Exploring StyleGAN Latent Space for Face Alignment with Limited Training Data

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

Exploring StyleGAN Latent Space for Face Alignment with Limited Training Data

Martin Dornier*[1,2], Philippe-Henri Gosselin†[1], Christian Raymond‡[2], Yann Ricquebourg§[2], and Bertrand Coüasnon¶[2]

[1]InterDigital
[2]Univ Rennes, CNRS, IRISA, France

Abstract

With deep learning models growing in size over the years, sometimes exceeding a billion parameters now, the need for large, annotated training datasets grows too. To alleviate this problem, the interest in self-supervised learning is also increasing. In this domain, with the rise of Generative Adversarial Networks (GANs) and particularly StyleGAN, the quality of image generation is significantly improving. In this paper, we propose to use StyleGAN to perform face alignment with limited training data instead of image generation. Our proposed framework Face Alignment using StyleGAN Embeddings (FASE) projects real images into StyleGAN latent space and then predicts facial landmarks from the latent vectors. Our method achieves state-of-the-art on multiple face alignment datasets in the few-shot setting.

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 HOG [8] or SIFT [33]. 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 main application of representation learning is reducing the number of annotated samples needed to train a model on a downstream task. Popular representation learning methods make use of auto-encoders [19, 27, 3], Generative Adversarial Networks (GANs) [7, 12, 16, 49], contrastive learning [6, 5, 52] or masked inputs [11, 18].

*martin.dornier@interdigital.com
†philippehenri.gosselin@interdigital.com
‡christian.raymond@irisa.fr
§yann.ricquebourg@irisa.fr
¶bertrand.couasnon@irisa.fr

Among GANs, StyleGAN is a popular approach thanks to its disentangled latent space that can be distilled into multiple styles [25, 26, 24]. Each style controls a semantic characteristic of the image such as color scheme, object shape or background. Because of its lack of 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 [17, 41, 45, 46, 1, 23, 49, 51, 38]. However, most of these works focus on image-to-image translation tasks, such as attribute editability, super-resolution or inpainting, and only a few experiment on StyleGAN latent space for discriminative tasks such as expression recognition, age estimation or facial landmark detection [49, 38].

In this work, we study the use of StyleGAN for the facial landmark detection task (also called face alignment) with limited training data. Indeed, labeling facial keypoints is time-consuming and can be challenging on images with large poses or occlusions. Thus, 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 on the training dataset. However, is it still possible to train a facial landmark detector only with a few samples and get good accuracy and generalization? To resolve this issue, we propose a new architecture called Face Alignment using StyleGAN Embeddings (FASE). We use a pre-trained StyleGAN encoder to predict StyleGAN latent vectors and modify a StyleGAN generator to predict facial landmarks from these latent vectors.

By only slightly modifying the StyleGAN generator and fine-tuning its encoder, 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 based on StyleGAN to perform face alignment with limited training data.
- We experiment on multiple face alignment datasets to demonstrate the effectiveness of our method.

2. Related Work

2.1. Representation Learning

Representation learning can be divided into 2 categories. Supervised pre-training and self-supervised representation learning. In the supervised setup a network is trained on a specific task with the labels provided. A common example is the training of a network on the ImageNet dataset for object recognition [10]. The network can be used then as a backbone for other vision tasks. However, the features learned during the pre-training might be limited by the size and the diversity of the dataset. Indeed, annotating a dataset is time-consuming even though progress has been made over the years to get large automatically annotated datasets such as JFT-300M [43] or Instagram-1B [34] but with the cost of noisy labels.

On the other hand, self-supervised representation learning does not need annotation, or only weak annotation, and can then use very large amounts of unlabeled samples gathered from the internet. In order to train the network, specific architectures or training losses have to be used. A popular architecture is the auto-encoder [19, 27, 3] where an encoder first projects the sample into latent space and then a decoder is used to reconstruct the sample from the latent representation. Sometimes the decoder is replaced with a GAN [7, 12, 16, 49]. Some other methods follow the contrastive setup [6, 5, 52]: representations of the same object (data augmentations of an image, text/image pair, ...) are pushed closer, while representations of different objects are pushed apart. Recently, methods based on masking some parts of the input and asking the network to retrieve these missing parts have gained interest, first in Natural Language Processing [11] and then in Computer Vision too [18].

2.2. StyleGAN

StyleGAN [25] 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 c and at each layer, w is transformed by an affine transformation (different for each layer) into a style vector and injected into the current feature map via an AdaIn [21] operation.

Each style vector controls a specific aspect of the generated image, style vectors corresponding to low resolutions will control high-level attributes such as, for face images, pose, face shape or general hairstyle while high-resolution style vectors will control more fine-grained aspects such as the color scheme or microstructure.

2.3. StyleGAN Inversion

It is easy to semantically edit a *synthetic* image generated by StyleGAN by modifying some of its style vectors. To modify in the same way a *real* image, we need first to approximate its StyleGAN latent vector, this is called StyleGAN inversion. Methods that try to invert StyleGAN can be divided into three families. The *optimization-based* methods iteratively refine a latent code by minimizing the reconstruction error [17, 23]. The *encoder-based* methods train an encoder to predict the latent code [41, 45, 49, 51, 38, 46]. Finally, *hybrid* methods train an encoder to predict an initial latent code which is refined through optimization [1]. 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 latent code for each style: $w^+ = (w_1, w_2, ..., w_n) \in \mathcal{W}^+$, n being the number of styles [41, 45, 46, 38]. This gives more flexibility and improves the reconstruction error. Some methods go even further, not only do they predict w^+ but also a feature map $f \in \mathcal{F}$ which replaces the first layers of the generator [23, 51]. This feature map improves the reconstruction error but also makes it possible to encode images that do not follow the training datasets alignment (e.g., FFHQ and Celeba-HQ 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.

2.4. Face Alignment

Face alignment (also known as facial landmark detection) is the task of localizing a set of pre-defined facial anatomical keypoints (e.g., tip of the nose corners of the mouth, boundaries of the face, ...). Many downstream applications make use of the predicted landmarks such as face swapping or facial expression recognition.

Instead of directly predicting the positions of the keypoints, most of the current methods use neural networks to predict facial landmark heatmaps [37]. The final landmark position is inferred from the best local maximum.

2.5. Semi-supervised Face Alignment

Semi-supervised methods try to alleviate the problem of facial landmark annotations explained in Section 1. To do so they use annotated but also non annotated data during the training. [20] imposes equivariance of predicted landmarks to geometric transformations. [40] produces multiple images with different styles from an input image. [14] uses pseudo-labeling with a teacher-student method. Some methods are based on representation learning: [3] trains an auto-encoder to reconstruct face images and then modifies its decoder to generate facial landmark heatmaps, [15] improves [3] by adding skip-connections to the auto-encoder and using active learning. [52] uses masked image modeling and image-text contrastive learning on a large text/image pair dataset to pre-train a network then used for different facial downstream tasks including face alignment.

3. Method

3.1. Overview

We are interested in using StyleGAN [25] latent representation to predict facial landmark heatmaps. For this purpose, we propose a new architecture: Face Alignment using StyleGAN Embeddings (FASE). From an input face image, we first approximate its StyleGAN latent vector using a pre-trained Feature-Style encoder [51]. Rather than directly predicting facial landmark positions from the latent vector or styles as in [49] we follow 3FabRec [3] and SCAF [15] principles: we modify a pre-trained StyleGAN2 generator [26] by adding convolution layers interleaved with its convolution blocks and train these additional layers to generate facial landmark heatmaps. Our full architecture can be seen in Fig. 1.

3.2. StyleGAN inversion encoder

We needed to choose a StyleGAN inversion model for our architecture. We don't consider optimization-based and hybrid models because their computational time would make our architecture training too slow. Among the encoder-based methods we chose Feature-Style [51] for its ability to reconstruct not aligned images, such as images of face alignments datasets, by predicting not only an extended latent vector w^+ but also a feature map code f.

3.3. Modified StyleGAN generator

[3] and then [15] 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 a sort of style transfer by modifying the already trained generator to generate not RGB face images but facial landmark heatmaps instead. To do so, as in [3] and [15] we add interleaved transfer layers (ITLs) inside the generator. The architecture is explained in detail in the following paragraphs

As seen in Section 2.2, the StyleGAN generator starts from 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 [21] 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 the image generation, the first layers of the generator 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.

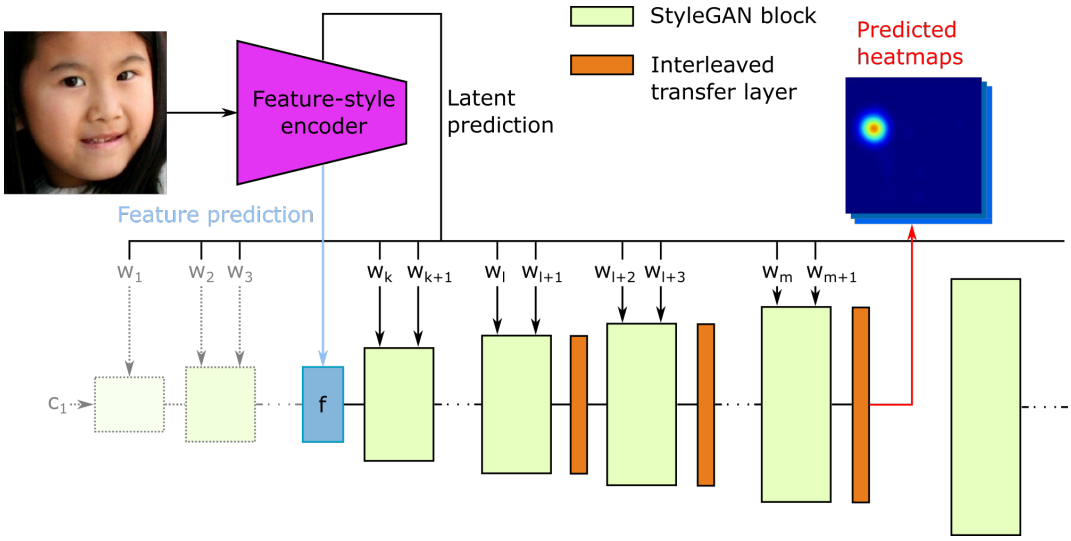

Figure 1: *Our network architecture. We use a pre-trained Feature-style encoder and StyleGAN2 generator. Additional convolution layers are interleaved with some generator convolution blocks and are trained to predict landmark heatmaps.*

We modify this architecture to predict facial landmark heatmaps. From a chosen generator block until the landmark heatmap resolution is reached, we interleave the generator blocks with Interleaved Transfer Layers (ITLs). An ITL is a 3 x 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.

4. Experiments

4.1. Datasets

We evaluated our method on three datasets commonly used to evaluate a face alignment model.

AFLW

This dataset [29] contains 24,386 face images annotated with 21 landmarks. Following usual practice [14, 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 [42] is an aggregation of five face alignment datasets: LFPW [2], AFW [53], HELEN [32], XM2VTS [35] and IBUG [42]. All these datasets have been re-annotated using 68 landmarks. Following the common splits [14, 3], the training set contains 3148 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 containing 135 images.

WFLW

This dataset introduced in [48] contains 7,500 training images and 2,500 testing images from the WIDER FACE dataset [50] 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. 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:

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

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 ($NME_{inter\text{-}ocular}$). For AFLW, because of the large number of profile face, we report both the NME normalized with the diagonal of the image (NME_{diag}) or the geometric mean of the ground truth bounding box width and height ($\sqrt{w_{bbox} * h_{bbox}}$) ($NME_{box}$). We also report the AUC at 7% NME_{box} (AUC^{7}_{box}) for this dataset.

4.3. Architecture and training parameters

The encoder is a Feature-style encoder [51] pre-trained on FFHQ [25]. The generator is StyleGAN2 generator [26] pre-trained on the same dataset. Input images are resized to 256 x 256 pixels. We use 5 ITLs (see Section 4.5 for details) and outputs heatmaps of size 128 x 128 pixels.

Our models are implemented with PyTorch [39]. 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. As optimizer, we use Adam [28] ($\beta_1 = 0.9$, $\beta_2 = 0.99$) 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 ($\pm$30°), translation ($\pm$4%), scaling ($\pm$5%), occlusions (p=50%, PyTorch default settings for the bounding box size), Gaussian blur (p=20%), brightness (p=45%, $\pm$80%) and contrast changes (p=45%, 30%-200%) as data augmentations.

When training with limited data, the training samples are chosen randomly 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.

4.4. Comparison with state-of-the-art

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. 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 [52]. 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 is closer to FFHQ, the training dataset for the StyleGAN encoder and generator. While these results are interesting, our goal with our method is to train the model with limited training. These results are discussed in the next subsection.

	300-W			WFLW
Method	Com.	Chal.	Full	Full
SAN [13]	3.34	6.60	3.98	5.22
LAB [48]	2.98	5.19	3.49	5.27
AVS [40]	3.21	6.49	3.86	4.39
DeCaFa [9]	2.93	5.26	3.39	4.62
AWing [47]	2.72	4.52	3.07	4.36
LUVLi [30]	2.76	5.16	3.23	4.37
HiH [31]	2.93	5.00	3.36	4.18
SHR-FAN [4]	2.61	**4.13**	2.94	**3.72**
ADNet [22]	**2.53**	4.58	**2.93**	4.14
FaRL [52]	2.56	4.45	**2.93**	3.96
FASE (Ours)	2.97	5.30	3.42	4.62
FASE std	0.01	0.09	0.02	0.03

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

AFLW dataset				
	NME_{diag} ↓		NME_{box} ↓	AUC^{7}_{box} ↑
Method	Full	Frontal	Full	Full
DSRN [36]	1.86	-	-	-
SAN [13]	1.91	1.85	4.04	0.540
LAB [48]	1.25	1.14	-	-
HR-Net [44]	1.57	1.46	-	-
LUVLi [30]	1.39	1.19	2.28	0.680
3FabRec [3]	-	-	1.84	-
SHR-FAN [4]	1.31	1.19	2.14	0.700
FaRL [52]	**0.94**	**0.82**	**1.33**	**0.813**
FASE (Ours)	1.02	0.90	1.45	0.791
FASE std	<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.*

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 (3148 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

WFLW dataset					
Method	Training set size				
	100%	20%	10%	5%	50
AVS [40]	**4.39**	6.00	7.20	-	-
3FabRec [3]	5.62	6.51	6.73	7.68	8.39
SCAF [15]	5.50	6.07	6.28	6.72	8.06
FASE (Ours)	4.62	**5.09**	**5.44**	**5.80**	**7.78**
FASE std	0.03	0.08	0.07	0.08	0.20

Table 3: *$NME_{inter\text{-}ocular}$ (%) (↓) when training with limited training set size on the WFLW Full test set.*

training set size goes down, even for 50 samples.

For WFLW (see Table 3), we also beat other methods when training on limited data. However, we have noticed that the performance gap is smaller on the very little training set size of 50. This may be explained by the many occlusions and difficult poses of WFLW which make it very different from FFHQ, the original training dataset of our StyleGAN encoder and generator, so our method might need more samples to train correctly compared to 300-W.

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 Frontal test sets. For the NME_{box}, we compared our method with FaRL [52], 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. Ablation studies

Encoder fine-tuning

Table 6 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. As we can see in Figure 2, if the Feature-style encoder [51] is not fine-tuned during the supervised training, the reconstructed images do not change and are close to the original images. However, the predicted heatmaps might be poor (particularly for the middle image) leading to mediocre landmark predictions. If we fine-tune the encoder, the reconstruction error increases, only the pose and the shape of the face remain in the re-

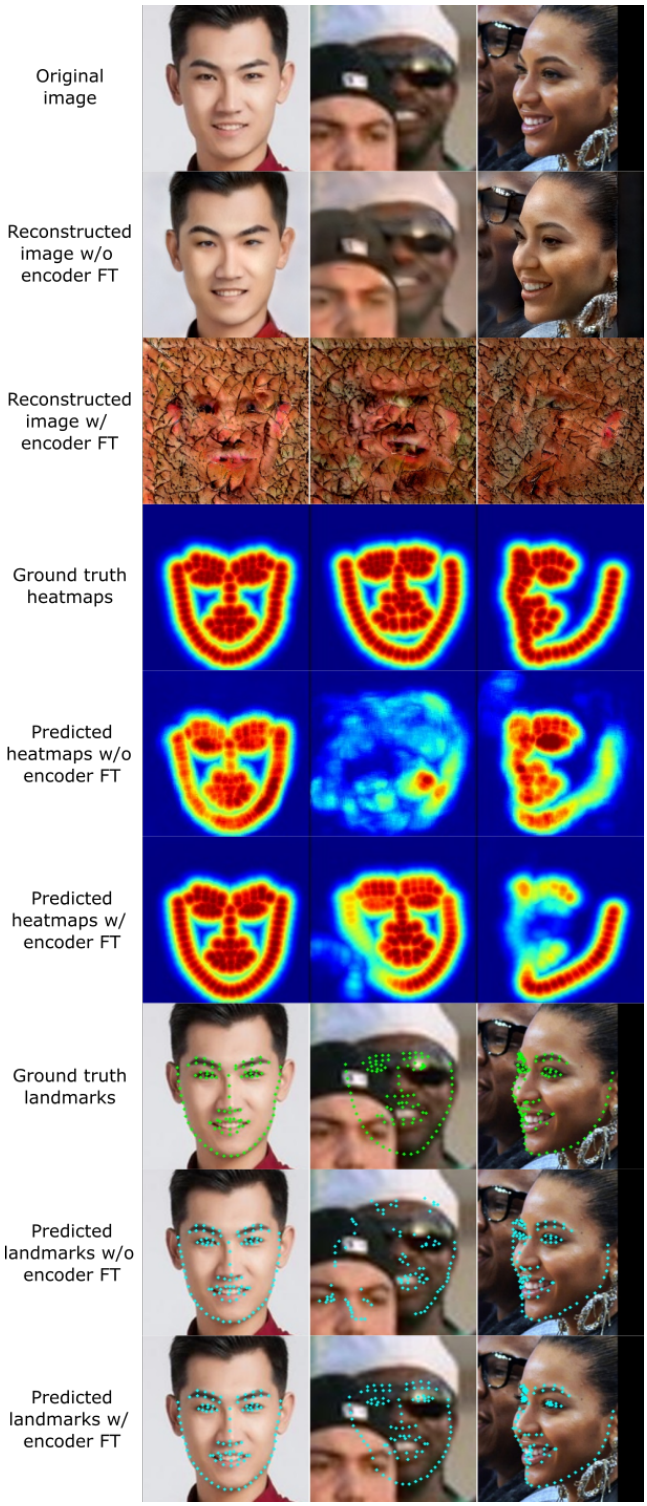

Figure 2: *Qualitative comparison with and without encoder fine-tuning (FT) on some images from the WFLW Full test set.*

300-W dataset															
Method	Training set size														
	100%			20%			10%			5%			50 (1.5%)		
RCN+ [20]	3.00	**4.98**	3.46	-	6.12	4.15	-	6.63	4.47	-	9.95	5.11	-	-	-
AVS [40]	3.21	6.49	3.86	3.85	-	-	4.27	-	-	6.32	-	-	-	-	-
TS^3 [14]	**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 [15]	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
FASE (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**
FASE 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

Table 4: *$NME_{inter\text{-}ocular}$ (%) (↓) 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).*

AFLW dataset												
Method	Training set size											
	100%		20%		10%		5%		1%		50 (0.25%)	
	NME_{box} (%) ↓											
RCN+ [20]	1.61	-	-	-	-	-	2.17	-	2.88	-	-	-
TS^3 [14]	-	-	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
FASE (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**
FASE standard deviations	<.01	<.01	<.01	0.02	<.01	0.01	<.01	0.03	0.01	0.01	0.03	0.02
	NME_{diag} (%) ↓											
FaRL [52]	**0.94**	**0.82**	-	-	**1.15**	-	-	-	1.35	-	-	-
FASE (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**
FASE standard deviations	<.01	<.01	<.01	0.01	<.01	0.01	<.01	0.02	0.01	<.01	0.02	0.02

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).*

	300-W			WFLW		
	Training set size			Training set size		
	100%	10%	50	100%	10%	50
w/o FT	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/ FT	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 6: *$NME_{inter\text{-}ocular}$ (%) (↓) on 300-W and WFLW Full test sets for different training set sizes with and without encoder fine-tuning.*

constructed image, and all other attributes such as background, hair or skin color are removed. But the quality of generated heatmaps and thus predicted landmarks is most of the time improved. The rightmost image shows a rare failure case where the landmark predictions are worse when fine-tuning the encoder.

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 7. 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 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

With FASE, we have demonstrated that StyleGAN [25] latent space can be used not only for generative tasks such as image edition but also for discriminative ones like face alignment, even if both the encoder and generator have been pre-trained on a dataset with small diversity in terms of face pose (FFHQ [25]). By modifying the generator and fine-tuning the encoder, we achieve superior results to other semi-supervised methods when training with limited data. Another advantage compared to these methods is that we don't need to perform any computationally expensive unsupervised training on large databases [3, 15, 52], prior to the supervised training, thanks to the abundance of already pre-trained StyleGAN generators and encoders available on the internet.

Num. ITLs	300-W			WFLW		
	Training set size			Training set size		
	100%	10%	50	100%	10%	50
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 7: *$NME_{inter\text{-}ocular}$ (%) (↓) on 300-W and WFLW Full test sets, depending on the number of Interleaved Transfer Layers (ITLs) and the training set size.*

Future work

Thanks to the Feature-Style encoder [51], 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.

Acknowledgments

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

References

[1] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Restyle: A residual-based stylegan encoder via iterative refinement. In *Proceedings of the IEEE/CVF*

International Conference on Computer Vision, pages 6711–6720, 2021.

[2] Peter N Belhumeur, David W Jacobs, David J Kriegman, and Neeraj Kumar. Localizing parts of faces using a consensus of exemplars. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2930–2940, 2013.

[3] Bjorn Browatzki and Christian Wallraven. 3fabrec: Fast few-shot face alignment by reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6110–6120, 2020.

[4] Adrian Bulat, Enrique Sanchez, and Georgios Tzimiropoulos. Subpixel heatmap regression for facial landmark localization. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2021.

[5] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020.

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[7] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2172–2180. Curran Associates, Inc., 2016.

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893 vol. 1, 2005.

[9] Arnaud Dapogny, Kevin Bailly, and Matthieu Cord. Decafa: Deep convolutional cascade for face alignment in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6893–6901, 2019.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina 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)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[12] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. *International Conference on Learning Representations*, 2016.

[13] Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Style aggregated network for facial landmark detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 379–388, 2018.

[14] Xuanyi Dong and Yi 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*, pages 783–792, 2019.

[15] Martin Dornier, Philippe-Henri Gosselin, Christian Raymond, Yann Ricquebourg, and Bertrand Coüasnon. Scaf: Skip-connections in auto-encoder for face alignment with few annotated data. In *International Conference on Image Analysis and Processing*, pages 425–437. Springer, 2022.

[16] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Olivier Mastropietro, Alex Lamb, Martin Arjovsky, and Aaron Courville. Adversarially learned inference. *International Conference on Learning Representations*, 2017.

[17] Jinjin Gu, Yujun Shen, and Bolei Zhou. Image processing using multi-code gan prior. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3012–3021, 2020.

[18] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[19] Irina Higgins, Loïc Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew M. Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

[20] Sina Honari, Pavlo Molchanov, Stephen Tyree, Pascal Vincent, Christopher Pal, and Jan Kautz. Improving landmark localization with semi-supervised learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1546–1555, 2018.

[21] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017.

[22] Yangyu Huang, Hao Yang, Chong Li, Jongyoo Kim, and Fangyun Wei. Adnet: Leveraging error-bias towards normal direction in face alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3080–3090, 2021.

[23] Kyoungkook Kang, Seongtae Kim, and Sunghyun Cho. Gan inversion for out-of-range images with geometric transformations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13941–13949, 2021.

[24] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34, 2021.

[25] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[26] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.

[27] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, pages 2649–2658. PMLR, 2018.

[28] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[29] Martin Koestinger, Paul Wohlhart, Peter M Roth, and Horst 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)*, pages 2144–2151. IEEE, 2011.

[30] Abhinav Kumar, Tim K Marks, Wenxuan Mou, Ye Wang, Michael Jones, Anoop Cherian, Toshiaki Koike-Akino, Xiaoming Liu, and Chen Feng. Luvli face alignment: Estimating landmarks' location, uncertainty, and visibility likelihood. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8236–8246, 2020.

[31] Xing Lan, Qinghao Hu, and Jian Cheng. Revisiting quantization error in face alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1521–1530, 2021.

[32] Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, and Thomas S Huang. Interactive facial feature localization. In *European conference on computer vision*, pages 679–692. Springer, 2012.

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

[34] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)*, pages 181–196, 2018.

[35] Kieron Messer, Jiri Matas, Josef Kittler, Juergen Luettin, Gilbert Maitre, et al. Xm2vtsdb: The extended m2vts database. In *Second international conference on audio and video-based biometric person authentication*, volume 964, pages 965–966. Citeseer, 1999.

[36] Xin Miao, Xiantong Zhen, Xianglong Liu, Cheng Deng, Vassilis Athitsos, and Heng Huang. Direct shape regression networks for end-to-end face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5040–5049, 2018.

[37] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016.

[38] Yotam Nitzan, Rinon Gal, Ofir Brenner, and Daniel Cohen-Or. Large: Latent-based regression through gan semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19239–19249, 2022.

[39] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[40] Shengju Qian, Keqiang Sun, Wayne Wu, Chen Qian, and Jiaya Jia. Aggregation via separation: Boosting facial landmark detector with semi-supervised style translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10153–10163, 2019.

[41] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel 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*, pages 2287–2296, 2021.

[42] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 397–403, 2013.

[43] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 843–852, 2017.

[44] Ke Sun, Yang Zhao, Borui Jiang, Tianheng Cheng, Bin Xiao, Dong Liu, Yadong Mu, Xinggang Wang, Wenyu Liu, and Jingdong Wang. High-resolution representations for labeling pixels and regions. *CoRR*, abs/1904.04514, 2019.

[45] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021.

[46] Tengfei Wang, Yong Zhang, Yanbo Fan, Jue Wang, and Qifeng Chen. High-fidelity gan inversion for image attribute editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11379–11388, 2022.

[47] Xinyao Wang, Liefeng Bo, and Li Fuxin. Adaptive wing loss for robust face alignment via heatmap regression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6971–6981, 2019.

[48] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2129–2138, 2018.

[49] Yinghao Xu, Yujun Shen, Jiapeng Zhu, Ceyuan Yang, and Bolei Zhou. Generative hierarchical features from synthesizing images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4432–4442, 2021.

[50] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[51] Xu Yao, Alasdair Newson, Yann Gousseau, and Pierre Hellier. Feature-style encoder for style-based gan inversion. *arXiv e-prints*, pages arXiv–2202, 2022.

[52] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18697–18709, 2022.

[53] Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2879–2886. IEEE, 2012.