{\text{box}}$ ↑
Method	Full	Frontal	Full	Full
DSRN [58]	1.86	-	-	-
SAN [53]	1.91	1.85	4.04	0.540
LAB [45]	1.25	1.14	-	-
HR-Net [59]	1.57	1.46	-	-
LUVLi [55]	1.39	1.19	2.28	0.680
3FabRec [3]	-	-	1.84	-
SHR-FAN [57]	1.31	1.19	2.14	0.700
FaRL [8]	**0.94**	**0.82**	**1.33**	**0.813**
Ours	1.02	0.90	1.45	0.791
Ours (standard deviations)	<0.01	<0.01	<0.01	<0.001

Table 2: Comparison with state-of-the-art methods on the AFLW Full and Frontal test sets.

WFLW dataset					
Method	Training set size				
	7500 (100%)	1500 (20%)	750 (10%)	375 (5%)	50 (0.67%)
AVS [41]	**4.39**	6.00	7.20	-	-
3FabRec [3]	5.62	6.51	6.73	7.68	8.39
SCAF [39]	5.50	6.07	6.28	6.72	8.06
SCAF+AL [39]	-	-	6.24	6.59	7.60
Ours	4.62	5.09	5.44	5.80	7.78
Ours std	0.03	0.08	0.07	0.08	0.20
Ours+AL	-	**4.94**	**5.18**	**5.45**	**7.30**
Ours+AL std	-	0.04	0.03	0.04	0.12

Table 3: $\text{NME}_{\text{inter-ocular}}$ (%) (↓) when training with limited training set size on the WFLW Full test set. AL stands for Active Learning.

Frontal test sets. For the NME_{box}, we compared our method with FaRL [8], they only report results for 100%, 10%, and 1% sizes but we can see that even though they achieve better results when training with the full training set, their results are even with us on the 10% training set size and we surpass them on the 1% training set size.

4.5.3. Results with active learning

Figure 3 shows examples of images selected by the Negative Neighborhood Magnitude (NNM), the active learning acquisition function used. Heatmaps with low magnitude have various causes: network confusion between the outline of the face and hair or beard (first two pictures), unusual face pose (third picture), or almost closed eyes (fourth picture). Also, landmarks on the outline of the face being the most ambiguous, they usually are the ones with the lowest heatmap magnitude.

300-W dataset															
Method	Training set size														
	3148 (100%)			630 (20%)			315 (10%)			168 (5%)			50 (1.59%)		
	Com.	Ch.	Full	Com.	Ch.	Full	Com.	Ch.	Full	Com.	Ch.	Full	Com.	Ch.	Full
RCN+ [37]	3.00	**4.98**	3.46	-	6.12	4.15	-	6.63	4.47	-	9.95	5.11	-	-	-
AVS [41]	3.21	6.49	3.86	3.85	-	-	4.27	-	-	6.32	-	-	-	-	-
TS^3 [38]	**2.91**	5.90	3.49	4.31	7.97	5.03	4.67	9.26	5.64	-	-	-	-	-	-
3FabRec [3]	3.36	5.74	3.82	3.76	6.53	4.31	3.88	6.88	4.47	4.22	6.95	4.75	4.55	7.39	5.10
SCAF [39]	3.48	5.89	3.95	3.66	6.23	4.17	3.87	6.60	4.40	3.93	6.84	4.50	4.33	7.60	4.97
SCAF+AL [39]	-	-	-	-	-	-	3.99	6.49	4.48	4.19	6.78	4.70	4.29	6.93	4.81
Ours	2.97	5.30	**3.42**	**3.14**	5.66	3.64	**3.22**	5.87	3.74	**3.33**	6.05	3.86	**3.57**	6.62	4.16
Ours std	0.01	0.09	0.02	0.02	0.04	0.02	0.03	0.06	0.03	0.03	0.13	0.03	0.05	0.21	0.07
Ours+AL	-	-	-	**3.12**	**5.53**	**3.59**	**3.20**	**5.67**	**3.68**	**3.32**	**5.83**	**3.81**	**3.54**	**6.24**	**4.06**
Ours+AL std	-	-	-	0.03	0.05	0.02	0.02	0.05	0.02	0.03	0.04	0.03	0.01	0.13	0.02

Table 4: $\text{NME}_{\text{inter-ocular}}$ (%) (↓) when training with limited training set size on 300-W on the Common, Challenging and Full test sets (first, second and third columns respectively for each training set size). AL stands for Active Learning.

AFLW dataset												
Method	Training set size											
	20000 (100%)		4000 (20%)		2000 (10%)		1000 (5%)		200 (1%)		50 (0.25%)	
	Full	Fr.	Full	Fr.	Full	Fr.	Full	Fr.	Full	Fr.	Full	Fr.
NME_{box} (%) ↓												
RCN+ [37]	1.61	-	-	-	-	-	2.17	-	2.88	-	-	-
TS^3 [38]	-	-	1.99	1.86	2.14	1.94	2.19	2.03	-	-	-	-
3FabRec [3]	1.87	1.59	1.96	1.74	2.03	1.74	2.13	1.86	2.38	2.03	2.74	2.23
Ours	**1.45**	**1.28**	**1.60**	**1.39**	**1.63**	**1.41**	**1.66**	**1.43**	1.79	**1.53**	**2.05**	**1.71**
Ours std	<.01	<.01	<.01	0.02	<.01	0.01	<.01	0.03	0.01	0.01	0.03	0.02
Ours+AL	-	-	-	-	-	-	**1.66**	1.49	**1.77**	1.56	**2.03**	**1.75**
Ours+AL std	-	-	-	-	-	-	<.01	0.02	<.01	<.01	0.03	0.04
NME_{diag} (%) ↓												
FaRL [8]	**0.94**	**0.82**	-	-	**1.15**	-	-	-	1.35	-	-	-
Ours	1.02	0.90	**1.13**	**0.98**	**1.15**	**1.00**	**1.17**	**1.01**	1.27	**1.08**	**1.45**	**1.21**
Ours std	<.01	<.01	<.01	0.01	<.01	0.01	<.01	0.02	0.01	<.01	0.02	0.02
Ours+AL	-	-	-	-	-	-	**1.17**	1.05	**1.25**	1.11	**1.44**	**1.24**
Ours+AL std	-	-	-	-	-	-	<.01	0.02	<.01	<.01	0.02	0.03

Table 5: Comparison with other semi-supervised methods when training with limited training set size on AFLW on the Full and Frontal test sets (first and second columns respectively for each training set size). AL stands for Active Learning

On 300-W, while the use of active learning greatly reduces the NME on the Challenging test, especially when the training size is very small, there is no real improvement on the Common test set. It proves the efficiency of active learning to select hard samples from the unlabeled training set, similar to the ones found in the Challenging test set.

For WFLW, using active learning always improves the performance of our model, proving again its efficiency. One should also notice that for both datasets, the use of active learning also reduces the NME standard deviations; that would tend to demonstrate that the samples selected by active learning are similar across the runs.

On AFLW, this time the active learning does not improve much the performance on the Full test set. It decreases it a bit on the Frontal test set, meaning active learning improves the performance on challenging images such as profile faces but with the cost of a slightly reduced accuracy on easy faces.

4.6. Results on 3D face alignment

We compare our model to other methods which train on 300-W-LP and evaluate on the AFLW2000-3D dataset. Although our method only predicts landmark heatmaps and does rely on estimating a 3D face model like most of the recent 3D face alignment models [34, 35, 36], our method still obtains decent results. For the low pose test subset ($[0°, 30°]$), we obtain a similar performance to the state-of-the-art. However, for larger poses, our model falls behind recent methods. This may be explained by the fact there is no profile face in the pre-training dataset (FFHQ [16]) and because our method relies on heatmaps, it may be difficult for some images to detect precisely occluded landmarks. Our model still obtains decent results on many profile

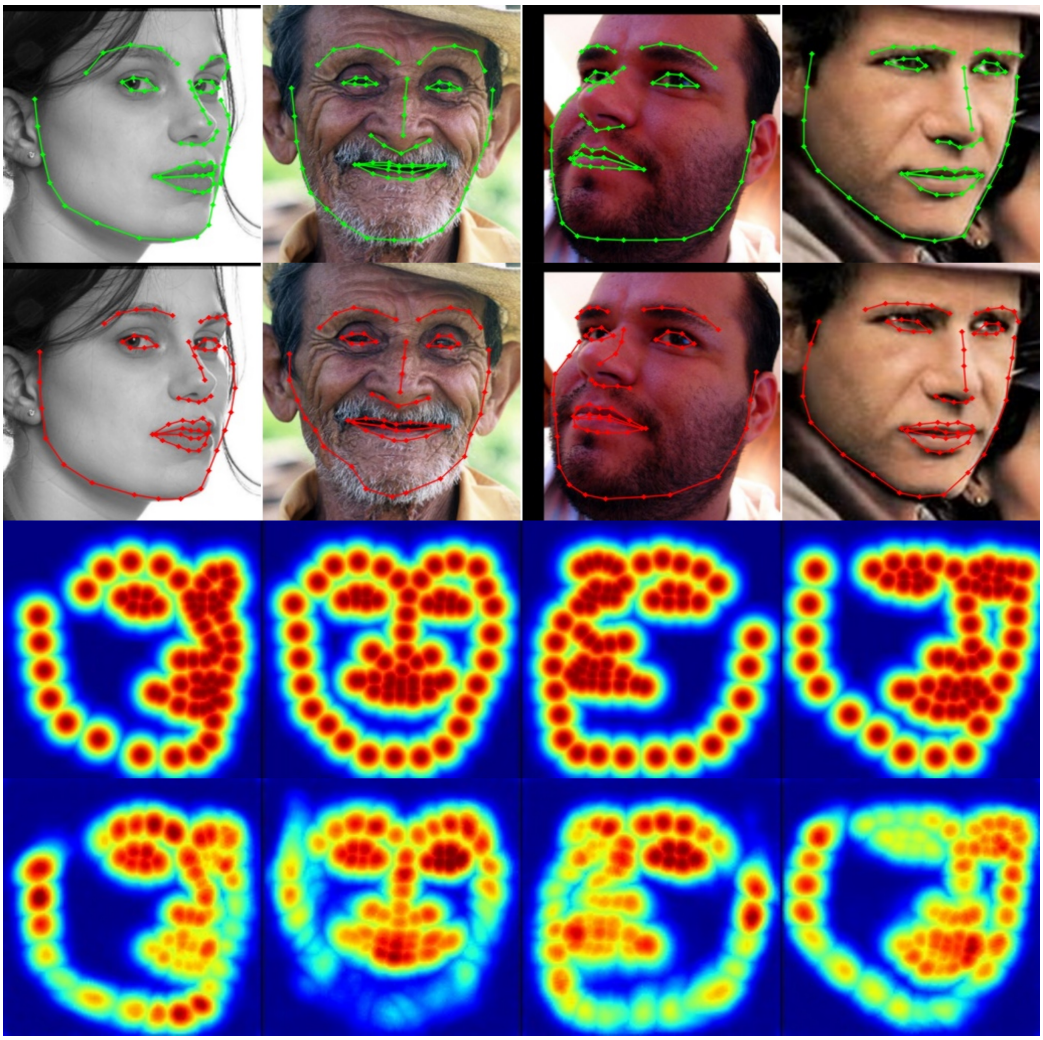

Figure 3: Some images of 300-W selected by the NNM, the active learning acquisition function, after a first training on 10 random images. First row: ground truth landmarks. Second row: predicted landmarks. Third row: ground truth heatmaps. Fourth row: predicted heatmaps. The NNM successfully selects images where some landmark predictions are not accurate. These samples will be added to the training labeled set for the next training iteration and should improve the model performance.

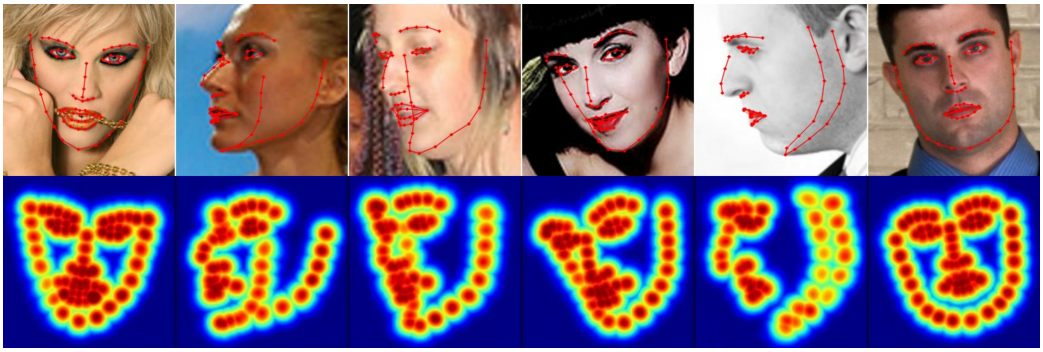

Figure 4: Our model predictions on some images of AFLW2000-3D. The top row shows landmark predictions and the bottom row shows heatmap predictions. Even occluded landmarks are correctly predicted.

face images (see Figure 4).

4.6.1. Results when training with limited training data

We evaluate our model on AFLW2000-3D after training on data sampled from 300-W-LP. We use three different sampling methods, "Random": fully random sampling, "Balanced": random sampling but the low (0-30°), medium (30-60°) and large pose (60-90°) subsets must have equal size and "Active": sampling using active learning (same procedure as used for 2D face alignment). We use different training set sizes, 300, 150, and 48, all divisible by 3 so we can have perfectly balanced training sets for the "Balanced" sampling method. Table 7 reports the NME_{box} of the models on the AFLW2000-3D pose subsets. It also reports the face pose distribution of the training datasets sampled from 300-W-LP. The whole 300-W-LP contains 25% low pose images, 37% medium pose images, and 38% large pose images so our "Random" samplings are close to this distribution. "Balanced" samplings have by definition a (33%, 33%, 33%) pose distribution. In the

AFLW2000-3D dataset					
Method	0-30°	30-60°	60-90°	Balanced	Mean
3DDFA [33]	3.78	4.54	7.93	5.42	6.03
3D-FAN [32]	3.16	3.53	4.60	3.79	-
3DDFA-PAMI [60]	2.84	3.57	4.96	3.79	-
PRNet [61]	2.75	3.51	4.61	3.62	3.26
3DDFAv2 [34]	**2.63**	3.42	4.48	3.51	-
SADRNet [35]	2.66	**3.30**	4.42	3.46	3.05
SynergyNet [36]	2.65	**3.30**	**4.27**	**3.41**	-
Ours	2.65	3.62	4.89	3.72	3.14
Ours std	0.03	0.07	0.11	0.07	0.03

Table 6: NME_{box} (%) (↓) on different subsets of AFLW2000-3D divided by face pose (yaw angle). "Balanced" column is the average of the first 3 columns. "Mean" column reports the mean NME over the whole AFLW2000-3D dataset.

case of "Active" sampling, the table shows that active learning heavily favors the large pose images when selecting samples. It represents more than half of the final training dataset for all training sizes. Thus, models trained with active learning perform better on AFLW2000-3D large pose subset, especially when training with 48 samples where the NME is decreased by 8% compared to "Random" and 11% compared to "Balanced", but a bit worse on the low pose and medium pose subsets. Also, compared to training with the 122,450 images of the whole 300-W-LP, the performance of the models does not degrade much. For example, there is only a 0.6% increase of NME on the large pose subset of AFLW2000-3D when training with 300 samples (0.25% of 300-W-LP size) with active learning compared to training on the whole dataset.

300-W-LP/AFLW2000-3D									
Sampling	Training set size (300-W-LP)								
	300			150			48		
	0-30°	30-60°	60-90°	0-30°	30-60°	60-90°	0-30°	30-60°	60-90°
	Test NME_{box} (%) ↓ (AFLW2000-3D)								
Random	2.77	**3.73**	5.00	2.90	**3.92**	5.25	3.43	4.73	6.08
std	0.04	0.06	0.09	0.05	0.08	0.11	0.18	0.12	0.21
Balanced	**2.74**	3.76	5.05	**2.86**	3.93	5.25	**3.23**	**4.58**	6.35
std	0.03	0.04	0.06	0.03	0.06	0.09	0.06	0.08	0.27
Active	2.81	3.75	**4.92**	2.98	3.94	**5.09**	3.45	4.70	**5.59**
std	0.03	0.06	0.06	0.05	0.08	0.10	0.05	0.03	0.08
	Training face pose yaw distribution (300-W-LP)								
Random	0.25	0.38	0.37	0.28	0.37	0.35	0.26	0.38	0.38
std	0.01	0.02	0.02	0.02	0.03	0.02	0.03	0.03	0.04
Balanced	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33
std	0	0	0	0	0	0	0	0	0
Active	0.16	0.34	0.50	0.15	0.33	0.52	0.20	0.26	0.54
std	0.01	0.03	0.04	0.02	0.02	0.02	0.05	0.02	0.06

Table 7: Comparison of different sampling methods for several training set sizes according to the face pose yaw angle. This table reports the NME of models trained on some 300W-LP samples and evaluated on the AFLW2000-3D subsets. It also reports the face pose yaw distribution of the training datasets. "Random": training samples are chosen randomly. "Balanced": training samples are random but the low, medium and large face pose subsets must have equal size. "Active": training samples are selected using active learning.

4.7. Ablation studies

4.7.1. Encoder fine-tuning

Table 8 reports results on the 300-W and WFLW Full test sets for different training set sizes whether we fine-tune or not the encoder while training the ITLs. Fine-tuning improves the performance for all training set sizes on both datasets, especially on WFLW.

4.7.2. Number of Interleaved Transfer Layers

We made experiments to know if there is an optimal number of Interleaved Transfer Layers (ITLs), results are reported in Table 9. When training with the full training set, using only one layer makes the model perform a bit worse (also confirmed by Wilcoxon's signed tests not reported in this table). From

	300-W			**WFLW**		
	Training set size			Training set size		
	3148 (100%)	315 (10%)	50 (1.59%)	7500 (100%)	750 (10%)	50 (0.67%)
w/o fine-tuning	4.54	4.86	5.99	8.94	9.42	14.36
std	0.02	0.03	0.06	0.08	0.06	0.33
w/ fine-tuning	3.42	3.74	4.16	4.62	5.44	7.78
std	0.02	0.03	0.07	0.03	0.08	0.20

Table 8: $\text{NME}_{\text{inter-ocular}}$ (%) (↓) on 300-W and WFLW Full test sets for different training set sizes with and without encoder fine-tuning.

Num. ITLs	**300-W**			**WFLW**		
	Training set size			Training set size		
	3148 (100%)	315 (10%)	50 (1.59%)	7500 (100%)	750 (10%)	50 (0.67%)
1	3.54	3.76	4.26	4.75	5.68	8.72
std	0.03	0.02	0.09	0.04	0.11	0.35
2	3.44	3.72	4.22	4.65	5.50	8.20
std	0.04	0.03	0.03	0.02	0.04	0.32
3	3.43	3.74	4.13	4.66	5.46	7.65
std	0.04	0.01	0.02	0.08	0.05	0.23
4	3.46	3.73	4.16	4.63	5.43	7.94
std	0.04	0.02	0.03	0.05	0.04	0.27
5	3.42	3.74	4.16	4.62	5.44	7.78
std	0.02	0.03	0.07	0.03	0.07	0.20
6	3.45	3.70	4.15	4.61	5.42	7.82
std	0.03	0.01	0.06	0.4	0.03	0.29

Table 9: $\text{NME}_{\text{inter-ocular}}$ (%) (↓) on 300-W and WFLW Full test sets, depending on the number of Interleaved Transfer Layers (ITLs) and the training set size.

2 to 6 ITLs, the NMEs are very close and most of the time the Wilcoxon's test null hypothesis can't be rejected when comparing two models. For small training set sizes, such as 50 samples, we need at least 3 layers to have the best performance. We suppose that when training with limited data, the encoder can't be totally fine-tuned because of its large number of parameters compared to ITLs so more ITLs make the training easier. However, from 3 to 6 layers the standard deviations are overlapping, and again the Wilcoxon's tests can't tell apart the models, so there is no clear winner. For other experiments in this paper, we used 5 ITLs.

5. Conclusion

In this paper, we have demonstrated that StyleGAN [16] can be used, not only for generative tasks such as attribute edition or single attribute regression like age estimation but also for complex discriminative tasks like face alignment, even when the training data is limited. Our method can generate accurate facial landmark heatmaps and obtains competitive results on multiple face alignment datasets for both 2D and 3D facial landmarks. It also systematically surpasses other semi-supervised methods when training with limited data. Another advantage of our method is that we don't need to perform any dedicated computationally expensive unsupervised training on large databases [3, 39, 8], prior to the supervised training because our model is based on StyleGAN and pre-trained weights are already available.

5.1. Future work

Thanks to the Feature-Style encoder [25], we were able to perform our training and testing on unaligned images which do not belong to the original StyleGAN generative distribution. An interesting work would be to align face alignment dataset images to make them follow FFHQ alignment. Would this improve performance because images would lie closer to the original generative distribution, or make it worse because of less face pose diversity during training, is an open question. Also, because we were more interested in proving the efficiency of our method rather than beating the state-of-the-art, we didn't use any specific trick used in recent face alignment models, such as facial boundaries heatmaps [45], complex training loss [26, 27], or attention maps [35] to boost our model. They could be easily applied to our

method to improve its performance.

Acknowledgments

This work was granted access to the HPC resources of IDRIS under the allocation 2022-AD011012376 made by GENCI.

About the authors

Martin Dornier received his engineering degree from IMT Atlantique, France, in 2018. He is currently working as a Ph.D. student at InterDigital, France, in partnership with Univ. Rennes, CNRS, IRISA, France. His Ph.D. focuses on training neural networks with limited annotated data.

Philippe-Henri Gosselin received a Ph.D. in image and signal processing in 2005 and joined the ETIS Lab as an assistant professor in 2007. In 2012, he was promoted to Full Professor. He left academics in 2018 to join Technicolor as a Principal Scientist. His research focuses on computer vision.

Christian Raymond received a Ph.D. in computer science from the University of Avignon, France, in 2005. He was appointed in 2009 as an Associate Professor at INSA Rennes, France, and joined the IRISA Lab. His research focuses on speech understanding, machine learning for natural language processing, and data-driven stochastic approaches.

Yann Ricquebourg received his Ph.D. in computer science from the Université de Rennes 1, France, in 1997. He obtained a position as an Assistant Professor at Université de Bretagne Occidentale, France, and then ENSAI, France, before joining INSA Rennes and IRISA. His research focuses on document analysis and gesture recognition.

Bertrand Coüasnon received his Ph.D. in computer science from the University of Rennes 1, France, in 1996. Since 1996 he is an Associate Professor at INSA Rennes, France. His research interests are knowledge formalization, generic methods for document structure recognition, combining deep learning and syntactical methods, and handwriting recognition.

References

[1] D. G. Lowe, Distinctive image features from scale-invariant keypoints, International journal of computer vision 60 (2) (2004) 91–110.

[2] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), Vol. 1, 2005, pp. 886–893 vol. 1. doi:10.1109/CVPR.2005.177.

[3] B. Browatzki, C. Wallraven, 3fabrec: Fast few-shot face alignment by reconstruction, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6110–6120.

[4] I. Higgins, L. Matthey, A. Pal, C. P. Burgess, X. Glorot, M. M. Botvinick, S. Mohamed, A. Lerchner, beta-vae: Learning basic visual concepts with a constrained variational framework, in: International Conference on Learning Representations, 2017.

[5] H. Kim, A. Mnih, Disentangling by factorising, in: International Conference on Machine Learning, PMLR, 2018, pp. 2649–2658.

[6] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, A. Joulin, Unsupervised learning of visual features by contrasting cluster assignments, Advances in Neural Information Processing Systems 33 (2020) 9912–9924.

[7] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: International conference on machine learning, PMLR, 2020, pp. 1597–1607.

[8] Y. Zheng, H. Yang, T. Zhang, J. Bao, D. Chen, Y. Huang, L. Yuan, D. Chen, M. Zeng, F. Wen, General facial representation learning in a visual-linguistic manner, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 18697–18709.

[9] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American

Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.
URL https://aclanthology.org/N19-1423

[10] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick, Masked autoencoders are scalable vision learners, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16000–16009.

[11] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, P. Abbeel, Infogan: Interpretable representation learning by information maximizing generative adversarial nets, in: D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, R. Garnett (Eds.), Advances in Neural Information Processing Systems 29, Curran Associates, Inc., 2016, pp. 2172–2180.

[12] J. Donahue, P. Krähenbühl, T. Darrell, Adversarial feature learning, International Conference on Learning Representations (2016).

[13] V. Dumoulin, I. Belghazi, B. Poole, O. Mastropietro, A. Lamb, M. Arjovsky, A. Courville, Adversarially learned inference, International Conference on Learning Representations (2017).

[14] Y. Xu, Y. Shen, J. Zhu, C. Yang, B. Zhou, Generative hierarchical features from synthesizing images, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 4432–4442.

[15] Y. Nitzan, R. Gal, O. Brenner, D. Cohen-Or, Large: Latent-based regression through gan semantics, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 19239–19249.

[16] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2019, pp. 4401–4410.

[17] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, T. Aila, Analyzing and improving the image quality of stylegan, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 8110–8119.

[18] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, T. Aila, Alias-free generative adversarial networks, Advances in Neural Information Processing Systems 34 (2021).

[19] Y. Alaluf, O. Patashnik, D. Cohen-Or, Restyle: A residual-based stylegan encoder via iterative refinement, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 6711–6720.

[20] J. Gu, Y. Shen, B. Zhou, Image processing using multi-code gan prior, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 3012–3021.

[21] K. Kang, S. Kim, S. Cho, Gan inversion for out-of-range images with geometric transformations, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 13941–13949.

[22] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, D. Cohen-Or, Encoding in style: a stylegan encoder for image-to-image translation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2287–2296.

[23] O. Tov, Y. Alaluf, Y. Nitzan, O. Patashnik, D. Cohen-Or, Designing an encoder for stylegan image manipulation, ACM Transactions on Graphics (TOG) 40 (4) (2021) 1–14.

[24] T. Wang, Y. Zhang, Y. Fan, J. Wang, Q. Chen, High-fidelity gan inversion for image attribute editing, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11379–11388.

[25] X. Yao, A. Newson, Y. Gousseau, P. Hellier, Feature-style encoder for style-based gan inversion, arXiv e-prints (2022) arXiv–2202.

[26] X. Wang, L. Bo, L. Fuxin, Adaptive wing loss for robust face alignment via heatmap regression, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 6971–6981.

[27] Y. Huang, H. Yang, C. Li, J. Kim, F. Wei, Adnet: Leveraging error-bias towards normal direction in face alignment, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 3080–3090.

[28] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, K. Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 27, Curran Associates, Inc., 2014, pp. 2672–2680.

[29] X. Huang, S. Belongie, Arbitrary style transfer in real-time with adaptive instance normalization, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 1501–1510.

[30] T. Karras, T. Aila, S. Laine, J. Lehtinen, Progressive growing of gans for improved quality, stability, and variation, International Conference on Learning Representations (2018).

[31] A. Newell, K. Yang, J. Deng, Stacked hourglass networks for human pose estimation, in: European conference on computer vision, Springer, 2016, pp. 483–499.

[32] A. Bulat, G. Tzimiropoulos, How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks), in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1021–1030.

[33] X. Zhu, Z. Lei, X. Liu, H. Shi, S. Z. Li, Face alignment across large poses: A 3d solution, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 146–155.

[34] J. Guo, X. Zhu, Y. Yang, F. Yang, Z. Lei, S. Z. Li, Towards fast, accurate and stable 3d dense face alignment, arXiv preprint arXiv:2009.09960 (2020).

[35] Z. Ruan, C. Zou, L. Wu, G. Wu, L. Wang, Sadrnet: Self-aligned dual face regression networks for robust 3d dense face alignment and reconstruction, IEEE Transactions on Image Processing 30 (2021) 5793–5806.

[36] C.-Y. Wu, Q. Xu, U. Neumann, Synergy between 3dmm and 3d landmarks for accurate 3d facial geometry, in: 2021 International Conference on 3D Vision (3DV), IEEE, 2021, pp. 453–463.

[37] S. Honari, P. Molchanov, S. Tyree, P. Vincent, C. Pal, J. Kautz, Improving landmark localization with semi-supervised learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1546–1555.

[38] X. Dong, Y. Yang, Teacher supervises students how to learn from partially labeled images for facial landmark detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 783–792.

[39] M. Dornier, P.-H. Gosselin, C. Raymond, Y. Ricquebourg, B. Coüasnon, Scaf: Skip-connections in auto-encoder for face alignment with few annotated data, in: International Conference on Image Analysis and Processing, Springer, 2022, pp. 425–437.

[40] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, ICLR (2021).

[41] S. Qian, K. Sun, W. Wu, C. Qian, J. Jia, Aggregation via separation: Boosting facial landmark detector with semi-supervised style translation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 10153–10163.

[42] E. Wood, T. Baltrusaitis, C. Hewitt, M. Johnson, J. Shen, N. Milosavljevic, D. Wilde, S. Garbin, C. Raman, J. Shotton, T. Sharp, I. Stojiljkovic, T. Cashman, J. Valentin, 3d face reconstruction with dense landmarks (2022). doi:10.48550/ARXIV.2204.02776.
URL `https://arxiv.org/abs/2204.02776`

[43] E. Wood, T. Baltrušaitis, C. Hewitt, S. Dziadzio, M. Johnson, V. Estellers, T. J. Cashman, J. Shotton, Fake it till you make it: Face analysis in the wild using synthetic data alone (2021). arXiv:2109.15102.

[44] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, M. Pantic, 300 faces in-the-wild challenge: The first facial landmark localization challenge, in: Proceedings of the IEEE international conference on computer vision workshops, 2013, pp. 397–403.

[45] W. Wu, C. Qian, S. Yang, Q. Wang, Y. Cai, Q. Zhou, Look at boundary: A boundary-aware face alignment algorithm, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 2129–2138.

[46] A. Kirsch, J. Van Amersfoort, Y. Gal, Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning, Advances in neural information processing systems 32 (2019) 7026–7037.

[47] Y. Gal, R. Islam, Z. Ghahramani, Deep bayesian active learning with image data, in: International Conference on Machine Learning, PMLR, 2017, pp. 1183–1192.

[48] D. Yoo, I. S. Kweon, Learning loss for active learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 93–102.

[49] O. Sener, S. Savarese, Active learning for convolutional neural networks: A core-set approach, arXiv preprint arXiv:1708.00489 (2017).

[50] M. Koestinger, P. Wohlhart, P. M. Roth, H. Bischof, Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization, in: 2011 IEEE international conference on computer vision workshops (ICCV workshops), IEEE, 2011, pp. 2144–2151.

[51] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: An imperative style, high-performance deep learning library, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems 32, Curran Associates, Inc., 2019, pp. 8024–8035.

[52] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

[53] X. Dong, Y. Yan, W. Ouyang, Y. Yang, Style aggregated network for facial landmark detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 379–388.

[54] A. Dapogny, K. Bailly, M. Cord, Decafa: Deep convolutional cascade for face alignment in the wild, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6893–6901.

[55] A. Kumar, T. K. Marks, W. Mou, Y. Wang, M. Jones, A. Cherian, T. Koike-Akino, X. Liu, C. Feng, Luvli face alignment: Estimating landmarks' location, uncertainty, and visibility likelihood, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8236–8246.

[56] X. Lan, Q. Hu, J. Cheng, Revisiting quantization error in face alignment, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 1521–1530.

[57] A. Bulat, E. Sanchez, G. Tzimiropoulos, Subpixel heatmap regression for facial landmark localization, in: Proceedings of the British Machine Vision Conference (BMVC), 2021.

[58] X. Miao, X. Zhen, X. Liu, C. Deng, V. Athitsos, H. Huang, Direct shape regression networks for end-to-end face alignment, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5040–5049.

[59] K. Sun, Y. Zhao, B. Jiang, T. Cheng, B. Xiao, D. Liu, Y. Mu, X. Wang, W. Liu, J. Wang, High-resolution representations for labeling pixels and regions, CoRR abs/1904.04514 (2019).

[60] X. Zhu, X. Liu, Z. Lei, S. Z. Li, Face alignment in full pose range: A 3d total solution, IEEE transactions on pattern analysis and machine intelligence 41 (1) (2017) 78–92.

[61] Y. Feng, F. Wu, X. Shao, Y. Wang, X. Zhou, Joint 3d face reconstruction and dense alignment with position map regression network, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 534–551.